Social media platforms and challenges for democracy, rule of law and fundamental rights Policy Department for Citizens' Rights and Constitutional Affairs Directorate-General for Internal Policies PE 743.400 -April 2023

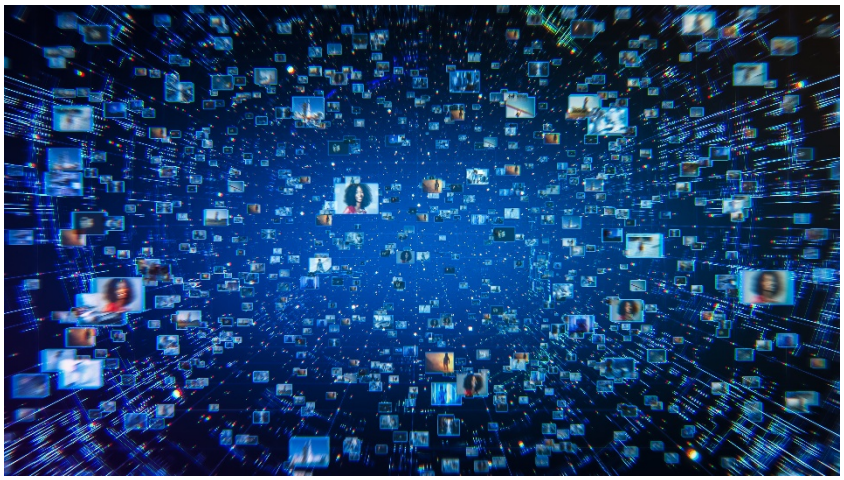
Beatriz Botero Arcila, Rachel Griffin

STUDY

Requested by the LIBE committee

European Parliament

Social media platforms and challenges for democracy, rule of law and fundamental rights



Policy Department for Citizens' Rights and Constitutional Affairs
Directorate-General for Internal Policies
PE 743.400 – April 2023

EN

Social media platforms and challenges for democracy, rule of law and fundamental rights

Abstract

This study, commissioned by the European Parliament's Policy Department for Citizens' Rights and Constitutional Affairs at the request of the LIBE Committee, examines risks that contemporary social media - focusing in particular on the most widely-used platforms - present for democracy, the rule of law and fundamental rights. The study focuses on the governance of online content, provides an assessment of existing EU law and industry practices which address these risks, and evaluates potential opportunities and risks to fundamental rights and other democratic values.

This document was requested by the European Parliament's Committee on Civil Liberties, Justice and Home Affairs

AUTHORS
Beatriz BOTERO ARCILA, Assistant Professor, Sciences Po Law School
Rachel GRIFFIN, PhD candidate at Sciences Po Law School

ACKNOWLEDGMENTS
Thank you to Giovanna Hajdu Hungria da Custódia, Ana Paula Rios Camarena, Ioan Paul Sipos and Andrea Ticchi for their invaluable editorial assistance and research assistance with some of the boxes. Ana Paula Rios Camarena also provided invaluable research assistance with Chapter 5 on media pluralism. Thank you, lastly, to Annina Claesson and Philipp Darius for their comments and advice on the literature review for Chapter 4 on disinformation.

ADMINISTRATOR RESPONSIBLE
Ina SOKOLSKA

EDITORIAL ASSISTANT
Ewelina MIAZGA

LINGUISTIC VERSIONS
Original: EN

ABOUT THE EDITOR
Policy departments provide in-house and external expertise to support EP committees and other parliamentary bodies in shaping legislation and exercising democratic scrutiny over EU internal policies.

To contact the Policy Department or to subscribe for updates, please write to:
Policy Department for Citizens' Rights and Constitutional Affairs
European Parliament
B-1047 Brussels
Email: poldep-citizens@europarl.europa.eu

Manuscript completed in April 2023

This document is available on the internet at:
http://www.europarl.europa.eu/supporting-analyses

CONTENTS

LIST OF ABBREVIATIONS

AI	Artificial intelligence
AVMSD	Audiovisual Media Services Directive
CD	Copyright Directive
CoC	Code of Conduct
CoP	Code of Practice
DSA	Digital Services Act
DSC	Digital Services Coordinator
EU	European Union
ECD	E-Commerce Directive
ECHR	European Convention on Human Rights
ECtHR	European Court of Human Rights
ECJ	European Court of Justice
GDPR	General Data Protection Regulation
IRU	Internet Referral Unit
NetzDG	German Network Enforcement Act
HLEG	High Level Expert Group on Fake News and Online Disinformation
IRU	Internet Referral Unit
LGBTQ+	Lesbian, gay, bisexual, transgender, queer and others
OECD	Organisation for Economic Cooperation and Development
NGO	Non-governmental organisation
PAR	Political Advertising Regulation
PTSD	Post-traumatic stress disorder
TCR	Terrorist Content Regulation

UK	United Kingdom
USA	United States of America
VLOP	Very large online platform

LIST OF BOXES

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

Background

Social media platforms are now key infrastructure of European information environments. The rise of social media has created vast opportunities to access and share information: they had a vital role in keeping families, friends and even workplaces connected during the peak of the Covid-19 pandemic, and have enabled and supported civic movements around the world. At the same time, they have also brought new challenges for democracy, rule of law and fundamental rights. Social media platforms have often been conduits for disinformation, undermining citizens' access to reliable information and the democratic process, and they have enabled the wider spread of hate speech, impacting the fundamental rights, dignity and safety of people in Europe. The challenge for policymakers is thus to strengthen accountability and oversight of social media in order to protect citizens against such threats, without curtailing access to the many benefits that they provide.

Aim

This study examines risks that contemporary social media - focusing in particular on the most widely-used platforms like Facebook, Instagram, TikTok, YouTube and Twitter - present for democracy, the rule of law and fundamental rights. Specifically, it focuses on the governance of online content: that is to say, the media and communications practices that take place on social media platforms, rather than issues such as how platform businesses are organised or how they handle user data. The study further provides an assessment of existing EU law and industry practices which address these risks, and evaluates potential opportunities and risks to fundamental rights and other democratic values. On this basis, the study makes recommendations for policymakers, relating both to the enforcement and implementation of existing law, and to possibilities for further legislative reform or other new policy initiatives.

The legal framework for social media content governance in the EU

Chapter 2 provides a high-level overview of the existing legal framework governing social media content, explaining how it shapes commercial platforms' content governance practices and highlighting recognised or potential threats to fundamental rights and democratic values. In this context, the study covers three broad areas: the overarching framework for content moderation set out in the 2022 Digital Services Act (DSA), the various other regulations that address content moderation in specific areas, and the nascent regulatory framework governing content recommendations and other aspects of platform design.

The DSA now sets out the ground rules for content moderation (platforms' enforcement of legal and voluntary standards about what content they will host, and how they will deal with harmful content). The baseline principle of content regulation is that social media companies are exempt from liability for hosting content that is illegal, so long as they do not participate in its production and remove illegal content as soon as they are made aware of it. The chapter explains the 'notice and takedown' framework for illegal content, and the 'due diligence' obligations that the Digital Services Act creates regarding the operation of platforms' moderation policies and procedures (which also apply to the voluntary moderation of legal content).

The Digital Services Act will coexist with a number of legislative and soft law instruments providing additional regulation for particular types of platform or content. The chapter highlights and briefly describes five: the 2019 Copyright Directive, the 2021 Terrorist Content Regulation, the 2018 updated Audiovisual Media Services Directive, and the co-regulatory Codes on Hate Speech and Disinformation.

Finally, EU law also - albeit to a more limited extent - regulates aspects of online content governance beyond moderation, such as how platforms recommend content to users. The chapter finishes by outlining how recommendations and other aspects of platform design are regulated in the Digital Services Act; how targeted advertising is regulated in the Digital Services Act and the proposed Regulation on the transparency and targeting of political advertising; and the promotion of 'safe design practices' in the 2022 updated Code of Practice on Disinformation.

Threats to fundamental rights and equal participation in democratic debate associated with online hate speech

Chapter 3 provides an in-depth analysis of the issue of hate speech on social media. The chapter briefly reviews existing empirical evidence on hate speech on social media and its consequences for fundamental rights, democracy and the rule of law, examines the challenges of addressing hate speech online, and evaluates the existing legal framework to deal with these issues.

The chapter notes that hate speech does not only violate the fundamental rights of those targeted, but also more broadly undermines equal participation in the public sphere and in democratic debate. Existing human rights law and jurisprudence suggests that banning and censoring such forms of harmful content via content moderation can sometimes be justified to protect the rights of others, as well as these broader social interests in safety and equality in the public sphere. However, content moderation is not a sufficient solution and raises its own fundamental rights concerns.

The chapter highlights three areas of particular concern. First, content moderation is highly unreliable. Platforms often fail to moderate serious hate speech, while removing large volumes of valuable and/or harmless content. In this regard, there are also significant geographic and linguistic disparities, with moderation far less effective for users in less wealthy and non-English-speaking markets. The study suggests several policy measures that could be taken within the Digital Services Act framework to address this. Second, content moderation is highly discriminatory, and disproportionately suppresses content from marginalised users. Consequently, simply trying to moderate ever more harmful content is not just inadequate to address hate speech, but will actively undermine fundamental rights and equality. Third, however, marginalised groups need more protection against online hate speech. To address this, instead of simply expanding moderation, platforms should focus on developing more holistic and systemic interventions, for example through design changes which can proactively discourage hate speech.

Informed by this literature review, the chapter evaluates the existing legal framework and its ability to address these issues. It highlights two main issues. First, the 2016 Code of Conduct on Online Hate Speech - which defines hate speech as incitement to violence or hatred based on race, religion, ethnicity or nationality - is too narrow to address the individual and social impacts of online hate. EU policy needs to recognise marginalisation based on other characteristics, such as gender identity and sexuality, as well as intersectional marginalisation, where people are targeted for reasons that cannot be reduced to a single protected identity category. Additionally, EU policymakers should broaden their focus to include abusive and exclusionary behaviour targeting marginalised groups which does not involve incitement to violence or hatred, such as harassment or privacy violations.

Second, the EU's encouragement of automated moderation (the use of software to automatically filter and remove user content, without direct human intervention) as a primary response to hate speech raises fundamental rights concerns. At the same time, the focus on moderation of individual pieces of harmful content does not give adequate weight to more structural, design-based interventions. However, the chapter highlights some aspects of the legislative framework which could promote more systemic interventions and recommends steps that regulators could take to maximise the benefits of

these provisions. In particular, developing a new Code of Conduct on Hate Speech could effectively incentivise and provide accountability for such systemic improvements.

Disinformation and its effects on public safety, fundamental rights and democratic debate

Chapter 4 analyses contemporary issues and recent developments in the spread of disinformation on social media. Regulating disinformation implicates the rights to freedom of speech and information, as well as raising broader concerns around the rule of law and democratic debate. Both platforms and regulators in the EU have had to grapple with the tension between protecting citizens against harmful disinformation and maintaining trust in the information environment, without threatening fundamental rights and political freedoms by centralising control over the 'truth' and speech.

Disinformation research is a vast and complex field, where fundamental questions about the causal effects of disinformation and the role of social media remain unresolved. The chapter starts by providing a necessarily brief overview of relevant empirical literature, drawing attention to tentative conclusions and unresolved questions. The expert consensus is that online disinformation should not be considered in isolation, but as one dynamic element of a broader social and political environment characterised by increasing polarisation and mistrust in institutions and the media. At the same time, widespread online disinformation also raises concerns around 'second-order effects' on trust in media and politics, even where it does not directly cause harm. This chapter's analysis and recommendations should thus be read in conjunction with Chapter 5 on how to strengthen democracy and trust in the media more generally.

The chapter then provides an overview of current responses to online disinformation by social media platforms, which include content moderation, fact-checking and design interventions. It also outlines the existing EU legal framework, which includes both hard regulation requiring platforms to remove certain forms of disinformation deemed illegal under national law, and soft law measures which encourage them to voluntarily moderate content or implement other preventive measures.

The chapter's key argument, however, is that disinformation which directly encourages violence or harmful behaviour, or which is spread by organised strategic disinformation operations, present the greatest threats to fundamental rights and democracy, and disinformation policy should be targeted towards these areas. To strengthen fundamental rights protection, the chapter suggests that the Digital Services Act should be amended to introduce stronger safeguards against removal of speech based only on assessments of accuracy. The 2022 updated Code of Practice on Disinformation includes positive elements, such as promoting a more systemic approach to discouraging disinformation through 'safe design practices', as well as some that are more concerning, such as the promotion of 'brand safety' measures through which advertisers can influence platforms to suppress content they consider inappropriate. The chapter suggests how policymakers can build on its positive elements, in collaboration with civil society, industry and independent researchers, to promote effective and fundamental rights-respecting measures against disinformation.

Finally, the chapter notes the relevance of micro-targeted political advertising to disinformation, as well as to trust, polarisation and inclusion in democratic debate more generally. For example, targeting narrowly-defined audiences with political messages can undermine constructive political debate based on shared understandings of the political landscape, and weaken political figures' accountability. The chapter recommends incorporating stronger restrictions on targeted advertising into the proposed Regulation on the transparency and targeting of political advertising to address this.

Risks to the capacity of news media to support pluralist political debate and promote democratic participation and accountability

Chapter 5 analyses how the growing popularity and influence of social media has impacted media pluralism in Europe, focusing on the news media due to its particular importance for democratic processes. Like the above chapters, it provides a brief review of the empirical literature and evaluates the existing EU legal framework.

Social media are now a major source of audiences and traffic for news publishers, meaning that platforms increasingly influence what journalists cover and how. At the same time, the rise of digital advertising has threatened existing business models in the news industry, and made publishers increasingly dependent on digital advertising intermediaries - a market heavily dominated by Google and (Facebook and Instagram owner) Meta. Lack of transparency and competition in digital advertising, and social media generally, exacerbates this dependence.

These trends have particularly undermined local journalism, with concerning implications for political participation and accountability. New business models such as paywalls and subscriptions, with which publishers have attempted to compensate for lost advertising revenue, have often favoured the biggest and best-known news brands, reducing pluralism. The study examines these developments in the context of wider economic trends, such as the financial difficulties faced by publishers since the 2008 financial crisis and the consolidation of large media companies.

The analysis of recent regulatory developments, notably the European Media Freedom Act and the new press publishers' right introduced by the Copyright Directive, suggests that they do not adequately address the structural trends favouring consolidation and threatening smaller-scale and local journalism. Consequently, the chapter advocates for the expansion of subsidy programmes for independent media, especially local and regional media, and discusses how EU institutions could promote new pilot schemes and the exchange of knowledge and best practices between member states.

Summary of recommendations

Chapter 6 summarises the detailed recommendations presented in each in-depth chapter. These can broadly be grouped in three areas.

DSA enforcement

The Digital Services Act leaves many open questions - for example, regarding the contours of very large platforms' obligations to assess and mitigate systemic risks, which will be essential in addressing systemic issues such as hate speech and disinformation. The study presents a number of detailed recommendations as to how national regulators and the Commission can effectively implement and expand on relevant provisions to ensure that platforms take effective measures against disinformation and hate speech, while respecting users' fundamental rights.

Legislative reform

The study identifies gaps in the Digital Services Act framework where further legislative reform could strengthen the protection of fundamental rights and democratic processes. These relate in particular to three areas: the regulation of content moderation labour (for example, the capacities, training and working conditions of moderators); strengthened safeguards against state-mandated censorship of content deemed potentially illegal under national laws which do not adequately respect freedom of expression; and more stringent restrictions on the personalised targeting of political advertising (which

could be introduced in the ongoing legislative process for the proposed Political Advertising Regulation).

Funding and policy programmes

Finally, funding and support from the EU can play a vital role in strengthening the broader ecosystem of civil society and media which will be essential in supporting healthy democratic debate in the age of social media. In particular, the report highlights three priority areas: subsidising independent media, especially local media; promoting the development of professional associations for platform staff working on security, equality and other important areas of content policy; and supporting and expanding media literacy programmes.

1. SOCIAL MEDIA, DEMOCRACY, FUNDAMENTAL RIGHTS AND THE RULE OF LAW

1.1. Introduction

Social media platforms are now a key part of the infrastructure of European information environments. Indeed, according to the Eurobarometer's News & Media Survey of 2022, 49% of respondents use social media for communication purposes and 45% use social media for information purposes, that is, to stay updated on the news and current events.[1] Facebook is the most popular social media platform, mentioned by almost 70% of respondents, but the landscape is varied; TikTok, for example, is used by 49% of 15-24 year-olds in Europe.[2]

The rise of social media has created vast opportunities to access and share information. Social media platforms had a vital role in keeping families, friends and even workplaces connected during the peak of the Covid-19 pandemic; they have also enabled and supported civic movements around the world. At the same time, they have also brought new challenges for democracy, rule of law and fundamental rights. The freedoms and rights associated with expressing, accessing and receiving information are central to democracy, the rule of law and the exercise of many other fundamental rights. However, social media platforms are often conduits to amplify mis - and disinformation, undermining citizens' access to reliable information and the democratic process, and are thought by many experts to have weakened the capacity of traditional media to support informed and constructive political debate.[3] They have also enabled new forms of hate speech, abuse and harassment, affecting the fundamental rights, equality and safety of people in Europe.

This study examines risks that today's most widely-used social media platforms present for democracy, the rule of law and fundamental rights. It further provides an assessment of existing EU law and industry practices which address these risks. On that basis, it presents recommendations for legal reform and enforcement, focusing in particular on how EU and national authorities could best implement and build on the legislative framework established in the 2022 the Digital Services Act (DSA).

The focus of this study is on regulation of online content by the EU, and on challenges to the rule of law, democracy and fundamental rights which are associated with the production and dissemination of information by social media companies. These are not the only challenges to democracy posed by social media - for example, there are also widely-discussed concerns around the huge size and market power of leading social media companies, and their data collection and analytics capacities. However, this report only addresses these aspects of the social media industry and business models tangentially and insofar as they relate directly to the governance of online content.

The rest of this introductory chapter will outline the understanding of technologies, the rule of law, democracy and fundamental rights that provides the general framework for this study. Chapter 2 then reviews the legal framework governing social media platforms in the EU. The chapters that follow provide in-depth analyses of three main areas of risk to democracy, fundamental rights, and the rule of

---

[1] Eurobarometer, *Media & News Survey 2022,* n.d. July 2022, p. 30, available at: https://europa.eu/eurobarometer/surveys/detail/2832.

[2] Eurobarometer, *Media & News Survey 2022,* n.d. July 2022, pp. 29-30, available at: https://europa.eu/eurobarometer/surveys/detail/2832.

[3] For an overview of the conflicting evidence and academic debates in this area see Haidt, J., & Bail, C. 'Social media and political dysfunction: A collaborative review', New York, 2022, available at: https://docs.google.com/document/d/1vVAtMCQnz8WVxtSNQev_e1cGmY9rnY96ecYuAj6C548/edit.

law. Chapter 3 analyses threats to the rights to freedom of expression, equality, and democratic debate associated with hate speech on social media. Indeed, hate speech not only threatens victims' safety and dignity, but also has chilling effects over the freedom of expression of those targeted, affecting the abilities of marginalised groups to participate in online media and benefit from the opportunities they offer. Chapter 4 explores current issues around disinformation, and its effects on the right to access information and participate in democratic debate and social life. Chapter 5 analyses the risks that the increasing importance of social media platforms as intermediaries for the production and consumption of news and journalism poses to media pluralism, which is a key element of democratic accountability and nourishment of democratic debate. Finally, Chapter 6 summarises the recommendations made throughout the report to improve the EU legal framework and its enforcement, and the governance of social media content in general.

1.2. Fundamental rights, democracy, the rule of law: the concepts

Democracy, the rule of law and fundamental rights are key foundations of the EU, and democracy is a precondition for EU membership. The Charter of Fundamental Rights establishes that 'the Union is founded on the indivisible, universal values of human dignity, freedom, equality and solidarity; it is based on the principles of democracy and the rule of law.'[4]

The rule of law, the protection of fundamental rights, and democracy are interdependent concepts and institutions.[5] Although their definitions are contested,[6] democracy is generally rooted in the idea that each citizen is entitled to participate in the key decisions of common life, and democratic institutions set the rules and procedures for such participation. These fundamental entitlements are derived from the idea that all individuals are 'endowed with reason and conscience"[7] and 'are born free and equal in dignity and in rights'[8] - particularly human or fundamental rights.[9] At the same time, full and equitable participation requires more than formally equal rights to participate in political processes - it also requires state institutions to recognise and guarantee other fundamental rights and freedoms, such as freedom of thought,[10] freedom of expression and information.[11] It also requires the right to integrity[12],

---

4 European Parliament, European Council, European Commission, *Charter of Fundamental Rights of the European Union*, par. 2.

5 Leslie D., et. al. *Artificial Intelligence, Human Rights, Democracy, and The Rule of Law. A Primer.* The Alan Turing Institute, p. 12, available at: https://www.turing.ac.uk/sites/default/files/2021-03/cahai_feasibility_study_primer_final.pdf.

6 Spicer, M., 'What Do We Mean by Democracy? Reflections on an Essentially Contested Concept and Its Relationship to Politics and Public Administration' *Administration & Society,* Vol. 51, No. 5, 2018; Waldron, J. 'Is the Rule of Law a Essentially Contested Concept (In Florida)?' Vol. 21, No. 2, In the Wake of Bush v. Gore: Law, Legitimacy and Judicial Ethics, 2002.

7 General Assembly of the United Nation, *Universal Declaration of Human Rights*, Article 1.

8 General Assembly of the United Nation, *Universal Declaration of Human Rights*, Article 1.

9 Human rights generally refer to rights protected in international law which are intended to protect the dignity and fundamental interests of all individuals, regardless of social status. The term 'fundamental rights' is used more often than 'human rights' in the EU, where it refers specifically to human rights principles which have been developed as binding legal norms by the ECJ, and are now set out in the Charter of Fundamental Rights. As this study's focus in EU law, the term 'fundamental rights' will generally be used. However, EU fundamental rights law often aligns with and is influenced by international human rights law (and in particular by the European Convention of Human Rights and associated case law, which provide authoritative guidance on the interpretation of corresponding rights in the EU Charter).

10 European Parliament, European Council, European Commission, Charter of Fundamental Rights of the European Union, Article 10.

11 European Parliament, European Council, European Commission, Charter of Fundamental Rights of the European Union, Article 10.

12 European Parliament, European Council, European Commission, Charter of Fundamental Rights of the European Union, Article 3.

liberty and security,[13] as well as the material and economic conditions which are necessary for citizens to participate in society on equal terms.[14] In turn, the rule of law has been described as an essentially contested concept, meaning that it has no fixed definition and that employing the term necessarily involves making value claims for a particular understanding of the concept.[15] Broadly, however, it generally stands for the guarantee that individuals will be governed according to predictable rules[16] with the ultimate purpose of guaranteeing dignity and equal treatment for everyone.[17]

Within academic literature, several different understandings of the rule of law can be identified. Some of the seminal work on the concept in legal philosophy takes a formalist approach, in which the rule of law essentially consists of governance in accordance with consistent and predictable rules.[18] On the other hand, in modern legal and political debates, it is commonly understood in a broader and more substantive sense, requiring not only consistent and predictable application of the law, but also some degree of respect for human rights.[19] In recent years, this approach has gained increasing prominence in the literature on social media governance, through its development and application by scholars positioning themselves within the field of digital constitutionalism.[20] A further development of the concept, which can be described as a more teleological approach, holds that the rule of law should be defined in accordance with the values it is ultimately meant to serve, most importantly human dignity and equality. Substantive respect for the rule of law must thus go beyond formal equality and individual rights, to create the structural and institutional foundations for equal participation in society.[21]

The version of the rule of law endorsed and operationalised by EU policy can be regarded as drawing from all three approaches. The EU's regular reporting on respect for the rule of law in individual Member States focuses on four key areas: institutional checks and balances, independence of the media, independence of the judiciary, and anti-corruption measures.[22] Through improving policy in these areas, the EU aims not only to ensure fair and consistent application of the law, but also to guarantee substantive protection of individual rights, and to establish the structural and institutional factors for a basic level of equal participation in political processes and democratic debate, thus integrating all three of the approaches discussed above. Extending this holistic understanding of the rule of law to the social media context, efforts to strengthen the rule of law should not be limited to protecting individual rights and guaranteeing predictable application of rules, but should encompass

---

13 European Parliament, European Council, European Commission, *Charter of Fundamental Rights of the European Union*, Article 5.

14 Rawls, J., *Justice as Fairness: A Restatement*, Belknap Press, 2001.

15 Waldron, J. 'Is the Rule of Law a Essentially Contested Concept (In Florida)?' Vol. 21, No. 2, In the Wake of Bush v. Gore: Law, Legitimacy and Judicial Ethics, 2002.

16 Fuller, L., The Morality of Law: Revised Edition, Yale University Press, New Haven, U.S., 1969; Raz J., The authority of law: Essays on law and morality, Oxford University Press, Oxford, U.K., 1979.

17 Kampourakis, I. et. al., 'Reappropriating the Rule of Law: Between Constituting and Limiting Private Power', Jurisprudence 2022, n.d.

18 Fuller, L., The Morality of Law: Revised Edition, Yale University Press, New Haven, U.S., 1969; Raz J., The authority of law: Essays on law and morality, Oxford University Press, Oxford, U.K., 1979.

19 Bingham, T., The Rule of Law, Penguin Books Limited, 2011.

20 De Gregorio, G., Digital Constitutionalism in Europe. Reframing Rights and Powers in the Algorithmic Society., Cambridge University Press, Cambridge, U.K., 2022.

21 Kampourakis, I. et. al., 'Reappropriating the Rule of Law: Between Constituting and Limiting Private Power,' Jurisprudence 2022.

22 European Commission, 2022 Rule of law report, n.d. July 2022, available at: https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/upholding-rule-law/rule-law/rule-law-mechanism/2022-rule-law-report_en.

substantive protection of fundamental rights – including positive and collective rights, such as media freedom and pluralism – as well as policies aimed at promoting equal participation in the online public sphere.

Democracy is also an essentially contested concept. Derived from the Greek *demos* 'people' and *kratos* 'rule', it essentially refers to a form of government in which the people who are ruled in a given polity hold ultimate political authority to decide on their governance.[23] In modern usage, dating back to the French Revolution, democracy generally refers to the type of government that gives everyone, regardless of education or property, a right to vote and participate in decision-making.[24] However, this minimal definition is not necessarily compatible with individual liberties and political freedom, as majorities can easily vote to limit the freedom and political participation of minorities. A more contemporary usage of the word democracy is constitutional democracy, which can be traced back to the American constitution and liberal constitutions drafted in Europe in the 19th century, and refers to systems of government based on majority rule, which is however limited by the equal rights of citizens.[25] Such rights are usually enshrined in constitutions and binding bills of rights.[26] Upholding fundamental rights principles which place constraints on legislative power is a central aspect of EU governance, as discussed in more detail below.

Political theorist Bernard Crick argued that the conditioning factors for democracy include the official rules of the electoral system, government and legislative institutions but also - importantly for the purposes of this study - attitudes to knowledge and the diffusion of information.[27] Crick argues that in autocracies, knowledge is seen as a unified instrument of political power, such as an unpublished reason of state or undebatable moral truths. In contrast, in modern democracies, knowledge is fragmented and contingent: social truths should be open to public debate. At the same time, modern states are crucially reliant on the production and dissemination of knowledge about the societies they govern,[28] and democratic government depends on the existence of civil society and media institutions which allow people to inform themselves, discuss and form public

opinions on social and political issues.[29] In particular, the news media have the role of providing reliable information, enabling political debate and discussion, and facilitating political accountability. Ideally, they should also prevent the spreading of false information.

---

[23] Crick, B., '*Democracy: A Very Short Introduction*,' Oxford, Oxford University Press, 2022, p. 1. Aristotle, *Nicomachean Ethics*, Book VIII, Chapter 10 (1160a.31-1161a.9). Internet Classics Archive. Retrieved 21 June 2018.

[24] This understanding has evolved considerably over time: for example, early liberal democracies in Europe generally did not extend the vote to women and those without property. See Crick, B., '*Democracy*: *A Very Short Introduction*,' Oxford, Oxford University Press, 2022, p. 13.

[25] There are two main modes of modern democracy: participatory democracy privileges opportunities for citizens to make direct contributions to decision-making, via referendum or plebiscite on a regular basis, or for certain key for decisions representative democracy privileges the election of representatives - parliamentarian, or a president and representatives who will make most of the governmental decisions. Many modern democracies today have elements of both.
Schiller, T. *'Direct Democracy'*, *Encyclopedia Britannica*, 07 Oct. 2022, available at: https://www.britannica.com/topic/direct-democracy.

[26] Dahl, R. "Democracy". *Encyclopedia Britannica*, 10 Jan. 2023, available at: https://www.britannica.com/topic/democracy.

[27] Crick, B., '*Democracy*: *A Very Short Introduction*,' Oxford, Oxford University Press, 2022, p. 97.

[28] Illustrating this, the term 'statistics' originally meant the science of statecraft. The production of data and knowledge about populations was central to the development of modern states, and contemporary democracies continue to invest heavily in centres of learning to produce and disseminate knowledge, such as state-sponsored universities or research grants and centres. See Scott, J.C, *Seeing Like a State*, Yale University Press, 2008; Crick, B., '*Democracy*: *A Very Short Introduction*,' Oxford, Oxford University Press, 2022.

[29] Habermas, J. The structural transformation of the public sphere: An inquiry into a category of bourgeois society. Cambridge, U.K.,: Polity, 1989.

Crick goes so far as to say that '[t]he effective working of democratic regimes comes to depend more and more on people having access to reasonably accurate information about how the state is run and on the state being able to assess public needs and reactions reasonably accurately. Hence the objective need for neutrality and objectivity in official publications, in stark comparison to all knowledge being seen as either propaganda or as secrets of state in totalitarian regimes.'[30] In this context, a key question for democracies in the age of social media is how to strengthen journalism and the diffusion of knowledge and information that is neutral and objective - but also, perhaps, what are the new institutional forms or actors that will deliver and diffuse objective and quality information in a trustworthy way in today's networked, information-rich media environment.

EU fundamental rights law echoes the central role of knowledge, diffusion of information, and political debate in democratic governance. Indeed, Article 11 of the Charter establishes the right to freedom of expression, and the right to receive and impart information without the interference of public authority. Article 11(2) further establishes that 'the freedom and pluralism of the media shall be respected.' In this context, freedom of expression is not just about individual liberties, but requires the structural and material conditions for a free and diverse media ecosystem. The EU has already recognised that social media companies pose risks to some of these values, for example by encouraging increasing concentrations of power and lack of diversity in the news media, and facilitating the spread of disinformation and hate speech. In recent years, it has developed a policy framework and several specific legislative and soft law measures to address these issues. As the following chapters will discuss in more detail, this will require not only strengthening existing institutions that have been challenged but understanding some of the new dynamics of online media, and strengthening and sponsoring new institutions that can promote and protect the flow of trustworthy information.

Finally, the protection of fundamental rights is today a constitutive element of modern democracies, and a foundational principle of the EU. Fundamental rights are a narrow category of rights that should attach to everyone in a given polity and have a high degree of protection due to their importance. They echo democratic principles like dignity, fairness and equality, and recognise special, basic interests associated with those values. International human rights treaties aim to recognise fundamental rights as pre-political individual entitlements and institutionalise them as the outer boundaries of state action,[31] protecting people from the 'tyranny of the majority'.[32] Later generations of rights evolved out of the realities of urban and industrial economies and the realisation that freedom, dignity and democratic participation are only possible when basic needs are fulfilled. Socioeconomic rights are thus oriented towards the guarantee of basic material needs, like healthcare and fair wages.[33]

Fundamental rights protection became a defining feature of European constitutions adopted after World War II, often involving specialised constitutional courts which review the compatibility of statutes with the constitution and its fundamental rights.[34] Like other modern constitutional democracies, the EU has today a complex system of fundamental rights protection. EU institutions and Member States implementing EU policy are limited by European fundamental rights principles. These were originally developed in the case law of the ECJ but are now, since the 2009 Treaty of Lisbon, authoritatively set out in the Charter of Fundamental Rights, which has equal legal status with the other

---

30 Crick, B., '*Democracy*: *A Very Short Introduction*,' Oxford, Oxford University Press, 2022, p. 98.

31 Fabbrini, F., '*Fundamental Rights in Europe,'* Oxford, Oxford University Press, Oxford, 2014, p. 16.

32 Clapham, A. '*Human Rights: A Very Short Introduction (2nd Edn.)* Oxford, Oxford University Press, 2015, p. 4.

33 United Nations - General Assembly, 'International Covenant on Economic, Social, and Cultural Rights,' *General Assembly Resolution 220A (XXI)*, 1966.

34 Fabbrini, F., '*Fundamental Rights in Europe,'* Oxford, Oxford University Press, Oxford, 2014, p. 8.

EU Treaties. At the same time, human rights are proclaimed in Member States' constitutions, and in the European Convention for Human Rights - to which all EU Member States are parties, and which provides authoritative guidance on the interpretation of corresponding Charter rights. This multilevel regime is characterised by a tension between national and supranational law, which can create challenges to pre-existing national institutions as well as horizontal differences between Member States about the scope and extent of protection of certain rights. In the social media context, for example, EU law offers broad general principles on the scope and definition of freedom of expression and information, but different countries conceptualise the limits of freedom of speech very differently, as Section 2.2.2 discusses. Ultimately, instead of attempting to establish uniform rules on sensitive cultural issues like the appropriate limits of free speech, EU content moderation law defers to national law to define illegal speech - which has benefits, but also raises concerns around the adequacy of fundamental rights protection, as Chapter 2 will discuss.

1.3. Technology and society: A framework for the rest of this study

Information, and the technologies through which information is produced and transmitted, have been central to the development of institutions like the rule of law and democracy and to the kind of human flourishing pursued by fundamental rights. Today, they are also associated with some of its most pressing risks. This does not mean that technologies build, or undo, democracy or the rule of law - a determinist view of technology. Nor does it mean that technologies are just tools that are employed by different people and actors in a way that depends only on their context. Rather, this study analyses how digital technology influences society through its affordances - a term from engineering which seeks to convey that different technologies make certain actions and interactions easier or harder to perform. All things being equal, things that are easier to do given particular affordances are likelier to be done, and harder things are less likely.[35] Consequently, the development, regulation and design of technologies influences how society may be more likely to develop, and is thus an important site of contestation and political action.

Social media and the internet are not the first information technologies to affect institutions like democracy and the rule of law. The wide adoption of printing in the 15th century was also a major shift, providing new affordances which enabled large-scale, long-distance and durable forms of communication. Over the following centuries, the use of print technologies contributed to shaping, transforming and challenging ideas like democracy, the rule of law and the idea of fundamental rights. As documented by Elizabeth Eisenstein, the use of the printing presses in countries where religion encouraged individual reading, such as Prussia or Scotland, enabled priests, scholars, and artisans to move beyond the limits of hand copying imposed. This happened to a lesser degree in countries where individual reading was less common, such as France or Spain. In Italy, print also allowed the cultural awakenings of the 14th and 15th centuries to be sustained and widely disseminated, while other classical revivals had necessarily been transitory and limited in scope.[36] These new social and economic practices affected some of the main cultural movements that gave rise to the modern world: the Renaissance, the Reformation, and the rise of modern science. However, as Eisenstein highlights, the print was an agent, with different effects in different social contexts, and not the sole agent of these transformations.[37]

---

[35] *See* Benkler, Y., *The Wealth of Networks*, Yale University Press, New Haven, US. p. 17; Winner L., 'Do Artifacts have politics?', *Daedalus*, Vol. 109, No. 1, 1980.

[36] Eisenstein, E., *The Printing Revolution in Early Modern Europe*, Cambridge University Press, Cambridge, U.K., p. 139.

[37] Eisenstein, E., *The Printing Revolution in Early Modern Europe*, Cambridge University Press, Cambridge, U.K.

Like social media and the internet, the printing press facilitated disruptive changes in people's access to information, undermining some established institutions while also allowing new powerful institutions to emerge. Eisenhower documents that the popularisation of the press created significant confusion: 'The same publicity system that enabled instrument makers to advertise their wares and contribute to public knowledge also encouraged an output of more sensational claims. Discoveries of philosophers' stones, the keys to all knowledge, the cures to all ills were proclaimed by self-taught and self-professed miracle workers who often proved to be more adept at press agency than at any of the older arts.'[38] The knowledge that had been before transmitted and guarded by authoritative institutions was hard to distinguish and recognise in this cacophony.

Media pluralism and the existence of an independent media as a building block of democracy can also be traced back to this transformation. In the Protestant North of Europe, the printed press loosened the Church's power over information and knowledge production, and facilitated the emergence of early print shops that were independent both from the church and from princes.[39] In the following centuries, as movements for democracy and the 'rights of man' - prefiguring what we would now call human or fundamental rights - spread across Europe, they relied on a lively and diverse ecosystem of printed newspapers and pamphlets, and on institutions such as coffee houses where individuals could not only access these information sources, but also discuss them with others.[40] However, as Yochai Benkler explains, as the print media developed, new institutions and concentrations of power emerged: 'Over the past century and a half, these early printers turned into the commercial mass media: A particular type of market-based production— concentrated, largely homogenous, and highly commercialised— that came to dominate our information environment by the end of the twentieth century.'[41]

It goes beyond the scope of this study to dive deep in how the societies of the Renaissance overcame the challenges brought about by the press and capitalised on its opportunities to build the modern world. The point is, rather, that they to a large extent succeeded. The internet, and in particular social media, also poses challenges to what are now our traditional institutions - fundamental rights, democracy and the rule of law - and requires them to adapt. Indeed, already in the early days of the internet, scholars identified that the decentralized mode of production and information dissemination that the internet enabled would challenge key incumbents of the industrial information economy like the mass media, but also top-down approaches to government and production focused on firms and rather monolithic entities. In the early 2000s, scholars argued that the internet would facilitate the emergence of a new form of information production, decentralised, socially driven, and diverse. This revolution would change how we see and interact in and with the world.[42] At the time, many saw these developments as holding great promise: the internet would be a platform for better democratic participation, a medium to foster a more critical culture, and to improve human development.[43]

It is often believed that those early internet scholars were wrong or naive but, in fact - even if they were indeed hopeful - they were mostly right, even as some of their main fears also became a reality. The internet did enable an information environment where production is more decentralised, widely accessible and diverse than ever before. Anyone with access to the internet and a smartphone can open a social media account, and reach audiences that were unimaginable before. As such, social

---

38 Eisenstein, E, *The Printing Revolution in Early Modern Europe*, Cambridge University Press, Cambridge, U.K. p. 158.

39 Eisenstein, E., *The Printing Revolution in Early Modern Europe*, Cambridge University Press, Cambridge, U.K.Press, 2005. Benkler, Y., *The Wealth of Networks*, Yale University Press, New Haven, US, 2006. p. 17, 33.

40 Habermas, J., The Structural Transformation of the Public Sphere, Polity, 1962 (trans 1989).

41 Benkler, Y., *The Wealth of Networks*, Yale University Press, New Haven, US, 2006, p. 33.

42 Benkler, Y., *The Wealth of Networks*, Yale University Press, New Haven, US, 2006, p. 34.

43 Benkler, Y., *The Wealth of Networks*, Yale University Press, New Haven, US, 2006.

media do not just present threats, but also many opportunities to strengthen and enhance democracy, the rule of law and fundamental rights. Their affordances offer new opportunities for political activism, community formation, self-expression and access to information. Regulation and policy should thus always be attentive to the risks of curtailing these benefits and restricting people's rights, for example by creating unjustified barriers to freedom of expression online. Regulatory interventions should also aim to ensure more equal and inclusive access to these benefits: for example, by preventing discriminatory treatment by social media platforms, and by tackling issues like online hate speech which can exclude minority groups from the public sphere.

However, the challenges are also significant, and they are the main focus of this study. With this decentralisation and democratisation of information production came a loss of consensus and monopoly about key narratives and facts - something that may not be entirely new, but has become more widespread and influential. Social media and the internet also created the possibility to inflict new forms of harm, and exercise new forms of control, in the digital environment.

Consequently, in recent years there has been a growing consensus in research, politics and the media that the rise of social media poses important challenges to the rule of law, fundamental rights and democracy. Hate speech, harassment and other forms of online violence and abuse are one major issue which threatens individual rights, as well as social equality and equal participation in democratic debate. At the same time, developments in the regulation of online speech pose well-recognised risks to freedom of expression and information. Online public spaces are largely governed by powerful private companies, which can suppress speech and influence public discourse in opaque and unaccountable ways; they also collaborate with and are influenced by state institutions and regulatory frameworks, creating further possibilities for unaccountable interventions in public debate. Moreover, the structural power of major platform companies creates risks for media pluralism and independence: significant power over media production and distribution now rests with these companies, in particular Google and Meta, which control most of the online advertising industry. Finally, while the evidence base is complex and uncertain, there is widespread concern about the broader, more diffuse effects that social media platforms might have on the nature of online conversations and public debate. Most prominently, they appear to have created new and very effective channels for the dissemination of disinformation, which threatens to distort political debate, exacerbate social divisions and in some cases endanger public safety.

While inevitably selective, the following chapters aim to offer a more detailed and context-specific picture of these issues by providing a more in-depth analysis of three key issues: online hate speech, disinformation, and media pluralism. These areas have been selected because they raise significant concerns for the various aspects of the rule of law discussed above: fair application of the law, fundamental rights, equal participation, and democratic debate. In addition, each of these areas is currently in a state of flux. New risks to the rule of law are emerging and developing, and changes to the EU regulatory landscape are underway. Regulatory change presents opportunities to address the problems discussed, yet some aspects of regulation create new threats to fundamental rights, the rule of law and democracy.

Accordingly, the following chapters aim to provide a detailed overview of current policy problems in each of the three areas discussed, as well as the currently-applicable regulatory regime and upcoming changes. They identify key issues and threats relating to the rule of law, fundamental rights and democracy, with a particular focus on ensuring equal and inclusive participation in social media and online public debate. Finally, each chapter proposes regulatory reforms and other policy interventions which could address the problems identified.

2. THE EUROPEAN LEGISLATIVE AND REGULATORY BACKGROUND

This chapter provides an overview of the relevant legal instruments relating to platform governance and online content moderation, and flags the most important provisions for issues around social media content governance, the rule of law and fundamental rights. The following sections will then build on this outline with more detailed analysis of this legal regime's implications for online hate speech, disinformation, and democratic debate and media pluralism, along with policy recommendations.

Section 2.2 sets out the legal framework governing content moderation in the EU. Sections 2.2.3 and 2.2.4 first describe the generally-applicable framework for intermediary liability (platforms' liability for hosting illegal content) and due diligence (other obligations relating to how platforms run their moderation systems) set out in the ECD and DSA. Sections 2.3.3 and 2.3.4 then briefly describe area-specific legislation or soft law measures governing five particular types of platform or content: the 2019 Copyright Directive (CD), the 2021 Terrorist Content Regulation (TCR), the 2018 updated Audiovisual Media Services Directive (AVMSD), the 2022 updated Code of Practice on Disinformation (CoP), and the 2016 Code of Conduct on Hate Speech (CoC). Finally, as the rule of law issues discussed in this study are not only affected by content moderation, but also by many other social, institutional and technical aspects of contemporary social media – including platform design, business and funding models – Section 2.3 briefly highlights aspects of the DSA, the CoP on Disinformation and the proposed Political Advertising Regulation which regulate these aspects.

2.1. Introduction

Social media platforms create both tensions and synergies between different European values. They raise questions about the development of the internal market, the protection and guarantee of fundamental rights and freedoms (particularly freedom of speech and information, the freedom to conduct a business, the right to non-discrimination and the attainment of high levels of consumer protection), and about Europe's ambitions to maintain and strengthen economic growth, investment and innovation.[44] Consequently, EU-level regulation of social media platforms has evolved into a complex framework involving numerous overlapping legal and soft-law instruments. For social media companies these include corporate law frameworks from their country of origin (such as US corporate law, in the case of Meta and many other leading platforms), self-regulatory frameworks like the Global Network Initiative principles,[45] as well as European corporate and economic law and sector-specific regulation.[46] As explained in the introduction, this study's main focus is the EU regulatory framework governing content moderation and associated challenges to democracy, the rule of law and fundamental rights.

Content moderation, as defined in Article 3(t) DSA, refers to measures that platforms take to enforce the law or their own in-house policies regarding what content they will host and how they address harmful content. This includes, for example, deleting content or user accounts, demonetising content (meaning it does not run with ads), or demoting it (showing it less prominently in algorithmic

[44] See 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, 15.12.2020, numerals, 1, 3.

[45] Global Network Initiative, 'The GNI Principles', Global Network Initiative, n.d., available at: https://globalnetworkinitiative.org/gni-principles/ .

[46] See Gorwa, R., ' What Is Platform Governance?', *Information, Communication & Society, Vol. 22, No. 6*, 2019, pp. 854–871.

recommendations).[47] The baseline regime on content moderation in the EU was established by the 2000 E-Commerce Directive (ECD), which established key principles that remain in place today. Under EU law, intermediary services, including social media, are exempt from liability for hosting content that is illegal, so long as they do not participate in its production and remove illegal content as soon as they are made aware of it.[48] However, not all content that is legal is allowed on social media, as platforms can set their own standards as to what content they will permit, and within this framework, European policymakers have encouraged them to participate in self- and co-regulatory measures to address harmful content.[49] This framework has enabled platform companies to shape online experience and govern user content and communications through their own Terms of service and Community guidelines, while at the same time being subject to regulatory, economic and reputational pressures from governments and other stakeholders.[50] This ecosystem of interacting actors exercising different forms of influence over online content is known as platform governance.[51]

The DSA entered into force on 16 November 2022 and is now the most important legal instrument in the EU setting out the generally-applicable ground rules for the regulation of online content. It maintains the key principles of the ECD, but significantly updates and expands it with new obligations which aim to address the dissemination of illegal content on platforms while protecting users' fundamental rights. The DSA will be directly applicable across the EU and will apply fifteen months or from 1 January 2024, whichever comes later, after entry into force.[52] Although billed as a 'first comprehensive rulebook for the online platforms that we all depend on',[53] far from comprehensively regulating online content, the DSA will function in tandem with a number of area-specific regulations which detail further obligations for platforms regarding specific types of content. A high-level overview of these various instruments is provided in Table 1. The DSA also envisages a number of co-regulatory and industry codes and best practices which will further specify its obligations,[54] creating a complex legal framework.

---

47 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, 15.12.2020, art 3(t).

48 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, 15.12.2020, art 6. See also Wilman, F., 'The EU's System of Knowledge-Based Liability for Hosting Service Providers in Respect of Illegal User Content – between the e-Commerce Directive and the DSA', Vol. 12, 2021, pp. 317–341.

49 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"), 17 July 2000; see also De Streel, A. et al. Online Platforms' Moderation of Illegal Content Online: Law, Practices and Options for Reform, Policy Department for Economic, Scientific and Quality of Life Policies, June 2020.

50 Gorwa, R., 'Who Are the Stakeholders in Platform Governance?', Yale Information Society Project, Platform Governance Terminologies Essay Series, October 2022.

51 See Gorwa, R., ' What Is Platform Governance?', *Information, Communication & Society, Vol. 22, No. 6,* 2019, pp. 854–871.

52 See European Commission, The Digital Services Act package, available at: https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package .

53 'Commission Welcomes European Parliament´s Adoption of Digital Services Package | Shaping Europe's Digital Future', available at: https://digital-strategy.ec.europa.eu/en/news/commission-welcomes-european-parliaments-adoption-digital-services-package.

54 For a comprehensive list see Jaursch, J., 'Overview of DSA Delegated Acts, Reports and Codes of Conduct', Stiftung Neue Verantwortung, September 12, 2022, available at: https://www.stiftung-nv.de/en/publication/overview-dsa-delegated-acts-reports-and-codes-conduct.

Table 1: Scope of relevant EU regulatory measures

Context	Legal obligations	Legal obligations
All intermediaries and all illegal content	ECD/DSA notice and takedown regime	
All content moderation by online platforms	DSA procedural protections and transparency rules	
Very large online platforms, all content and technical/organisational decisions	DSA systemic risk provisions and additional transparency rules	Global Network Initiative
Copyright-infringing content	Copyright Directive	
Terrorist content	Terrorist Content Regulation	Global Internet Forum to Counter Terrorism
Hate speech	ECD/DSA notice and takedown regime (for illegal hate speech)	Code of Conduct on Hate Speech
Disinformation	ECD/DSA notice and takedown regime (for illegal speech)	Code of Practice on Disinformation
Video-sharing platforms, as regards illegal speech, hate speech and speech harmful to minors	Audiovisual Media Services Directive	

Source: Authors' own elaboration

2.2. Content moderation

2.2.1. Intermediary liability in the E-Commerce Directive and the DSA

Intermediary liability is a broad term describing the legal liability of services which host, transmit or distribute information for content generated by their users. In principle, since social media and other online platforms host user-generated content on their servers and publish it via their websites and apps, they could be primarily liable for distributing any content which is illegal. It is widely recognised that leaving this default situation in place in the social media context would raise serious concerns for fundamental rights, such as freedom of expression and information and privacy, as the risk of liability would incentivise platforms to closely surveil and censor users to prevent any potentially-illegal activity.[55] It would also have very significantly hampered the growth of today's platform economy, as such liability risks and surveillance systems would have been unfeasibly expensive for platform

[55] Wilman, F., 'The EU's System of Knowledge-Based Liability for Hosting Service Providers in Respect of Illegal User Content – between the e-Commerce Directive and the DSA', Vol. 12, 2021, pp. 317–341.

companies.[56] Thus, in 2000, the ECD established broad intermediary liability immunities, under which internet intermediaries are only liable for user-generated content under certain specified conditions. Importantly, since liability rules themselves are not entirely harmonised, the types of content that can attract liability and the precise contours of that liability depend on national law.

Articles 12-14 ECD, now replaced by Articles 4-6 DSA, establish intermediary liability immunity for three types of service: mere conduits, caching and hosting. Social media primarily involves hosting user content. Article 6 DSA establishes that platforms are not liable for hosting illegal information, on condition that they either do not know about the illegal content, or remove it promptly on becoming aware. This effectively creates a notice-and-takedown regime, in which platforms are not obliged to proactively look for illegal content, but if specific illegal content is notified to them by a third party - which could be any individual or organisation - they must act expeditiously to remove it or disable it.[57] Additionally, Article 8 prohibits the imposition of general monitoring obligations, meaning that platforms cannot be obliged 'actively to seek facts or circumstances indicating illegal activity'.

Building on this established liability regime, the DSA further specifies how notice-and-takedown systems should operate in Articles 9 and 16. Article 16 requires all hosting services to establish easily-accessible, user-friendly mechanisms for users to report illegal content. Sufficiently substantiated reports create knowledge for the purposes of Article 6, which will generally mean that the platform has to remove the content to avoid potential liability. Article 9 further authorises national judicial and administrative authorities to order intermediaries to remove illegal content.

'Illegal content' is defined in Article 3(h) DSA as any content which is illegal under EU or national law, either in itself (e.g. hate speech or defamation) or in relation to illegal products or services (e.g. offering to sell contraband goods). As the following sections describe in more detail, types of content which are illegal under EU law include (i) child sexual abuse material, (ii) racist and xenophobic hate speech; (iii) terrorist content and (iv) intellectual property infringement. Beyond this, Member States have their own national laws defining various types of illegal speech.

2.2.2. Illegal speech under national law

Member States' national civil and criminal laws remain central in regulating online speech because the ECD/DSA intermediary liability framework primarily defers to national law to define illegal speech. In effect, the definition of 'illegal content' differs in each member state. Thus, the notice-and-takedown framework requires platforms to remove reported content in a given Member State if it is illegal under that state's national law. Additionally, the ECD/DSA framework leaves it to national courts to issue injunctions against social media companies which can require them to remove illegal content.

This deferral to national law to define illegal content has many benefits: it allows the particular social and cultural conditions of each Member State to be taken into account, and avoids regulating culturally and politically sensitive issues which involve complex rights-balancing exercises in an undifferentiated way across the EU. However, it also raises two major issues for fundamental rights and the rule of law.

First, platforms can and sometimes do respond to notice-and-takedown obligations by 'geoblocking' content: making it unavailable in a specific country, while continuing to make it available elsewhere.

---

[56] Savin, A., EU Internet Law, Third edition., Elgar European Law, Edward Elgar Publishing, Cheltenham, UK ; Northampton, MA, USA, 2020.

[57] Kuczerawy, A., 'From 'Notice and Take Down' to 'Notice and Stay Down': Risks and Safeguards for Freedom of Expression', The Oxford Handbook of Intermediary Liability Online, 2019. See Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (DSA) and Amending Directive 2000/31/EC, COM(2020) 825 Final, 15.12.2020. Art 5 and 15.

However, they generally prefer to operate consistent rules across all markets, as this is simpler and minimises costs.[58] Thus, content which is illegal in one Member State may often be removed across the EU (and globally). This may create a 'levelling down' effect, where platforms regulate speech in line with whatever Member State's laws are most restrictive – including in countries where different social and political contexts mean such restrictions are not appropriate or proportionate.

Second, many Member States have speech laws in place which, even in their original contexts, raise serious fundamental rights concerns and cannot be regarded as proportionate. For example, several Member States have very broad prohibitions on sharing false news, which are difficult to reconcile with European and international human rights jurisprudence establishing that false information is also protected by freedom of expression.[59] Germany broadly criminalises 'insults', which has been used to investigate trivial insults of politicians on social media, threatening to seriously chill political dissent and debate.[60] Hungary in 2021 banned sharing information with under-18s which could encourage homosexuality or gender transition, violating LGBTQ+ people's freedom of expression and young people's rights to access information which may be vital for their mental health and wellbeing.[61]

Additionally, Article 9 DSA empowers Member States to actively require platforms to remove all such information in their own countries, which also creates serious risks that content could be removed worldwide under problematic laws like those mentioned above. Recital 25 DSA confirms that – as under the ECD – the intermediary liability immunity provisions preclude criminal and monetary liability, except where content has been specifically notified to platforms, but Member State courts can still issue injunctions requiring platforms to remove content (with Article 9 setting out further details as to how platforms must respond to judicial orders).

In recent years, the European Court of Justice (ECJ) has permitted national courts to use injunctions to impose increasingly strict obligations on platforms, in particular regarding proactive monitoring and filtering to prevent reuploads of illegal content. In its earlier *SABAM* decisions, the ECJ had held that platforms could not be required to check all user uploads for copyright-infringing content, as Article 15 ECD prohibits general monitoring obligations (maintained in Article 8 DSA), and such obligations would violate platforms' fundamental rights to conduct a business and users' freedom of expression and privacy rights.[62] This changed in 2019, with the ECJ's ruling in *Glawischnig-Piesczek*, an Austrian case involving an anonymous Facebook user who shared an article and a defamatory comment against politician Eva Glawischnig-Pieszek. Ms. Glawischnig-Piesczek obtained an injunction against Facebook (now Meta) which required them not only to remove the post that had been found to be defamatory, but also to continue removing identical or equivalent content on an ongoing basis. The ECJ found that such an injunction did not violate the general monitoring prohibition, provided that the content to be monitored is defined in specific terms and can be identified automatically without manual

---

58 Heldt, A., 'Reading between the Lines and the Numbers: An Analysis of the First NetzDG Reports', *Internet Policy Review, Vol. 8, No. 2*, June 12, 2019.

59 Ó Fathaigh et al., 'The Perils of Legally Defining Disinformation', *Internet Policy Review, Vol. 10, No. 4,* November 4, 2021.

60 Drügemöller, L., 'Pimmel-Gate in Hamburg: Unterhalb der Schwelle', *Die Tageszeitung*, August 8, 2022, sec. TAZ, Nord.

61 Rankin, J., 'Hungary Passes Law Banning LGBT Content in Schools or Kids' TV', *The Guardian*, June 15, 2021, sec. World news.

62 Judgment of the Court (Third Chamber) of 24 November 2011, *Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM)*, Case C-70/10; Judgment of the Court (Third Chamber) of 24 November 2011, *Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM) v. Netlog NV*, Case C-360/10.

intervention.[63] The Court also held that the ECD did not preclude Member States from ordering such removal worldwide, and left it to the Member States to determine the geographic scope of the restriction within the framework of the relevant national and international laws.[64] More recently, in *Poland v Parliament and Council*, the ECJ reaffirmed and clarified the principle established in *Glawischnig-Piesczek,* holding that legal obligations to monitor all content uploaded on a platform for illegal content do not constitute impermissible general monitoring obligations, provided that the content which platforms must search for is specifically defined and can effectively be identified automatically, without requiring manual assessments of content.[65]

This reinterpretation raises serious fundamental rights concerns, as automated filtering tools are inevitably imprecise and will block legal content, as well as requiring further surveillance and analysis of user data.[66] Similarly, the digital rights NGO Article 19 pointed out that the *Glawischnig-Piesczek* ruling's tolerance for extraterritoriality set a dangerous precedent where courts in one country can control what internet users in another country see, which could be open to abuse.[67] It has also laid the groundwork for the imposition of further obligations - or strong incentives - for platforms to expand automated filtering in the Copyright Directive (CD) and Terrorist Content Regulation (TCR). These provisions have concerning implications for fundamental rights, discussed in Section 2.2.4.

2.2.3. Due diligence obligations in content moderation

Under the intermediary liability framework, platforms must moderate illegal content if they receive a sufficiently substantiated notice. Building on this, Chapter II of the DSA establishes a tiered set of due diligence obligations as to how their moderation systems must operate, aiming to ensure the removal of illegal content, while also preventing excessive censorship.

This includes obligations for all providers of intermediary services (Articles 11-15), obligations for providers of hosting services (Articles 16-18), and obligations for providers of online platforms (Articles 19-28).[68] Generally, social media platforms fall into all three of these categories, so must comply with all these obligations. In addition, Articles 33-43 create further obligations for providers of 'very large online platforms' (defined as platforms whose average number of monthly users in the EU is 45 million or more, i.e. 10% of the EU population) and 'very large online search engines' (Articles 33-43). The most relevant obligations in each tier are briefly described below.

a. Obligations for all intermediaries

Article 14 requires platforms to publish their contractual content policies 'in clear, plain, intelligible, user-friendly and unambiguous language'. Article 14(4) further requires them to 'act in a diligent, objective and proportionate manner in applying and enforcing the restrictions…with due regard to the rights and legitimate interests of all parties involved'. Platforms must thus consider fundamental

---

63 Jütte, B.J., and G. Priora, 'On the Necessity of Filtering Online Content and Its Limitations', Kluwer Copyright Blog, July 20, 2021, available at:https://copyrightblog.kluweriplaw.com/2021/07/20/on-the-necessity-of-filtering-online-content-and-its-limitations-ag-saugmandsgaard-oe-outlines-the-borders-of-article-17-cdsm-directive/.

64 Jütte, B.J., and G. Priora, 'On the Necessity of Filtering Online Content and Its Limitations', Kluwer Copyright Blog, July 20, 2021, available at: https://copyrightblog.kluweriplaw.com/2021/07/20/on-the-necessity-of-filtering-online-content-and-its-limitations-ag-saugmandsgaard-oe-outlines-the-borders-of-article-17-cdsm-directive/.

65 Republic of Poland v European Parliament and Council of the European Union, ECJ 2022.

66 Keller, D., 'Facebook Filters, Fundamental Rights, and the ECJ's Glawischnig-Piesczek Ruling', *GRUR International, Vol. 69, No. 6*, June 1, 2020, pp. 616–623.; Chowdhury, N., Automated Content Moderation: A Primer, Cyber Policy Center, Stanford University, March 19, 2022.

67 Article 19, 'CJEU Judgment in Facebook Ireland Case Is Threat to Online Free Speech', Article 19, October 3, 2019, available at: https://www.article19.org/resources/CJEU-judgment-in-facebook-ireland-case-is-threat-to-online-free-speech/.

68 Articles 26 to 32 include consumer protection provisions related to online platforms allowing consumers to conclude distance contracts with traders, which are outside the scope of this study.

rights when formulating and enforcing content policies. However, the abstract and indeterminate nature of this obligation means it may have limited impact in practice: what rights require in particular situations is generally open to interpretation, there are typically multiple competing rights involved, and platforms only have to 'have regard to' rights, not strictly respect them. Article 15 requires platforms to publish yearly transparency reports setting out how their moderation systems work (e.g. staffing, training, use of automation) and how much content is removed on different grounds.

b. Obligations for hosting providers

Article 16 requires hosting providers to operate a notice-and-takedown system where users can report content as illegal or incompatible with the platform's policies. Article 17 requires them to 'provide a clear and specific statement of reasons' to users whose content is restricted, explaining what measures have been taken and why (this excludes content removed by order of a public authority under Article 9, but Article 9(5) requires a similar statement of reasons in these cases). Restrictions of content in this sense can include deleting the content, but also geoblocking, deleting the users' account, suspending revenue-sharing programmes, and demoting or 'shadowbanning' (where content is not recommended to other users, or made more difficult to access).

c. Obligations for online platforms

Article 20 further requires platforms to institute an easily-accessible and user-friendly complaints-handling system in which users whose content is removed, or who have reported content which was not removed, can appeal the decision. Platforms are free to determine the procedures used,[69] but must review complaints 'in a timely, non-discriminatory, diligent and non-arbitrary manner' and not solely automatically. If decisions are shown to be unfounded (i.e. not justified either by law or by the platform's terms and conditions) the platform must reverse them (Article 20(4)). Article 21 additionally empowers users to further appeal such decisions to certified out-of-court dispute settlement institutions.

d. Exclusion for micro and small enterprises

Certain obligations, such as establishing an internal complaint-handling system and creating dispute resolution mechanisms (Articles 20-21) exclude micro or small enterprises, defined as those with under 250 employees and either annual turnover under €50 million or an annual balance sheet total under €43 million.[70] This is important, as compliance costs will generally be significant and highly burdensome for smaller platforms. Requiring them to implement costly compliance systems will tend to strengthen the market dominance of today's leading platforms – with negative implications for media pluralism and freedom of expression and information. However, the employee and revenue thresholds for micro and small enterprises are arguably too narrow to address these concerns: experts have suggested that medium-sized and rapidly-scaling platforms will still face significant barriers.[71]

---

[69] Ortolani, P., 'If You Build It, They Will Come: The DSA's "Procedure Before Substance" Approach', Verfassungsblog, November 7, 2022, available at: https://verfassungsblog.de/dsa-build-it/.

[70] European Commission (EC) (L 124/36) Recommendation 2003/361/EC, concerning the definition of micro, small and medium-sized enterprises (notified under document number C(2003) 1422), 6 May 2003, annex I, art 2.

[71] Kellner, D., 'The DSA's Industrial Model for Content Moderation', *Verfassungsblog*, February 24, 2022, available at: https://verfassungsblog.de/dsa-industrial-model/ ; Kellner, D., 'The EU's new Digital Services Act and the Rest of the World', *Verfassungsblog*, November 7, 2022. https://verfassungsblog.de/dsa-rest-of-world/.

The DSA's generally-applicable provisions regulating content moderation by online platforms have been summed up as centring 'procedure over substance',[72] in the sense that they regulate moderation procedures (applying clear policies, notifying users of decisions, etc.), but not the substantive rules that platforms apply. Platforms are free to set their own contractual rules about what speech they allow, as long as these are transparent and decisions are subject to appeal. This creates new avenues for aggrieved users to challenge arbitrary or biased moderation decisions, which are likely to be useful in particular to people who use social media in a (semi-)professional capacity and are thus informed and motivated to challenge decisions. However, evidence from the copyright context suggests that the majority of users are unlikely to utilise procedural protections such as appeals.[73] Given inequalities in time, resources and digital literacy, they may particularly fail to protect the freedom of expression of marginalised groups.[74]

Regulating 'procedure over substance' also means these provisions do not address some important concerns around fundamental rights and media pluralism. Platforms' substantive policies can still significantly restrict freedom of expression and disproportionately affect marginalised social groups, even if they are applied in a procedurally fair and consistent way: examples include policies banning pseudonymous or anonymous accounts, which disproportionately censor vulnerable users who need anonymity[75] and demonetisation policies which remove adverts from content viewed negatively by advertisers, disincentivising or suppressing such content even though it may be neither illegal nor harmful.[76] However, in the case of 'very large online platforms' with over 45 million EU users, the DSA establishes additional, more substantive obligations, detailed below.

e. Obligations for very large online platforms

Articles 34 and 35 require very large online platforms to conduct regular risk assessments, and take measures to mitigate various systemic risks. Article 34 provides that these include '(a) the dissemination of illegal content through their services; (b) any negative effects for the exercise of fundamental rights such as private and family life, freedom of expression and information, the prohibition of discrimination and the rights of the child (...); [and] (c) any intentional manipulation of their service (...) with an actual or foreseeable negative effect on the protection of health, minors, civic discourse or actual or foreseeable effects related to electoral processes and public security.'[77] Article 35 requires platforms to take reasonable, proportionate and effective measures to mitigate the risks identified under Article 34. This could include for example adapting their content moderation or recommender systems, limiting the display of advertisements, and reinforcing internal processes or supervision.[78]

---

[72] Ortolani, P., 'If You Build It, They Will Come: The DSA's "Procedure Before Substance" Approach', Verfassungsblog, November 7, 2022, available at: https://verfassungsblog.de/dsa-build-it/.

[73] Urban, J.M. et al., 'Notice and Takedown in Everyday Practice', UC Berkeley Public Law, Research Paper No. 2755628, March 22, 2017, pp. 1–182.

[74] Griffin, R., 'Rethinking Rights in Social Media Governance: Human Rights, Ideology and Inequality', *Forthcoming in European Law Open, Vol. 2, No. 1*, March 23, 2022.

[75] Griffin, R., 'Public and Private Power in Social Media Governance: Multistakeholderism, the Rule of Law and Democratic Accountability', SSRN Electronic Journal, 2022.

[76] Kumar, S., 'The Algorithmic Dance: YouTube's Adpocalypse and the Gatekeeping of Cultural Content on Digital Platforms', *Internet Policy Review, Vol. 8, No. 2,* June 30, 2019.

[77] 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, art. 26.

[78] 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, art. 27.1.

Risk assessments and mitigation measures must be independently audited every year under Article 37, and will generally be overseen by the Commission.[79] Article 40 deals with the provision of data, documentation and information to regulators by very large online platforms, in order to facilitate oversight and enforcement of their substantive obligations. In addition, Article 40(4) provides that platforms must provide internal data on request to researchers vetted by national regulators, for the purpose of investigating systemic risks.

In principle, therefore, very large platforms will no longer be able to set whatever content policies they like, but must consider more holistically whether these policies and the systems implementing them create risks to fundamental rights, democratic debate and media pluralism, or fail to adequately protect against such risks. This could include, for example, changing policies which disproportionately suppress speech from marginalised groups; avoiding the use of biased automated moderation systems; or redesigning recommender systems and interfaces to discourage the dissemination of disinformation, hate speech or other harmful content.

However, since platforms are themselves responsible for deciding how to identify and conceptualise risks, and what to do in response, it remains uncertain whether they will take adequate action on all the rule of law issues detailed in this study – in particular where doing so would create significant costs. Making these provisions an effective safeguard for the rule of law and fundamental rights will require active oversight by the Commission, with clear policy goals, such as ensuring equal treatment for marginalised groups. Independent research and public scrutiny will also play a key role in identifying and understanding systemic risks and holding platforms accountable for how they address them. The following sections and policy recommendations provide more concrete suggestions as to how these goals could be achieved.

2.2.4. Area-specific regulation

a. The 2019 Copyright Directive

The 2019 Copyright Directive (CD)[80] is a complex regulation making various reforms to EU copyright law.[81] While it cannot be fully reviewed here, two provisions are particularly relevant to the rule of law, fundamental rights and democracy in the context of social media: Article 15, which creates new neighbouring rights for press publishers, and Article 17, which creates a new intermediary liability regime for copyright material.

Article 15 must be understood in the context of the financial struggles facing the news industry in general (discussed in detail in Chapter 5) and the long-running conflict between large news publishing corporations and online platforms.[82] Publishers have argued that social media and search engines unfairly use their content to drive user engagement (for example, by linking to and previewing articles) without compensating them, while platforms have argued that publishers benefit from the traffic this

[79] 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, 15.12.2020, art. 27.

[80] Directive (EU) 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, available at: https://eur-lex.europa.eu/eli/dir/2019/790/oj.

[81] Dusollier, S., 'The 2019 Directive on Copyright in the Digital Single Market: Some Progress, a Few Bad Choices, and an Overall Failed Ambition', *Common Market Law Review, Vol. 57, No. Issue 4*, August 1, 2020, pp. 979–1030.

[82] Papaevangelou, C., 'Funding Intermediaries: Google and Facebook's Strategy to Capture Journalism', Digital Journalism, January 13, 2023, pp. 1–22.

drives to their sites.[83] In this context, Article 15 creates a new neighbouring right for press publishers (not journalists) to control the reproduction and making available of their content for two years after publication, except for individual words/short quotes and hyperlinks.[84] In France, the first Member State to implement the CD, the competition authority has made it clear that platforms cannot simply decide they do not need to reproduce excerpts of news content, but are required to do so and to compensate publishers for it, issuing multiple fines against Google for abuse of dominance when it was deemed not to be negotiating in good faith.[85]

Article 15 is thus essentially an attempt to shift the balance of bargaining power between publishers and platforms,[86] effecting a transfer of revenue to the former in order to strengthen the ailing news industry. While evidence from Australia, which has implemented a similar reform, suggests this can be quite effective,[87] evidence from France also suggests that licensing negotiations favour bigger and better-resourced news publishers and associations, and may thus have the side effect of weakening media pluralism.[88]

Article 17, on the other hand, creates an exception from the generally-applicable intermediary liability immunities for online platforms which share copyright content with the public (including social media). Such platforms are now primarily liable for infringing material if they do not make best efforts to obtain a licence from the rightsholder or, in the absence of a licence, make best efforts to remove copyright works notified to them by rightsholders and prevent future uploads. The latter obligation effectively requires automated filtering of all user uploads to identify and block the notified copyright works. Again, this reform must be understood in context as a deliberate intervention to shift the balance of bargaining power between commercial actors. A detailed analysis by Annemarie Bridy shows how media industries successfully pushed for Article 17 as a way to force YouTube to pay more for licensing music and offer its existing Content ID filtering system for free, and other platforms to offer similar filtering software.[89]

Article 17 was heavily criticised on fundamental rights grounds, due to the evident risk of 'overblocking': no filtering software exists which can reliably distinguish between copyright-infringing and non-infringing content, in particular content which uses copyrighted material legally, under an

---

83 Radsch, C.C., 'Frenemies: Global Approaches to Rebalance the Big Tech v Journalism Relationship', Brookings, August 29, 2022, available at: https://www.brookings.edu/blog/techtank/2022/08/29/frenemies-global-approaches-to-rebalance-the-big-tech-v-journalism-relationship/.

84 Dusollier, S., 'The 2019 Directive on Copyright in the Digital Single Market: Some Progress, a Few Bad Choices, and an Overall Failed Ambition', *Common Market Law Review, Vol. 57, No. Issue 4*, August 1, 2020, pp. 979–1030.

85 Autorité de la concurrence, ' Decision 20-MC-01 of April 09, 2020 on requests for interim measures by the Syndicat des éditeurs de la presse magazine, the Alliance de la presse d'information générale and others and Agence France-Presse', Autorité de la concurrence, 9 April 2020, available at: https://www.autoritedelaconcurrence.fr/en/decision/requests-interim-measures-syndicat-des-editeurs-de-la-presse-magazine-alliance-de-la; Paevangelou, C., and N. Smyrnaios, 'Regulating Dependency: The Political Stakes of Online Platforms' Deals with French Publishers', HAL, August 2022.

86 Dusollier, S., 'The 2019 Directive on Copyright in the Digital Single Market: Some Progress, a Few Bad Choices, and an Overall Failed Ambition', *Common Market Law Review, Vol. 57, No. Issue 4*, August 1, 2020, pp. 979–1030.

87 Stoller, M., 'Should We Save Newspapers from Google?', *BIG by Matt Stoller*, September 8, 2022, available at: https://mattstoller.substack.com/p/should-we-save-newspapers-from-google.

88 Papaevangelou, C., and N. Smyrnaios, 'Regulating Dependency: The Political Stakes of Online Platforms' Deals with French Publishers', HAL, August 2022.

89 Bridy, A., 'The Price of Closing the Value Gap: How the Music Industry Hacked EU Copyright Reform', *Vanderbilt Journal of Entertainment & Technology Law, Vol. 22, No. 2*, 2020, pp. 323–358.

exception such as quotation.[90] In a 2022 judicial review, the ECJ held that Article 17 is only compatible with fundamental rights if interpreted narrowly, such that only content which is clearly unlawful and can reliably be identified through automated means can be automatically blocked.[91] Responsibility for implementing legal safeguards which limit filtering in this way remains with Member States, and exactly what restrictions on filtering they will implement largely remains unclear.[92]

b. The 2021 Terrorist Content Regulation

The TCR[93] does not reform intermediary liability rules, but creates new, parallel due diligence obligations regarding 'terrorist content'. Under Article 3, where law enforcement authorities issue a removal order relating to such content, platforms must remove it within an hour. Article 5 further provides that platforms requiring more than one such order per year (likely to include every major platform) can be designated by authorities as 'exposed to terrorist content', and thereby required to take 'specific measures' to address it. These measures are in the first instance up to the platform, but whether they are adequate will ultimately be determined by national authorities, who can issue decisions requiring further action (Article 5(6)). Article 5(2) provides that such measures could include, for example, increased moderation staff and technical resources and enhanced mechanisms for user reporting; Article 5(3) requires that any such measures must be effective, proportionate and applied with consideration of users' fundamental rights.

Notwithstanding this provision, these requirements also raise important fundamental rights concerns and are likely to lead to significant censorship of legal content. 'Terrorist content' is defined in Article 2(7) as content with any one of various effects (e.g. soliciting, incitement, glorification) in relation to one of the terrorist offences defined in the 2017 Directive on Combating Terrorism. As well as being complex, this definition has been criticised for being overly vague and broad (in particular because it lacks an intention requirement); it creates significant uncertainty about what will be deemed terrorist material, creating obvious risks of arbitrary and biased application, and could be used to target journalistic content or non-violent political advocacy.[94]

Particularly given the complex legal analysis required to apply such definitions, platforms receiving one-hour removal orders under Article 3 cannot be expected to carefully consider whether removing the content is justified before censoring it. It is likely that platforms will respond to their Article 5 obligations by expanding automated moderation, which is already widely used for terrorist content, partly due to pressure from European governments.[95] Given the unreliability of automated content

---

90 Dusollier, S., 'The 2019 Directive on Copyright in the Digital Single Market: Some Progress, a Few Bad Choices, and an Overall Failed Ambition', *Common Market Law Review, Vol. 57, No. Issue 4*, August 1, 2020, pp. 979–1030. ; Husovec, M., '(Ir)Responsible Legislature? Speech Risks under the EU's Rules on Delegated Digital Enforcement', September 17, 2021.

91 Republic of Poland v European Parliament and Council of the European Union, ECJ 2022.

92 Reda, F., and P. Keller, 'ECJ Upholds Article 17, but Not in the Form (Most) Member States Imagined', *Kluwer Copyright Blog*, April 28, 2022, available at: https://copyrightblog.kluweriplaw.com/2022/04/28/cjeu-upholds-article-17-but-not-in-the-form-most-member-states-imagined/.

93 Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online, available at: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex:32021R0784.

94 Van Hoboken, J., 'The Proposed EU Terrorism Content Regulation: Analysis and Recommendations with Respect to Freedom of Expression Implications', *Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression*, May 3, 2019, available at: https://www.ivir.nl/publicaties/download/TERREG_FoE-ANALYSIS.pdf.

95 Douek, E., 'The Rise of Content Cartels', Essays and Scholarships, Knight First Amendment Institute at Columbia University, February 11, 2020, available at: http://knightcolumbia.org/content/the-rise-of-content-cartels; Borelli, M., 'Social Media Corporations as Actors of Counter-Terrorism', *New Media & Society*, August 8, 2021.

recognition tools and prevalence of algorithmic bias,[96] expanding automated moderation will inevitably lead to arbitrary and over-broad censorship – which is likely to disproportionately target Muslims, Arabic speakers and other minorities who are stigmatised and stereotypically associated with terrorism in European society.

c. The updated Audiovisual Media Services Directive

The AVMSD[97], passed in 2010 and significantly amended in 2018, regulates a range of audiovisual media including traditional broadcasters. However, the 2018 updated version includes new provisions on 'video-sharing platforms', a category which includes many social media. This is defined in Article 1(a) as a service provided through electronic communications networks, where the main purpose or an essential functionality of the service or a dissociable section thereof is to disseminate audiovisual programmes or user-generated videos to the general public, and where the service provider does not have editorial responsibility for the videos but does control their organisation, including through automated means. As well as video-centric platforms like YouTube and TikTok, this definition includes platforms which have a dissociable section dedicated to user-generated videos, notably Instagram's Instagram Video and Reels and Facebook's Watch section, meaning these platforms must comply with the AVMSD in regard to those products.

With regard to online content regulation, the key obligations for video-sharing platforms are set out in Article 28b. Member States must ensure that video-sharing platforms take 'appropriate, practicable and proportionate' measures to protect minors using their services from content which could be mentally, physically or morally harmful, and to protect all users from content which is illegal or which incites hatred or violence against a group protected by Article 21 of the Charter on non-discrimination. These could include various technical or organisational changes, including allowing users to report harmful content for removal, but should not lead to general monitoring of all user content.

2.2.5. Self- and co-regulatory initiatives

a. Self-regulation

In addition to the multi-layered regulatory framework detailed above, platforms rely on their contractual terms of service and content policies (often termed community standards or guidelines) to regulate user-generated content beyond what is legally required, as well as further specifying how they moderate illegal content. These terms of service and policies do not necessarily reflect a specific legal system, but aim to prevent harm, create welcoming online environments for users, and serve commercial goals such as attracting advertisers.[98] For example, Facebook's Community Standards prohibit content that promotes or celebrates suicide and self-harm.[99]. Strictly speaking, both of these

---

[96] Llansó, E. et al., 'Artificial Intelligence, Content Moderation, and Freedom of Expression', *Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression*, February 26, 2020; Chowdhury N., 'Automated Content Moderation: A Primer', Program on Platform Regulation, Stanford, 2022.

[97] Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive), available at: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32010L0013.

[98] De Streel, A. et al., 'Online Platforms' Moderation of Illegal Content Online: Law, Practices and Options for Reform', Policy Department for Economic, Scientific and Quality of Life Policies, June 2020. ; Gillespie, T., *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*, Yale University Press, New Haven [Connecticut], 2018.

[99] Meta, 'Suicide and Self-Injury', Facebook Transparency Center, n.d., available at: https://transparency.fb.com/en-gb/policies/community-standards/suicide-self-injury/.

are legal forms of expression. As such, platforms are often stricter in identifying what kind of content is allowed than many national laws.[100]

The influence that platforms can exercise over online communications and media through their terms of service, their design, their algorithms and other technical structures has led internet scholars to stress that platforms are political actors who make important decisions about the networked infrastructure of democracy.[101] Platforms govern, albeit in a way that is influenced by other actors and directly informed by supranational multilayered levels of governance.[102] This is not necessarily a bad thing. Policymakers and academics generally recognise that platforms' self-governance has several advantages.[103] Large platforms can respond to developing situations quickly, with targeted technical interventions that could not be achieved through state regulation.[104] Additionally, keeping decisions about freedom of expression in their hands can mitigate concerns over government censorship and excessive state control of the media.[105]

That platforms do react to certain forms of public pressure has become evident at least since the Cambridge Analytica scandal in 2016, when leading social media companies have attracted increasing public criticism and scrutiny.[106] Since then, they have implemented a range of transparency and 'trust and safety' policies as a way to regain public trust.[107] For example, all major platforms now publish various transparency reports setting out information about their content moderation systems, other safety and security measures, and human rights initiatives.

However, researchers and other stakeholders have highlighted that the information made available is typically not detailed or specific enough for researchers to understand the actual scope of these problems, or platforms' actions and their effects.[108] Thus, the transparency measures required by the DSA are expected to significantly improve the ability of researchers, civil society, regulators and other stakeholders to understand and scrutinise governance by social media platforms.[109] In particular, Article 40(4) will allow researchers vetted by national regulators to request exactly what internal data they need from a platform to investigate a given topic, rather than relying on the predefined and

---

[100] De Streel, A. et al. 'Online Platforms' Moderation of Illegal Content Online: Law, Practices and Options for Reform', Policy Department for Economic, Scientific and Quality of Life Policies, June 2020.

[101] Gorwa, R., ' What Is Platform Governance?', Information, Communication & Society, Vol. 22, No. 6, 2019, pp. 854–871.

[102] Gorwa, R., ' What Is Platform Governance?', Information, Communication & Society, Vol. 22, No. 6, 2019, pp. 854–871.

[103] Douek, E., ' Verified Accountability: Self-Regulation of Content Moderation as an Answer to the Special Problems of Speech Regulation', Hoover Institution Aegis Series Paper 1903, 18 September 2019, available at: https://www.lawfareblog.com/verified-accountability-self-regulation-content-moderation-answer-special-problems-speech-0.

[104] Douek, E., 'Governing Online Speech: From 'Posts-as-Trumps' to Proportionality and Probability', Columbia Law Review, Vol. 121, No. 3, April 2021, pp. 759–833.

[105] Kaye, D. 'Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression,' United Nations, General Assembly, Human Rights Council A/HRC/38/35. 6 April 2018. https://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/38/35.

[106] Paul, K., 'A Brutal Year: How the 'Techlash' Caught up with Facebook, Google and Amazon', The Guardian, December 28, 2019, sec. Technology.

[107] Trust and safety is the industry term for content moderation, security measures, and other 'principles, policies, and practices that define acceptable behaviour and content online and/or facilitated by digital technologies': see Trust and Safety Professional Association. 'What We Do', Trust & Safety Professional Association, n.d. https://www.tspa.org/what-we-do/.

[108] Vermeulen, M.,' The Keys to the Kingdom', Knight First Amendment Institute Essays and Scholarship, 27 July 2021, available at: https://knightcolumbia.org/content/the-keys-to-the-kingdom.

[109] Vermeulen, M., 'Researcher Access to Platform Data: European Developments', Journal of Online Trust and Safety, Vol. 1, No. 4, September 20, 2022.

aggregated categories of data that platforms make publicly available. This should significantly aid independent research, scrutiny and regulation.

Another prominent self-regulatory initiative is Meta's Oversight Board, set up in 2020 as an independent expert body to review selected content moderation decisions across Meta's platforms (Facebook and Instagram) and provide policy advice. The Board now includes a number of prominent experts on freedom of expression, human rights and media freedom[110] and is operationally independent from Meta, although it is funded through a trust established by Meta[111] and Meta retains significant input into the selection of new members.[112] At the time of writing, the Board has reviewed and ruled on 35 content moderation decisions, as well as issuing two 'policy advisory opinions' in which it discusses Meta's content moderation systems at a more general level.[113] While these rulings only represent a tiny fraction of the moderation decisions appealed by Meta's users, let alone the total number of moderation decisions the company makes every day, the Board has used its rulings in particular cases as a means of publicising more information about how Meta's internal moderation processes work and providing recommendations for improvement. This can be understood as a step forward in terms of transparency and accountability, but it is ultimately a limited one given that information is provided to the public in an unsystematic and ad hoc way; that the public has no input into the Board's composition, decision-making procedures or the normative principles it applies;[114] and that Meta has no obligation to follow the Board's recommendations.

b. Co-regulation: The Codes on Hate Speech and Disinformation

The Commission has also pursued co-regulatory initiatives, agreeing codes of practice with leading industry actors on hate speech in 2016 and disinformation in 2018. These have to date only involved voluntary commitments. However, under Article 45 DSA, such industry codes can now have an official status. This means that they will factor into evaluations of very large online platforms' compliance with their risk mitigation obligations under Article 35, which gives them a quasi-binding character.

In 2022, the Commission agreed with leading platforms, adtech companies and marketing industry organisations on a significantly more detailed version of the CoP on Disinformation, which will become an official code under Article 45. The CoP aims to address disinformation on several fronts, including but not limited to content moderation. It does not expand platforms' duties to moderate defined types of content (though certain types of disinformation content will be deemed illegal under national law, meaning removal can be required under the DSA, as discussed in Section 2.2.1). Instead, platforms commit to take more action on 'manipulative behaviour' associated with strategic dissemination of disinformation, e.g. deleting fake accounts.[115] In addition, platforms and ad industry actors commit to various other actions including expanding 'brand safety' tools which demonetise disinformation;[116]

---

[110] Oversight Board, 'Our commitment', Oversight Board, n.d., available at: https://www.oversightboard.com/meet-the-board/.

[111] Klonick, K., 'The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression', *The Yale Law Journal, Vol. 129, No. 8,* 2020, pp. 2418–2499.

[112] Levy, S., 'Inside Meta's Oversight Board: 2 Years of Pushing Limits', Wired, November 8, 2022. https://www.wired.com/story/inside-metas-oversight-board-two-years-of-pushing-limits/.

[113] Oversight Board, 'Case decisions and policy advisory opinions', Oversight Board, n.d., available at: https://www.oversightboard.com/decision/.

[114] Dvoskin, B., 'Expert Governance of Online Speech', *Harvard International Law Journal, Forthcoming,* July 28, 2022.

[115] The Strengthened Code of Practice on Disinformation 2022, chapter IV.

[116] The Strengthened Code of Practice on Disinformation 2022, chapter II.

strengthening partnerships with independent fact-checkers;[117] and 'adopting safe design practices' in recommender systems to decrease the viral spread of disinformation.[118]

The 2016 Code of Conduct (CoC) on Hate Speech is much shorter.[119] It calls on companies to prohibit incitement to violence and hateful conduct in their terms and conditions, review the majority of reported hate speech in under 24 hours, and increase transparency reporting regarding the moderation of hate speech. These obligations are at this point largely superseded by the DSA's more detailed provisions. However, as Chapter 3 will detail, online hate speech remains a major problem and is not adequately addressed through the DSA's 'procedure over substance' approach. The possibility to establish a more detailed and comprehensive code of conduct under Article 45 DSA represents a promising avenue to improve protection against online hate speech, as outlined in Chapter 3.

2.3. Regulation of platform design and business models

Content policies and moderation systems are not the only factors that bear on fundamental rights and rule of law concerns associated with online content. As the following three sections will explore in more detail, user interactions and the spread of information on platforms are significantly shaped by platforms' design choices – like technical features that enable or constrain certain behaviours, and algorithmic recommender systems that organise and target information – and business models that create certain economic incentives for advertisers, users and platforms themselves. These are of course closely linked, since the product design choices of commercial platforms are ultimately geared towards maximising profit within the parameters of their (at present primarily advertiser-funded) business models.

The need to address platform design and business models has most prominently been discussed by academics and stakeholders in the context of disinformation, where it is widely recognised that just censoring any misleading content is neither effective nor desirable. As Chapter 4 discusses in more detail, many commentators believe it is instead necessary to address the design choices and algorithmic recommender systems that are designed to promote 'engagement' and revenue, which can in some cases lead to promoting the most divisive, sensationalist or controversial content.[120] Similar points have been made in relation to hate speech, abuse and harassment.[121]

It is important not to take a techno-deterministic perspective which blames social media for everything and sees technological change as an easy fix. For example, evidence suggests that the appeal of disinformation and hateful content reflects broader social and economic factors and that users are not

---

[117] The Strengthened Code of Practice on Disinformation 2022, chapter VII.

[118] The Strengthened Code of Practice on Disinformation 2022, commitment 18.

[119] European Commission, 'Code of conduct on countering illegal hate speech online', European Commission, 30 June 2016. Available at: 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.

[120] Marsden, C. et al.,'Platform Values and Democratic Elections: How Can the Law Regulate Digital Disinformation?', Computer Law & Security Review, Vol. 36, April 2020. ; Bennett, L. et al., 'Treating Root Causes, Not Symptoms: Regulating Problems of Surveillance and Personal Targeting in the Information Technology Industries', Hertie School, 2021; Bennett, O., 'The Promise of Financial Services Regulatory Theory to Address Disinformation in Content Recommender Systems', Internet Policy Review, Vol. 10, No. 2, May 11, 2021. ; Jaursch, J., 'Strengthening EU Proposals on Deceptive Platform Design', Stiftung Neue Verantwortung, March 15, 2022. https://www.stiftung-nv.de/en/publication/strengthening-eu-proposals-deceptive-platform-design.

[121] Suzor, N. et al., 'Human Rights by Design: The Responsibilities of Social Media Platforms to Address Gender-Based Violence Online: Gender-Based Violence Online', *Policy & Internet, Vol. 11, No. 1*, March 2019, pp. 84–103.; Griffin, R., 'New School Speech Regulation as a Regulatory Strategy against Hate Speech on Social Media: The Case of Germany's NetzDG', *Telecommunications Policy, Vol. 46, No. 9*, October 2022.

just passive consumers.[122] However, there is evidence that design interventions, like changing recommendation systems, can significantly mitigate the incidence and dissemination of harmful content, even if they do not address root causes.[123]

2.3.1. Regulating recommendation systems and design

The DSA includes some limited regulation of recommender systems. In particular, Article 27 requires all platforms using such systems to clearly explain, in their terms and conditions, the 'main parameters' used in those systems and the reasons for their relative importance. If multiple recommendation settings are available, they must make it easy and accessible for users to change those settings. Article 38 further requires very large online platforms to make available at least one option for each recommender system personalised based on user data. As regards interface design, Article 25 provides that platforms may not design, organise or operate their services in ways that deceive or manipulate users (for example, making it difficult to refuse consent to data collection).

These provisions appear to be based on the assumption that transparency and user choice will address problems with recommender systems. This assumption is questionable, since the spread of harmful information is driven by the dynamics of recommendation systems operating at scale; a few users opting out of the platform's default settings will make very little difference. The usefulness of public-facing transparency can also be questioned. Recommendation systems can be hugely complex, using tens of thousands of parameters;[124] they also do not produce results in a deterministic way, but in the course of complex, recursive interactions between the recommendation algorithm, the platform interface and millions or even billions of users.[125] Given this complexity, if the aim is to strengthen platforms' accountability for recommending harmful content, Article 40(4) – which allows vetted researchers to access platforms' internal data, enabling a better understanding of recommendation systems' operation and outcomes in the real world and at scale – may be more useful.[126]

2.3.2. Regulating advertising

Targeted advertising, a central part of the business model of social media platforms, is also recognised as raising important fundamental rights and rule of law issues. For example, it can facilitate political

---

[122] Roozenbeek, J. et al., 'Susceptibility to Misinformation about COVID-19 around the World', *Royal Society Open Science, Vol. 7,* No. 10, October 2020.; Lewis, B., 'All of YouTube, Not Just the Algorithm, Is a Far-Right Propaganda Machine', FFWD, January 9, 2020. https://ffwd.medium.com/all-of-youtube-not-just-the-algorithm-is-a-far-right-propaganda-machine-29b07b12430; Munger, K., and J. Phillips, 'Right-Wing YouTube: A Supply and Demand Perspective', *The International Journal of Press/Politics, Vol. 27, No. 1*, 2022, pp. 186–219.

[123] Copland, S., 'Reddit Quarantined: Can Changing Platform Affordances Reduce Hateful Material Online?', *Internet Policy Review, Vol. 9, No. 4*, 2020. ; Katsaros, M.,et al., 'Reconsidering Tweets: Intervening during Tweet Creation Decreases Offensive Content', *Proceedings of the International AAAI Conference on Web and Social Media, Vol. 16*, May 31, 2022, pp. 477–487.

[124] Liu, Z. et al., 'Monolith: Real Time Recommendation System With Collisionless Embedding Table', arXiv, September 27, 2022.

[125] Leerssen, P., 'The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems', *European Journal of Law and Technology, Vol. 11, No. 2,* February 24, 2020.; Thorburn, L., J. Stray, and P. Bengani, 'How to Measure the Causal Effects of Recommenders', Understanding Recommenders, November 23, 2022. https://medium.com/understanding-recommenders/how-to-measure-the-causal-effects-of-recommenders-5e89b7363d57.

[126] Leerssen, P., 'The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems', European Journal of Law and Technology, Vol. 11, No. 2, February 24, 2020.

disinformation and divisive campaign tactics,[127] and enable discrimination and privacy violations.[128] These are to a limited extent addressed in the DSA, again mostly through transparency obligations. Article 26 requires platforms presenting online adverts to clearly indicate to users that it is an advert, who paid for it, and the main parameters used for targeting. Article 39 requires very large platforms to establish comprehensive, searchable ad archives for researchers. These provisions again seem to be based on the questionable assumptions that if users are informed about adverts they will not be manipulated, and that public-facing transparency will be sufficient to ensure accountability.

Beyond transparency, risks to fundamental rights, democracy and the rule of law associated with targeted advertising have primarily been addressed through the 2016 General Data Protection Regulation (GDPR), as such advertising necessarily involves processing personal data and must therefore comply with the GDPR's various requirements. Targeted advertising requires a legal basis for processing this data, which will generally require either free and explicit consent from the user, a legitimate interest on the part of the company, or necessity for the performance of the contract.[129] In January 2023, the European Data Protection Board ruled that Meta could not continue targeting ads based on user's online activity without affirmative, opt-in consent, and that it could not rely on contract as a legal basis, as targeted advertising is not a core element of its service and thus not necessary for the performance of its contract with users.[130] This has been interpreted by many as a significant blow to Meta's business model (and that of many other platforms), although Meta is appealing the decision and it is not yet clear how much it will ultimately affect targeted advertising practices.[131]

Article 26(3) DSA also now entirely bans the targeting of adverts based on 'sensitive data' as defined in Article 9 GDPR (notably including race, religion, ethnicity, sexuality and political views). This is aimed at preventing discriminatory targeting, particularly for resources and opportunities such as job adverts. Unfortunately, however, it is unlikely to have much impact, as algorithmically-targeted advertising can produce highly discriminatory results without directly using sensitive data.[132]

Although its final outcome remains uncertain, the Commission's 2022 proposal for a Political Advertising Regulation (PAR)[133] would more strictly regulate transparency and disclosure for political adverts.[134] These are defined in Article 2(2) as messages which are placed by, for or on behalf of political actors or which are liable to influence electoral processes. The original proposal bans targeting based

---

127 Borgesius, F.J. et al. 'Online Political Microtargeting: Promises and Threats for Democracy', *Utrecht Law Review, Vol. 14, No. 1*, February 9, 2018, p. 82.

128 Griffin, R., 'Tackling Discrimination in Targeted Advertising: US regulators take very small steps in the right direction – but where is the EU?', *Verfassungsblog*, June 23, 2022.; Mühlhoff, R., and T. Willem, 'Social Media Advertising for Clinical Studies: Ethical and Data Protection Implications of Online Targeting', Pre-Print, *Forthcoming in Big Data & Society*, April 2022.

129 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1, art 6.

130 'Facebook and Instagram Decisions: "Important Impact on Use of Personal Data for Behavioural Advertising"', European Data Protection Board, January 12, 2023. https://edpb.europa.eu/news/news/2023/facebook-and-instagram-decisions-important-impact-use-personal-data-behavioural_hu.

131 MacCarthy, M., 'The European Data Protection Board (EDPB) goes after tech's personalized ad business model' Brookings, 1 February 2023, available at: https://www.brookings.edu/blog/techtank/2023/02/01/the-european-data-protection-board-goes-after-techs-personalized-ad-business-model/.

132 Griffin, R., 'Tackling Discrimination in Targeted Advertising: US regulators take very small steps in the right direction – but where is the EU?', *Verfassungsblog*, June 23, 2022.

133 Proposal for a Regulation of the European Parliament and of the Council on the transparency and targeting of political advertising,
COM/2021/731 final, available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0731.

134 Van Drunen, M. et al., 'New Actors and Risks in Online Advertising', European Audiovisual Authority, Strasbourg, 2022.

on sensitive data, as in Article 26(3) DSA, but several Member States have backed a proposal to ban all personalised targeting, except on broad criteria such as location.[135] Civil society groups have however criticised the very broad scope of the PAR, since it counts unpaid messages relating to political issues as advertising, which could place significant compliance burdens on civil society and restrict independent political advocacy and commentary.[136]

2.3.3. Safe design practices

Finally, the CoP on Disinformation introduces the concept of 'safe design practices' in relation to disinformation.[137] The commitments under this heading are much more extensive than just being transparent about recommender systems, and if fully enforced, will require platforms to much more significantly rethink their product design processes. Signatories commit to researching measures to reduce the spread of disinformation in their design processes, pre-testing recommender systems and other design features to avoid harmful impacts, and developing clear metrics to evaluate the success of these measures. These explicit requirements for platforms to thoroughly test for potential harms and to develop and evaluate mitigation measures throughout design processes could be much more effective than simply providing transparency about their ultimate outcomes and hoping that users and researchers could pressure platforms into doing something differently. Box 1 provides some context on the potential of safe design practices in platform governance.

However, these commitments are not formally binding, and much will depend on how they are implemented in practice. As they are relevant to assessing very large online platforms' compliance with Articles 34-35 DSA, the Commission can use the threat of a fine under these provisions to push for clear and thorough documentation of how platforms are implementing safe design practices. It would also be possible to develop similar codes demanding safe design practices to address other types of harmful content and behaviours, such as hate speech and harassment. How this will look in practice remains undecided.

Box 1: A systemic approach to content moderation

Contemporary scholarship on content moderation has increasingly argued that a governance framework that focuses on creating rules for deciding individual cases misses the systemic, and emergent, properties of the online ecosystem and the challenges that it poses. In her influential 2022 Harvard Law Review article 'Content Moderation as Systems Thinking', Evelyn Douek argued that the picture of content moderation as an aggregation of many individual cases where users' speech rights are adjudicated is misleading and incomplete.

Douek highlighted that 'the scale and speed of online speech means content moderation cannot be understood as simply the aggregation of many (many!) individual adjudications', and argues that as a result, focusing on correcting individual decisions or ensuring that platforms treat individuals fairly produces 'accountability theater rather than actual accountability.' Douek thus argued that content moderation should rather be seen as a 'project of mass speech administration', with a focus on the large-scale systems and processes developed to govern online speech at scale. Moreover, she stresses the need for lawmakers to adopt a 'second wave of regulatory thinking', focusing on ex ante

[135] Killeen, M., 'Germany Supports Ban on Personal Data for Political Ads', Euractiv, September 7, 2022. https://www.euractiv.com/section/digital/news/germany-supports-ban-on-personal-data-for-political-ads/.

[136] European Partnership for Democracy et al. 'Civil Society Public Letter on the Council's Proposed General Approach to the Regulation of Political Advertising, Calling for a Regulation That Delivers for Democracy and Fundamental Rights', October 28, 2022.

[137] The Strengthened Code of Practice on Disinformation 2022, commitment 18.

regulation of how these systems are designed, rather than ex post accountability for their outcomes in individual cases.

Along similar lines, the Integrity Institute, an organisation founded in 2020 by former Meta employees, advocates for a more holistic and systemic approach to content moderation, thinking less about individual decisions and more about incentives, information ecosystems and designing systems for integrity. Its founder Sahar Massachi describes this approach as similar to the traffic rules of a city, and suggests that like urban design, platform design can introduce friction which slows down or discourages certain forms of harmful interaction. For example, social media companies could (and should) impose escalating costs on actions associated with disinformation operations, such as creating many groups at once, or commenting on a thousand videos in an hour.

Sources: Massachi, S., How to save our social media by treating it like a city. MIT Technology Review, December 20, 2021, available at: https://www.technologyreview.com/2021/12/20/1042709/how-to-save-social-media-treat-it-like-a-city; Douek, E., 'Content Moderation as Systems Thinking', *Harvard Law Review, Vol. 136, No. 2*, December 2022, pp. 526–607.

3. HATE SPEECH

3.1. Introduction

This chapter addresses the moderation of hate speech on social media. It outlines how platforms currently moderate hateful content, as well as the relevant legal framework, and evaluates these existing measures from a fundamental rights perspective. Restricting hate speech limits freedom of expression. However, this must be balanced against the need to protect the freedom of expression and other fundamental rights of those targeted, as well as equal participation in democratic debate, since hate speech limits the opportunities of marginalised groups to participate in online media.

Currently the primary way platforms deal with hate speech is through content moderation: banning it in their terms and conditions and removing it where they detect it. This is also encouraged by the existing European legal framework, which requires platforms to delete some forms of illegal hate speech. Such measures are important from a fundamental rights perspective, to protect marginalised social groups against unchecked hate speech. At the same time, online hate speech cannot simply be addressed only through more and stricter content moderation. Efforts to censor hate speech are important in dealing with the most harmful content, but they are inevitably imperfect. Oftentimes, in practice, they lead to further suppression of marginalised users, while failing to protect them effectively. Thus, EU regulation of hate speech – in particular through very large online platforms' due diligence obligations under the DSA – should also place more emphasis on alternative interventions which could discourage hate speech and support affected users without simply removing more content.

Section 3.2 provides some necessary background on the concept of hate speech, discussing and comparing current definitions in EU law and outlining why it is harmful to democratic values. Section 3.2.3 briefly outlines the relevant European human rights framework and argues that a rights-respecting approach to hate speech would aim to ensure the most harmful content is moderated, but would also pursue solutions beyond moderation. Section 3.3 outlines how leading social media platforms currently approach hate speech moderation. Section 3.4 discusses the human rights implications of current moderation practices, highlighting three particular issues: the unreliability of existing moderation systems, which means that hate speech policies often are not enforced; bias and discrimination, which mean that efforts to moderate hate speech in practice disproportionately censor marginalised groups; and a failure to address the harms of hate speech in a more holistic way. Section 3.5 then discusses the limitations of the existing EU legal framework in relation to these issues, before Section 3.6 concludes with recommendations.

3.2. Background

3.2.1. Defining hate speech

There is no single, established definition of hate speech. EU Member States have their own hate speech laws.[138] To the extent that online content is criminalised under national laws, it is governed by the intermediary liability framework in the ECD and now DSA, under which social media platforms can become legally liable for hosting illegal content once they have been notified about that specific content. In EU law, the 2008 Council Framework Decision on Racism requires Member States to

[138] Iacob, O., "Hate Crime and Hate Speech in Europe: Comprehensive Analysis of International Law Principles, EU-Wide Study and National Assessments | European Website on Integration", 2015, available at: https://ec.europa.eu/migrant-integration/library-document/hate-crime-and-hate-speech-europe-comprehensive-analysis-international-law_en.

criminalise certain forms of hate speech, at least where they are threatening, abusive, insulting or liable to disturb public order.[139] The 2016 Code of Conduct on Hate Speech defines hate speech by reference to this provision, as 'all 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'.[140]

Compared to common understandings of hate speech, this definition is evidently incomplete. National hate speech laws vary widely, but they typically offer protection for characteristics which are important to people's identities and associated with prejudice and discrimination, commonly including aspects like gender, sexuality and disability as well as race, ethnicity and religion.[141] For example, Articles 32 and 33 of France's 1881 Press Law prohibit defamation or insults targeted at a person or group of persons based on their nationality, ethnicity, race, religion, sex, sexual orientation, gender identity or handicap. The exclusion of these other protected characteristics from the Code of Practice on Hate Speech, the EU's primary initiative regulating hate speech on social media, is puzzling—especially given that the possibility of a broader approach is illustrated in Article 28b(1)(b) AVMSD. This provision only applies to video-sharing platforms, but requires them to take preventive measures against 'incitement to violence or hatred directed against a group of persons or a member of a group based on any of the grounds referred to in Article 21 of the Charter'. Given clear evidence that hate speech based on other characteristics is a serious problem on social media—as the following subsection will discuss—this study will define hate speech in accordance with Article 28(1)(b) AVMSD.

3.2.2. Intersectionality and marginalisation

Beyond causing immediate harms to the dignity and emotional wellbeing of those targeted,[142] online hate speech has serious consequences for the rule of law, fundamental rights and democracy. The widespread presence of hate speech on social media can influence social norms, increasing the perceived acceptability of hateful rhetoric and ideas, and promoting social division and prejudices against marginalised groups.[143] The role of exposure to online hate content in individual political radicalisation remains debated.[144] However, online hate speech and disinformation targeting marginalised groups can encourage broader political mobilisations against such groups, which can in turn promote extremism and violence.[145] There is also some evidence from the UK and Germany suggesting causal links between the overall prevalence of online hate content and the incidence of

[139] 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.

[140] 'The EU Code of Conduct on Countering Illegal Hate Speech Online', p.1, available at: 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.

[141] Sellars, A., 'Defining Hate Speech', SSRN Scholarly Paper, Rochester, NY, December 1, 2016.

[142] *See Amnesty* International, 'Toxic Twitter - A Toxic Place for Women', *Amnesty.org,* 2018, available at: Toxic Twitter - A Toxic Place for Women.

[143] González-Bailón, S., and Y. Lelkes, 'Do Social Media Undermine Social Cohesion? A Critical Review', *Social Issues and Policy Review*, December 31, 2022.

[144] Klompmaker, N., 'Censor Them at Any Cost? A Social and Legal Assessment of Enhanced Action Against Terrorist Content Online', *Amsterdam Law Forum*, Vol. 11, No. 3, June 1, 2019, p. 3.

[145] Ritholtz, S., 'Fanning the Flames of Hate: The Transnational Diffusion of Online Anti-LGBT+ Rhetoric and Offline Mobilisation', *GNET*, available at: https://gnet-research.org/2022/11/29/fanning-the-flames-of-hate-the-transnational-diffusion-of-online-anti-lgbt-rhetoric-and-offline-mobilisation-2/.

violent hate crime[146] – although the complexity of these issues means that firmly establishing causal effects is difficult.[147]

Finally, hate speech also affects the ability of targeted groups and individuals to use and benefit from social media. Studies show that it often leads people to self-censor what they say in future, or withdraw from online discussions altogether[148]. Indeed, this may be a core motivation and function of such speech. Scholars have theorised online hate speech as a way of enforcing social norms which exclude or devalue certain groups, by punishing those who are perceived as speaking too prominently or otherwise violating community norms.[149] This restricts their ability to benefit from the social and economic opportunities offered by social media.[150] It also restricts their participation in political life and other public debates, ensuring the public sphere continues to privilege the views of straight, white, able-bodied cisgender men.[151]

One implication of this is that focusing only on hate speech, understood as involving 'incitement to violence or hatred', may be too narrow to address the impacts discussed in the preceding paragraphs. Most of the studies cited above do not focus only on hate speech in this sense, but on a wider range of abusive behaviours which target users or groups based on protected identity characteristics. These include threats, harassment and abusive language, which are often ongoing and/or coordinated between many users, and privacy violations, such as publication of personal information or intimate images.[152] These multiple forms of online abuse often occur together[153] and ultimately serve the same functions as hate speech: directly harming marginalised users and excluding them from online conversations. As such, policy and regulation concerned with online hate speech should arguably take

---

[146] Williams, M. et al., 'Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime', *The British Journal of Criminology*, July 23, 2019; Müller, K., and C. Schwarz, 'Fanning the Flames of Hate: Social Media and Hate Crime', SSRN Scholarly Paper, Rochester, NY, June 5, 2020.

[147] Buerger, C., 'Speech as a Driver of Intergroup Violence: A Literature Review', *Dangerous Speech Project*, June 16, 2021. https://dangerousspeech.org/wp-content/uploads/2021/06/Speech-and-Violence-Lit-Review.pdf.

[148] Franks, M. A., 'Beyond the Public Square: Imagining Digital Democracy', *The Yale Law Journal*, 2021, available at: https://www.yalelawjournal.org/forum/beyond-the-public-square-imagining-digital-democracy; Duguay, S. et al., 'Queer Women's Experiences of Patchwork Platform Governance on Tinder, Instagram, and Vine', *Convergence: The International Journal of Research into New Media Technologies*, Vol. 26, No. 2, April 2020, pp. 237–252; 'Amnesty Reveals Alarming Impact of Online Abuse against Women', *Amnesty International*, November 20, 2017; 'Measuring the Prevalence of Online Violence against Women', *Jigsaw Infographic*; Posetti, J. et al., 'The Chilling: global trends in online violence against women journalists', UNESCO, 2021,available at: https://unesdoc.unesco.org/ark:/48223/pf0000377223.

[149] Chemaly, S., 'Demographics, Design, and Free Speech: How Demographics Have Produced Social Media Optimized for Abuse and the Silencing of Marginalized Voices', by S. Chemaly, Free Speech in the Digital Age, *Oxford University Press*, 2019, pp. 150–169,
available at: https://academic.oup.com/book/27505/chapter-abstract/197448652?redirectedFrom=fulltext; Marwick, A.E., 'Morally Motivated Networked Harassment as Normative Reinforcement', *Social Media + Society*, Vol. 7, No. 2, April 2021. https://journals.sagepub.com/doi/full/10.1177/20563051211021378.

[150] Siapera, E., 'Online Misogyny as Witch Hunt: Primitive Accumulation in the Age of Techno-Capitalism', in D. Ging and E. Siapera (eds.), *Gender Hate Online*, Springer International Publishing, Cham, 2019, pp. 21–43.

[151] Franks, M. A., 'Beyond the Public Square: Imagining Digital Democracy', *The Yale Law Journal*, 2021, available at: https://www.yalelawjournal.org/forum/beyond-the-public-square-imagining-digital-democracy.

[152] Khoo, C., *Deplatforming Misogyny: Report on Platform Liability for Technology-Facilitated Gender-Based Violence*, Women's Legal Education and Action Fund (LEAF), 2021, available at: https://www.leaf.ca/wp-content/uploads/2021/04/Full-Report-Deplatforming-Misogyny.pdf; Posetti, J. et al., 'The Chilling: global trends in online violence against women journalists', UNESCO, 2021, available at: https://unesdoc.unesco.org/ark:/48223/pf0000377223; The Economist Intelligence Unit, 'Measuring the prevalence of online violence against women', *The Economist,* 2021, available at: https://onlineviolencewomen.eiu.com/.

[153] The Economist Intelligence Unit, 'Measuring the prevalence of online violence against women', *The Economist,* 2021, available at: https://onlineviolencewomen.eiu.com/.

a broader view, encompassing all abusive and harmful behaviour which targets people based on protected characteristics, not only incitement to hatred and violence.

To adequately recognise and address how online hate speech and harassment affect people in practice, it is also essential to understand it as intersectional. Broadly, intersectional theory argues that people often simultaneously face multiple forms of social marginalisation, discrimination and prejudice, which interact with one another and mean that people who share some identity characteristics may nonetheless be marginalised in very different ways. For example, Black women may have very different experiences to either white women or Black men. Consequently, pursuing equality requires an analysis of how multiple social structures and prejudices interrelate in particular situations, instead of dividing people into broad identity categories.[154]

Research on online hate speech has been criticised for failing to take an intersectional approach, generalising about misogynist or racist hate speech without considering the different experiences and vulnerabilities of different people affected by these issues.[155] With that caveat, there are numerous studies showing that people from various minority groups are highly likely to encounter hate speech and various other forms of abuse, threats, harassment and privacy violations on social media, and on the internet more broadly. Generally, LGBTQ+ people[156] and people of colour[157] are more likely to encounter online hate speech, abuse and harassment. People with disabilities also face high rates of ableist hate speech and abuse.[158] Evidence on whether women face more online hate speech and harassment than men is conflicting, but suggests that women are generally more impacted than men by threatening behaviour, and more likely to respond by self-censoring what they say online.[159] Some of these studies do use an intersectional lens, showing that people facing multiple forms of

---

[154] Crenshaw, K., 'Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics', *The University of Chicago Legal Forum*, 1989, p. 139-167.

[155] Vickery, J. R., and T. Everbach, eds., *Mediating Misogyny: Gender, Technology, and Harassment*, Softcover reprint of the hardcover 1st edition 2018, Palgrave Macmillan, Cham, Switzerland, 2019, available at: https://link.springer.com/book/10.1007/978-3-319-72917-6.

[156] Haimson, O.L., *Transgender Experiences in Online Harassment*, Microsoft Research Social Computing Symposium, 2016; Lenhart, A., M. Ybarra, K. Zickuhr, and M. Price-Feeney, 'Online Harassment, Digital Abuse, and Cyberstalking in America', *Data & Society*, November 21, 2016, available at: https://datasociety.net/library/online-harassment-digital-abuse-cyberstalking/; Plan International, 'Free to Be Online? Girls' and Young Women's Experiences of Online Harassment', *Plan International*, 2020, available at: https://plan-international.org/uploads/2022/02/sotwgr2020-commsreport-en-2.pdf; 'Digital Youth Index Report 2022', *Digital Youth Index*, 2022, available at: https://digitalyouthindex.uk/.

[157] Lenhart, K. et al., 'Online Harassment, Digital Abuse, and Cyberstalking in America', *Data & Society*, November 21, 2016, available at: https://datasociety.net/library/online-harassment-digital-abuse-cyberstalking/; Plan International, 'Free to Be Online? Girls' and Young Women's Experiences of Online Harassment', *Plan International*, 2020, available at: https://plan-international.org/uploads/2022/02/sotwgr2020-commsreport-en-2.pdf.

[158] 'Online Disability Hate Crimes Soar 33%', *Leonard Cheshire*, 2019, available at: https://www.leonardcheshire.org/about-us/our-news/press-releases/online-disability-hate-crimes-soar-33; Harmer, E., and Lumsden, K., eds., *Online Othering: Exploring Digital Violence and Discrimination on the Web*, 1st ed. 2019., *Palgrave Studies in Cybercrime and Cybersecurity*, Springer International Publishing : Imprint: Palgrave Macmillan, Cham, 2019, available at: https://link.springer.com/book/10.1007/978-3-030-12633-9; Plan International, 'Free to Be Online? Girls' and Young Women's Experiences of Online Harassment', *Plan International*, 2020, available at: https://plan-international.org/uploads/2022/02/sotwgr2020-commsreport-en-2.pdf.

[159] Lenhart, K. et al., 'Online Harassment, Digital Abuse, and Cyberstalking in America', *Data & Society*, November 21, 2016, available at: https://datasociety.net/library/online-harassment-digital-abuse-cyberstalking/; Chemaly, S., 'Demographics, Design, and Free Speech: How Demographics Have Produced Social Media Optimized for Abuse and the Silencing of Marginalized Voices', in Brison, S.J. and Gerber, K., *Free Speech in the Digital Age*, Oxford University Press, 2019, pp. 150–169, available at: https://academic.oup.com/book/27505/chapter-abstract/197448652?redirectedFrom=fulltext; Khoo, C., *Deplatforming Misogyny: Report on Platform Liability for Technology-Facilitated Gender-Based Violence*, Women's Legal Education and Action Fund (LEAF), 2021, available at: https://www.leaf.ca/wp-content/uploads/2021/04/Full-Report-Deplatforming-Misogyny.pdf.

marginalisation are particularly vulnerable to online hate speech and harassment,[160] and are likely to be targeted based on these intersecting identities.[161]

EU policy on hate speech must recognise this reality and the particular vulnerabilities of those facing multiple forms of marginalisation. Policies should thus define hate speech as targeting people based on any characteristic protected by Article 21 of the Charter of Fundamental Rights, or any combination of characteristics of which at least one is protected.

3.2.3. Hate speech and fundamental rights

As the above discussion suggests, the complex fundamental rights issues involved in regulating online hate speech cannot just be reduced to a binary trade-off between preventing hate speech, on the one hand, and protecting free speech, on the other.[162] Feminist and critical race theorists have long argued that free speech is not best served by permissive attitudes to hate speech, because hate speech itself limits the free speech of those who are marginalised, and who have already always had the fewest opportunities to speak freely.[163] It directly drives them out of public debates, and more generally, it also reinforces prejudices and stereotypes which mean they are less likely to be listened to and respected.[164] These arguments are now mainstream in scholarship on social media law. It is widely accepted that some moderation of hate speech (and other offensive content such as graphic violence) is necessary if social media platforms are to function as spaces for constructive or enjoyable interactions.[165]

In the EU, freedom of expression and information are protected by Article 11 of the Charter of Fundamental Rights and Article 10 of the European Convention on Human Rights (which provides authoritative guidance on the scope of corresponding Charter rights: see Article 52(3) of the Charter). The European Court of Human Rights (ECtHR) has developed an extensive case law on the scope of these rights. In general, it has aimed to strongly protect them and permit restrictions only where strictly

---

[160] Chemaly, S., 'Demographics, Design, and Free Speech: How Demographics Have Produced Social Media Optimized for Abuse and the Silencing of Marginalized Voices', in Brison, S.J. and Gerber, K., *Free Speech in the Digital Age*, Oxford University Press, 2019, pp. 150–169, available at: https://academic.oup.com/book/27505/chapter-abstract/197448652?redirectedFrom=fulltext; Plan International, 'Free to Be Online? Girls' and Young Women's Experiences of Online Harassment', *Plan International*, 2020, available at: https://plan-international.org/uploads/2022/02/sotwgr2020-commsreport-en-2.pdf; Posetti, J. et al., 'The Chilling: global trends in online violence against women journalists', UNESCO, 2021, available at: https://unesdoc.unesco.org/ark:/48223/pf0000377223.

[161] Starr, T.J., *The Unbelievable Harassment Black Women Face Daily on Twitter*, Alternet.Org, 2014. available at: https://www.alternet.org/2014/09/unbelievable-harassment-black-women-face-daily-twitter; 'Amnesty Reveals Alarming Impact of Online Abuse against Women', *Amnesty International*, November 20, 2017; The Economist Intelligence Unit, 'Measuring the Prevalence of Online Violence against Women', *The Economist Intelligence Unit,* 2021, available at: https://onlineviolencewomen.eiu.com/ ; Hackworth, L., 'Limitations of "Just Gender": The Need for an Intersectional Reframing of Online Harassment Discourse and Research', in Vickery, J.R. and Everbach, T. (eds.), *Mediating Misogyny: Gender, Technology & Harassment*, Palgrave MacMillan, 2018, pp. 51-70.

[162] Guney, G. et al., eds., *Towards Gender Equality in Law: An Analysis of State Failures from a Global Perspective*, Springer International Publishing, Cham, 2022, available at: https://link.springer.com/book/10.1007/978-3-030-98072-6.

[163] Matsuda, M. J., ed., *Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*, *New Perspectives on Law, Culture, and Society*, Westview Press, Boulder, Colo, 1993. https://scholarship.law.columbia.edu/books/287/; Franks, M. A., 'Beyond the Public Square: Imagining Digital Democracy', *The Yale Law Journal*, 2021, available at: https://www.yalelawjournal.org/forum/beyond-the-public-square-imagining-digital-democracy.

[164] Khoo, C., *Deplatforming Misogyny: Report on Platform Liability for Technology-Facilitated Gender-Based Violence*, Women's Legal Education and Action Fund (LEAF), 2021, available at: https://www.leaf.ca/wp-content/uploads/2021/04/Full-Report-Deplatforming-Misogyny.pdf.

[165] Review, C.L., 'Governing Online Speech: From "Posts-As-Trumps" To Proportionality And Probability', *Columbia Law Review*, n.d., available at: https://columbialawreview.org/content/governing-online-speech-from-posts-as-trumps-to-proportionality-and-probability/.

necessary.[166] Restrictions of these rights must satisfy the ECtHR's standard four-step test:[167] being legally prescribed, pursuing a legitimate aim, being necessary in a democratic society, and being a proportionate means to reach the aim pursued. Two particular elements in the relevant case law are worth highlighting in the context of online hate speech.

First, as noted above, binary oppositions between restricting hate speech and protecting free speech are misleading. Under the ECHR, states have positive obligations to guarantee media pluralism, and to endeavour to create an environment in which all can participate in public debate without fear.[168] As discussed above, hate speech and other forms of online harassment and abuse systematically deprive people from marginalised groups of opportunities to express themselves on social media. Thus, while censorship of online hate speech necessarily restricts the free speech of the user sharing it, it can also be necessary to protect the free speech of others. However, as the following subsections will discuss in more detail, given the inherent limitations and bias of systems for moderating hate speech, alternative interventions may be more effective in creating a safe online environment than stricter moderation obligations.

Second, under ECtHR jurisprudence, some forms of extreme hate speech are not protected by the right to free expression at all. In this context, the ECtHR has applied Article 17, which provides that the ECHR does not protect behaviour aimed at the destruction of the rights it establishes, to hold that the Article 10 protection does not cover certain forms of racist speech, such as extreme dehumanising language or calls for deportation aimed at racial minorities.[169] While it has used the term hate speech to describe such categories, it has never actually defined the term.[170]

Relevantly for the social media context, the ECtHR applied this principle to online intermediary liability in *Delfi v Estonia*,[171] in which it upheld an order for damages against an online news website for hosting comments expressing hatred and threats against an individual mentioned in one of its articles. The majority judgment held that because these comments qualified as 'manifest expressions of hatred and blatant threats to [the applicant's] physical integrity', they were not protected by Article 10 ECHR. While the judgment did represent an interference with the news publisher's Article 10 rights, given the extreme nature of the comments, it was proportionate to hold it liable for hosting them – even though it had not known about them and on being notified had removed them immediately. Since the comments in question, though offensive and threatening, did not target any person or group based on a protected characteristic, this adds further confusion as to how the ECtHR defines hate speech and when it falls outside the protection of Article 10.

The *Delfi* decision was also heavily criticised, including by two dissenting judges, for opening the door to strict intermediary liability laws and excessive censorship: if publishers can be strictly liable for hosting content of which they have not been notified, they will be incentivised to significantly restrict

---

166 McGonagle, T., 'Free Expression and Internet Intermediaries: The Changing Geometry of European Regulation', in G. Frosio (ed.), by T. McGonagle, *Oxford Handbook of Online Intermediary Liability*, Oxford University Press, 2020, pp. 466–485.

167 Bayer, J., and P. Bárd, 'Hate speech and hate crime in the EU and the evaluation of online content regulation approaches', Policy Department for Citizens' Rights and Constitutional Affairs, July 2020, available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/655135/IPOL_STU(2020)655135_EN.pdf.

168 Judgment of 14 September 2010, *Dink v. Turkey,* applications no. 2668/07, 6102/08, 30079/08, 7072/09 and 7124/09.

169 *Seurot v France* [2004]; *Norwood v UK* [2004].

170 McGonagle, T., 'Free Expression and Internet Intermediaries: The Changing Geometry of European Regulation', in G. Frosio (ed.), by T. McGonagle, Oxford Handbook of Online Intermediary Liability, Oxford University Press, 2020, pp. 466–485.

171 Judgment of October 10, 2013, *Delfi AS v Estonia* (application number 64569/09).

or filter what users can publish as the only way of avoiding liability.[172] In future cases with similar facts, the ECtHR has declined to apply *Delfi* on the grounds that the content involved, though defamatory and offensive, did not involve hate speech or incitement to violence.[173] However, a clear definition of hate speech or explanation of what takes content outside the protection of Article 10 is lacking.[174]

The *Delfi* approach is problematic for several reasons. First, holding that hate speech is not only capable of being legitimately and proportionately restricted, but is not protected by Article 10 at all, creates a concerning accountability gap. This is because it could mean laws restricting speech are not scrutinised by courts at all. Since the ECtHR's definition of hate speech is vague, this creates significant room for governments to restrict speech while claiming that Article 10 does not apply and no proportionality assessment is necessary, even where such restrictions affect valuable and protected speech. The vagueness of the ECtHR's definition of hate speech also means it could be applied in an arbitrary and biased way, so that the protection of Article 10 will not extend equally to everyone.

Moreover, it is especially concerning in the social media context to hold that restrictions on hate speech do not engage Article 10 at all, because there is no such thing as perfection in content moderation. Regardless of how objectionable the speech which is targeted, any prohibition will necessarily also censor some non-hateful speech, and so raises fundamental rights concerns which should be assessed for legality and proportionality. First, even assuming that there are clear and well-understood definitions of hate speech in theory, enforcement is always imperfect in practice. Any technical or manual processes developed to apply such standards at scale – even if they are highly accurate in percentage terms – will necessarily produce many false positives (mistakenly removing content) and false negatives (ignoring content that should be removed).[175] Moreover, in reality, any standards established to govern interactions between millions or billions of users must necessarily be indeterminate and non-exhaustive, leaving room for disagreement about individual cases.[176] What constitutes hate speech is inherently contestable, and attempts to identify it in practice are often imbued with bias against minority groups and forms of communication that are stigmatised.[177] Legal prohibitions on hate speech can be abused by governments to silence marginalised groups and dissenting voices.[178]

This cannot however be understood as an argument against any such prohibitions. Protecting the freedom of expression of marginalised groups, and ensuring fair and inclusive democratic debate, demands reasonably effective moderation of online hate speech. This requires the acceptance of some

---

[172] Keller, D., 'Policing Online Comments in Europe: New Human Rights Case Law in the Real World', 2016, *The Center for Internet and Society*, Stanford Law School, available at: https://cyberlaw.stanford.edu/blog/2016/04/policing-online-comments-europe-new-human-rights-case-law-real-world; McGonagle, T., 'Free Expression and Internet Intermediaries: The Changing Geometry of European Regulation', in G. Frosio (ed.), by T. McGonagle, *Oxford Handbook of Online Intermediary Liability*, Oxford University Press, 2020, pp. 466–485.

[173] MTE v Hungary [2016], Pihl v Sweden [2017].

[174] McGonagle, T., 'Free Expression and Internet Intermediaries: The Changing Geometry of European Regulation', in G. Frosio (ed.), by T. McGonagle, *Oxford Handbook of Online Intermediary Liability*, Oxford University Press, 2020, pp. 466–485; Sottiaux, S., 'Conflicting Conceptions of Hate Speech in the ECtHR's Case Law', *German Law Journal*, Vol. 23, No. 9, December 2022, pp. 1193–1211.

[175] Douek, E., 'Governing Online Speech: From 'Posts-as-Trumps' to Proportionality and Probability', *Columbia Law Review*, Vol. 121, No. 3, April 2021, pp. 759–833. ; Chowdhury, N., *Automated Content Moderation: A Primer, Cyber Policy Center*, Stanford University, March 19, 2022.

[176] Leerssen, P., 'An End to Shadow Banning? Transparency rights in the Digital Services Act between content moderation and curation', *Computer Law & Security Review, Vol. 48*, 105790.

[177] Hare, I., and J. Weinstein, eds., *Extreme Speech and Democracy*, 1st ed., Oxford University Press, Oxford, 2009.

[178] Sottiaux, S., 'Conflicting Conceptions of Hate Speech in the ECtHR's Case Law', *German Law Journal*, Vol. 23, No. 9, December 2022, pp. 1193–1211.

false positives. The crucial questions are not just whether the right balance is struck between freedom of expression and protecting marginalised groups from hate speech, but rather whose freedom of expression is protected, what kinds of error rates can be accepted, and which groups are most likely to bear the costs of errors.[179] A fundamental rights-respecting approach to hate speech regulation would thus be one that bans some particularly harmful kinds of speech entirely, but does not simply respond to every type of harmful content by banning it. Instead, it should focus on ensuring that platforms operate effective moderation systems that can reliably identify harmful speech, that they are sensitive to the particular needs and vulnerabilities of marginalised users, and that they work to design online environments that discourage abusive behaviour and enable inclusive participation.

3.3. Current approaches to hate speech moderation

Before analysing how moderation of hate speech is currently regulated in Europe, it is useful to understand how it typically works in practice. An influential typology by Robyn Caplan identifies three broad approaches to moderating social media content.[180] In artisanal moderation, small teams of employees manually examine content flagged as potentially problematic to apply and develop community standards in a flexible, context-dependent way. In community moderation, user moderators are permitted to develop and enforce their own standards for particular groups or forums on the platform. In industrial moderation, larger teams – sometimes platform employees, but more often contracted through outsourcing companies – work in a highly routinised way, supported by software tools, to review content according to standardised rules.

Industrial moderation is the predominant model at the largest platforms:[181] artisanal moderation is too resource-intensive to easily scale for millions- or billions-strong user bases, while community moderation may play a role in maintaining norms of particular communities (notably for Reddit forums and Facebook's Groups) but cannot be relied on to enforce legal obligations and broader social expectations that platforms should never host certain types of content. The 'industrial model' of content moderation has also effectively been codified and established as an industry standard by the DSA, which requires all platforms to develop and enforce standardised content policies in their contractual terms and conditions, and to implement standardised reporting and appeals procedures.[182] This section will thus focus on the industrial moderation approach as the most relevant for major platforms dealing with hate speech at scale.

Under the European intermediary liability regime, platforms can become liable for illegal content once they have actual knowledge of it, typically because it has been reported by a third party, which then gives them a strong legal incentive to take it down. Traditionally, industrial content moderation has largely followed this 'notice-and-takedown' model, even for content which is not illegal: users are given the option to report content they dislike, and reported content is then reviewed by moderation staff

[179] Ananny, M., 'Probably Speech, Maybe Free: Toward a Probabilistic Understanding of Online Expression and Platform Governance', Knight First Amendment Institute at Columbia University, 2019, available at: http://knightcolumbia.org/content/probably-speech-maybe-free-toward-a-probabilistic-understanding-of-online-expression-and-platform-governance; Douek, E., 'Governing Online Speech: From 'Posts-as-Trumps' to Proportionality and Probability', *Columbia Law Review,* Vol. 121, No. 3, April 2021, pp. 759–833.

[180] Caplan, R., 'Content or Context Moderation?', *Data & Society*, November 14, 2018, available at: https://datasociety.net/library/content-or-context-moderation/.

[181] Caplan, R., 'Content or Context Moderation?', *Data & Society*, November 14, 2018, available at: https://datasociety.net/library/content-or-context-moderation/.

[182] Keller D., 'The DSA's Industrial Model for Content Moderation', Verfassungsblog, 2022 ; Douek, E., 'Content Moderation as Systems Thinking', Harvard Law Review, Vol. 136, No. 2, December 2022, pp. 526–607.

for illegality or incompatibility with the platform's content policies.[183] This approach offers companies numerous advantages: it allows platforms to sort through large volumes of content by effectively harnessing free labour from users; offers insight into what types of content are likely to put users off using the platform; and serves a legitimating function by allowing platforms to frame their moderation policies as a reflection of user preferences, while also being able to ignore user reports when they have other priorities.[184] Article 16 DSA further entrenches this approach by requiring all platforms to allow users to report illegal content. Sufficiently substantiated reports can trigger intermediary liability under Article 4, thus effectively requiring platforms to remove the content.

In parallel, platforms have also increasingly been automating moderation decisions.[185] This is partly due to technological advances and economic incentives to scale up moderation while saving on labour costs, but is also influenced by pressure from regulators. For example, industry-wide initiatives to coordinate on the automated removal of child sexual abuse material and terrorist content were strongly encouraged by EU policymakers.[186] Automated filtering of copyright-infringing content is now legally required by Article 17 of the 2019 Copyright Directive. While it is not strictly required for other types of content, it is effectively encouraged by Article 5 of the TCR and Article 35 DSA, which require platforms to take proactive measures to reduce the dissemination of terrorist content and to mitigate other systemic risks. Both provisions explicitly mention the use of technical tools to expeditiously remove illegal content as an example measure platforms could take. Given that automated moderation is already in place at major platforms, is easily-scalable and relatively low-cost, and does not require significant changes to their business models or commercial practices, it will be more attractive for them than many other possible measures and will likely be a key part of their compliance efforts.

Platforms use two primary techniques to automatically identify and remove content: hash-matching and AI classifiers. Hash-matching is used to detect and remove exact or near-exact copies of previously removed content. Cryptographic techniques are used to encode a unique identifier for a piece of content, such as an image; if the same image is posted again, it will produce the same code, triggering automatic removal.[187] Hash-matching only works for content which has already been identified and removed by human moderators. For previously-unknown content, platforms use artificial intelligence (AI) classifiers which 'learn' from large datasets of content which was removed to identify new content with similar characteristics.[188] A large majority of content removals on leading platforms are now executed through such automated means.[189] In addition, such tools feed into manual moderation

---

183 Crawford, K., and T. Gillespie, 'What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint', *New Media & Society*, Vol. 18, No. 3, March 2016, pp. 410–428.

184 Crawford, K., and T. Gillespie, 'What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint', *New Media & Society*, Vol. 18, No. 3, March 2016, pp. 410–428.

185 Gorwa, R. et al., 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance', *Big Data & Society*, Vol. 7, No. 1, January 2020.

186 Douek, E., 'The Rise of Content Cartels', Essays and Scholarships, *Knight First Amendment Institute at Columbia University*, February 11, 2020, available at: http://knightcolumbia.org/content/the-rise-of-content-cartels.

187 Gorwa, R. et al., 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance', *Big Data & Society*, Vol. 7, No. 1, January 2020.

188 Gorwa, R. et al., 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance', *Big Data & Society*, Vol. 7, No. 1, January 2020; Gillespie, T., 'Content Moderation, AI, and the Question of Scale', *Big Data & Society*, Vol. 7, No. 2, July 2020; Llansó, E. et al., 'Artificial Intelligence, Content Moderation, and Freedom of Expression', *Transatlantic Working Group*, February 26, 2020.

189 Appelman, N., and P. Leerrssen, 'On "Trusted" Flaggers', Yale-Wikimedia Initiative on Intermediaries & Information, July 12, 2022,
available at: https://law.yale.edu/sites/default/files/area/center/isp/documents/trustedflaggers_ispessayseries_2022.pdf.

processes, as content which is flagged by software as potentially problematic, but without enough certainty to immediately be removed, can instead be reviewed by human moderators.[190]

3.4. Balancing human rights in hate speech moderation

As discussed in Section 3.3, no system for hate speech moderation can be perfect – both because hate speech is an indeterminate and contested concept, so there is no objectively 'correct' way to enforce the rules, and because even where rules are clear, large-scale moderation systems inevitably make technical errors. Thus, a fundamental rights-respecting approach to hate speech moderation cannot be based on the expectation that platforms get every decision right, as traditional case-by-case approaches to fundamental rights suggest.[191] Rather, it must ask whether they are doing enough to keep errors to a minimum, whether they are adequately protecting marginalised users who are most vulnerable to hate speech, and whether all user groups are being treated fairly.

As this section will outline, current moderation practices fall short in all these respects. First, both manual and automated moderation are highly unreliable, so existing rules on hate speech are generally not accurately or effectively enforced. Second, efforts to address hate speech often censor content from marginalised users, preventing them from speaking out to challenge hateful ideas and from participating in social media more generally. Third, moderation systems are primarily designed to focus on identifying and removing individual pieces of content, which limits their capacity to adequately prevent hate speech or support victims. This section explores these issues in more detail and identifies relevant gaps in the EU regulatory framework, which form the basis for the policy recommendations set out in Section 3.6.

3.4.1. Unreliability

Major social media companies do not appear capable of reliably and effectively enforcing their stated policies. Ex-Meta employees have estimated that less than 5% of hate speech[192] on Facebook is removed.[193] This figure is particularly low considering that Facebook is an industry-leading platform owned by one of the world's biggest and wealthiest companies, meaning its moderation capabilities are presumably better than those of most other platforms. There are several obvious steps that could be taken to improve the reliability of content moderation processes, so that where hate speech clearly breaks the law or platforms' policies, it would be more likely to be identified and removed. However,

---

[190] Bellanova, R., and M. de Goede, 'Co-Producing Security: Platform Content Moderation and European Security Integration', *JCMS: Journal of Common Market Studies*, Vol. 60, No. 5, September 2022, pp. 1316–1334.

[191] Llansó et al., 'Artificial Intelligence, Content Moderation, and Freedom of Expression', *Transatlantic Working Group*, February 26, 2020.

[192] Meta's community standards define hate speech as 'a direct attack against people – rather than concepts or institutions – on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. We define attacks as violent or dehumanising speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation' (see Meta, 'Hate speech', Meta Transparency Center, not dated, https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/. This is essentially similar to the definition adopted here based on the AVMSD, although the definition of 'attacks' is broader than just incitement to hatred or violence, and the lists of protected characteristics are slightly different: the AVMSD includes language, property, birth, genetics, age and political opinion, but does not explicitly mention caste or gender identity. However, the definitions are similar enough that the 5% figure can be taken as indicative.

[193] Silverman, C. and R. Mac, 'After The US Election, Key People Are Leaving Facebook And Torching The Company In Departure Notes', *BuzzFeed News*, 2020, available at: https://www.buzzfeednews.com/article/ryanmac/facebook-rules-hate-speech-employees-leaving.

these would generally require significant investments of resources from platforms, and are therefore unlikely to happen in the absence of stronger regulatory incentives than currently exist in EU law.

A first issue – highly relevant for Europe given its linguistic diversity – is that major platforms systematically underinvest in moderation staff and resources for languages other than English. Detailed data on platforms' moderation capabilities and staff for different languages is not available, but leaks from within Meta and Twitter have shown that they lack basic moderation capabilities for even very widely-spoken languages, such as Spanish and most dialects of Arabic.[194] New features designed to improve content moderation and user safety are often only rolled out for a few particularly widely-spoken languages.[195] Journalistic investigations suggest that Meta's resources for European languages like Romanian and Lithuanian are very limited, with the result that obvious hate speech goes unremoved, while speech which does not violate platform policies may be censored.[196]

This is likely also true of other platforms, due to common structural factors. AI language analysis in general[197] and automated hate speech detection in particular[198] remain dominated by English, so research and technological resources (datasets, models, etc.) for other languages are scarce. Social media platforms are also incentivised to invest more in moderation for bigger and wealthier markets, where users are more valuable targets for advertisers (the US being by far the most valuable market[199]). Thus, hate speech in languages other than English is particularly likely to be overlooked by moderation systems. Requiring platforms to invest more in staff and automated moderation tools for other European languages, as well as languages such as Arabic which are widely spoken in Europe, is an obvious way to improve hate speech moderation.

Second, moderators' working conditions are notoriously poor.[200] They are generally employed through outsourcing companies and relatively poorly paid, often in Global South countries with lower labour

---

[194] Scheck et al., 'Facebook Employees Flag Drug Cartels and Human Traffickers. The Company's Response Is Weak, Documents Show.', *Wall Street Journal*, September 16, 2021, available at: https://www.wsj.com/articles/facebook-drug-cartels-human-traffickers-response-is-weak-documents-11631812953#refreshed?mod=series_facebookfiles; Dwoskin et al., 'Twitter can't afford to be one of the world's most influential websites', *Washington Post*, September 4, 2022, available at: https://www.washingtonpost.com/technology/2022/09/04/twitter-mudge-alethea-resources/; Paul, K., 'Disinformation in Spanish Is Prolific on Facebook, Twitter and YouTube despite Vows to Act', *The Guardian*, October 6, 2022.

[195] See e.g. Ghaffary, S., 'Instagram's Surprising Strategy for Bullies: Tell Them to Be Nice', *Vox*, October 20, 2022, available at: https://www.vox.com/recode/2022/10/20/23413581/instagram-nudging-meta-creators-wellbeing-bullying-harassment; Bobrowsky, M., 'Elon Musk Champions Twitter Fact-Checking Feature That Corrects Him', *WSJ*, 2022, available at: https://www.wsj.com/articles/elon-musk-champions-twitter-fact-checking-feature-that-corrects-him-11669436937.

[196] Marinescu, D., 'Facebook's Content Moderation Language Barrier', *New America*, September 8, 2021, available at: https://www.newamerica.org/the-thread/facebooks-content-moderation-language-barrier/; Kayser-Bril, N., 'Facebook's Moderation Is Wreaking Havoc in Lithuanian Public Discourse', *Algorithm Watch*, 2022, available at: https://r.algorithmwatch.org/nl3/21f8gzjqud4-SLZEdp95Ng.

[197] Joshi, P. et al., 'The State and Fate of Linguistic Diversity and Inclusion in the NLP World', *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, 2020, pp. 6282–6293, available at: https://arxiv.org/pdf/2004.09095.pdf.

[198] Vidgen, B. and L. Derczynski, 'Directions in abusive language training data, a systematic review: Garbage in, garbage out', *PLoS ONE*, Vol. 15, No. 12, 2020, Article e0243300.

[199] 'Oversight Board Publishes Policy Advisory Opinion on Meta's Cross-Check Program', *Oversight Board*, 2022, available at: https://oversightboard.com/news/501654971916288-oversight-board-publishes-policy-advisory-opinion-on-meta-s-cross-check-program/.

[200] Roberts, S.T., *Behind the Screen: Content Moderation in the Shadows of Social Media*, Yale University Press, New Haven, 2021; Newton, C., 'The Secret Lives of Facebook Moderators in America', *The Verge*, February 25, 2019, available at: https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona; Newton, C., 'Three Facebook Moderators Break Their NDAs to Expose a Company in Crisis', *The Verge*, June 19, 2019, available at: https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-

costs. Their jobs are fast-paced, closely tracked to enforce performance quotas, and highly stressful: TikTok moderators reportedly have to watch between three and ten videos simultaneously to save time.[201] Often, this work involves viewing graphically violent, hateful and/or distressing content.[202] Moderators have frequently reported developing post-traumatic stress disorder , depression and other mental health problems; this has resulted in multiple class action lawsuits against platform companies for unsafe working conditions.[203]

These labour conditions are, in themselves, a pressing public policy issue. However, as regards the quality of hate speech moderation, it is obviously concerning that the staff responsible for identifying and removing it are poorly paid, have little training, work in extremely stressful and sometimes unsafe conditions, and often come from entirely different cultural and linguistic backgrounds to the users they are moderating. Moderators with little training, psychological support or understanding of particular linguistic and cultural contexts are likely to make frequent errors and to overlook hate speech expressed in subtle or culturally-specific terms. Their stressful and fast-paced conditions are also likely to increase the influence of stereotypes and unconscious bias. For example, snap decisions about whether someone appears violent or threatening tend to be strongly influenced by racial stereotypes.[204]

Third, existing technological approaches used for automated moderation have very limited capacities, even for English-language content. Hash-matching can only identify nearly exact copies of already-removed content. For this purpose it is quite effective, but it can be evaded by informed users,[205] and also frequently removes content that does not violate policies, because it cannot identify uses which are legitimate or harmless when considered in context (such as a critical discussion of unlawful material[206]). AI classifiers – the main way of assessing previously-unknown content – are even more unreliable. Given the complexity and social knowledge required to understand intention and context, even text is still very difficult for AI tools to analyse – especially when the goal is to apply inherently

trauma-ptsd-cognizant-tampa; Ahmad, S., and M. Greb, 'Automating Social Media Content Moderation: Implications for Governance and Labour Discretion', *Work in the Global Economy*, Vol. 2, No. 2, November 2022, pp. 176–198.

201 Vincent, J., 'TikTok Sued by Former Content Moderator for Allegedly Failing to Protect Her Mental Health', *The Verge*, December 24, 2021, available at: https://www.theverge.com/2021/12/24/22852817/tiktok-content-moderation-lawsuit-candie-frazier.

202 Newton, C., 'Three Facebook Moderators Break Their NDAs to Expose a Company in Crisis', *The Verge*, June 19, 2019, available at: https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-trauma-ptsd-cognizant-tampa; McIntyre, N., Bradbury, R., and Perrigo, B., 'The Traumatized Content Moderators Behind TikTok's Boom', *Time*, 2022, available at:

https://time.com/6223340/tiktok-content-moderators-latin-america/.

203 Newton, C., 'Facebook Will Pay $52 Million in Settlement with Moderators Who Developed PTSD on the Job', *The Verge*, May 12, 2020, available at: https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health; Vincent, J., 'TikTok Sued by Former Content Moderator for Allegedly Failing to Protect Her Mental Health', *The Verge*, December 24, 2021, available at:

https://www.theverge.com/2021/12/24/22852817/tiktok-content-moderation-lawsuit-candie-frazier; Reuters, 'Ex-Facebook Moderator in Kenya Sues over Working Conditions', *The Guardian*, May 10, 2022, available at: https://www.theguardian.com/technology/2022/may/10/ex-facebook-moderator-in-kenya-sues-over-working-conditions.

204 Keller, D., 'Daphne Keller and ACLU File Comment to Meta Oversight Board in 'UK Drill Music' Case', *ACLU*, August 23, 2022.

205 Ralton, G., 'Proposed Mechanisms to Detect Illegal Content Can Be Easily Evaded, Study Finds Imperial News Imperial College London', *Imperial News*, 2022, available at: https://www.imperial.ac.uk/news/239291/proposed-mechanisms-detect-illegal-content-easily/.

206 Gillespie, T., Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media, Yale University Press, 2018, p.12.

vague and abstract criteria, such as whether a post incites hatred.[207] Multimedia and especially video content pose even more difficulties.[208] Major platforms still appear to rely to a significant degree on fairly simplistic techniques, like indiscriminately removing content which uses certain keywords.[209] Consequently, innocuous content is often wrongly removed, while hate speech often goes undetected.

While platforms probably could improve these systems to some extent – for example, by investing more in AI tools for underrepresented languages – these problems largely reflect basic limitations of currently-existing technology.[210] For this reason, several authors have argued that reliance on automated content moderation inevitably raises human rights concerns, particularly around the unjustified censorship of 'false positives', and should be kept to a minimum, with content moderation instead undertaken by adequately trained and resourced human staff.[211] This view is supported by the ECJ's recent judgment in *Poland v Parliament and Council*,[212] which held that where automated moderation is mandated by law, it should only be used for content so obviously illegal that it can reliably be identified by software without human intervention.

3.4.2. Discrimination

Unreliable moderation also does not affect everyone equally. Evidence suggests that overinclusive censorship disproportionately affects minority groups, so not only are they exposed to harmful hate speech, but their own fundamental rights to freedom of expression are also restricted.[213] This directly undermines the goal of hate speech moderation, which is to create safe and inclusive online environments.

---

[207] Duarte, E. et al., 'Mixed Messages? The Limits of Automated Social Media Content Analysis', *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 2018, available at: https://cdt.org/wp-content/uploads/2017/12/FAT-conference-draft-2018.pdf.

[208] Zeng, J., and D.B.V. Kaye, 'From Content Moderation to *Visibility Moderation* : A Case Study of Platform Governance on TikTok', *Policy & Internet*, Vol. 14, No. 1, March 2022, pp. 79–95; Thakur, D., and E. Llansó, 'Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis', *Center for Democracy and Technology*, 2022, available at: https://cdt.org/insights/do-you-see-what-i-see-capabilities-and-limits-of-automated-multimedia-content-analysis/; Chowdhury, N., *Automated Content Moderation: A Primer*, Cyber Policy Center, Stanford University, March 19, 2022.

[209] Lux, D., and Lil Miss Hot Mess, 'Facebook's Hate Speech Policies Censor Marginalized Users', Wired, 2017, available at: https://www.wired.com/story/facebooks-hate-speech-policies-censor-marginalized-users/,

Alexander, J., 'YouTube Moderation Bots Punish Videos Tagged as "Gay" or "Lesbian," Study Finds', *The Verge*, September 30, 2019, available at: https://www.theverge.com/2019/9/30/20887614/youtube-moderation-lgbtq-demonetization-terms-words-nerd-city-investigation; Debre, I., and F. Akram, 'Facebook, available at: While Black: Users Call It Getting 'Zucked,' Say Talking about Racism Is Censored as Hate Speech', *USA TODAY*, April 24, 2021.

https://www.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/; Eckert, S., C. Felke, and O. Vitlif, 'TikTok schränkt mit Wortfiltern Meinungsfreiheit ein', *tagesschau.de*, 2022, available at: https://www.tagesschau.de/investigativ/ndr/tik-tok-begriffe-101.html.; Grison, T., and V. Julliard, 'Les Enjeux de La Modération Automatisée Sur Les Réseaux Sociaux Numériques: Les Mobilisations LGBT Contre La Loi Avia', Communication, Technologies et Développement, No. 10, May 20, 2021.

[210] Duarte, E. et al., 'Mixed Messages? The Limits of Automated Social Media Content Analysis', *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 2018, available at: https://cdt.org/wp-content/uploads/2017/12/FAT-conference-draft-2018.pdf;%20 Chowdhury, N., *Automated Content Moderation: A Primer*, Cyber Policy Center, Stanford University, March 19, 2022.

[211] Duarte, E. et al., 'Mixed Messages? The Limits of Automated Social Media Content Analysis', *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 2018, available at: https://cdt.org/wp-content/uploads/2017/12/FAT-conference-draft-2018.pdf; Llansó, E. et al., 'Artificial Intelligence, Content Moderation, and Freedom of Expression', *Transatlantic Working Group*, February 26, 2020.

[212] Republic of Poland v European Parliament and Council of the European Union, ECJ 2022.

[213] Griffin, R., 'The Sanitised Platform', *JIPITEC*, Vol. 3, No.1, 2022, available at: https://www.jipitec.eu/issues/jipitec-13-1-2022/5514/citation.

Both automated and manual moderation processes often exhibit bias against marginalised groups, meaning that their content will more often mistakenly be identified as hate speech. As discussed in the previous subsection, human moderators will inevitably hold (conscious or unconscious) biases and stereotypes – which are particularly likely to influence their decisions given their fast-paced and stressful working conditions. Algorithmic bias - a broad term for algorithmic systems which produce disparate or unfair outcomes for different social groups - is also pervasive. For example, image- and facial-recognition software tends to be significantly less accurate for women and racial minorities.[214] Language analysis is also less accurate for minority groups whose dialects and slang terms are typically underrepresented in training data, and may be associated with aggression or 'inappropriate' behaviour due to social stigma.[215] For example, experiments have shown that Google's widely-used Perspective moderation software (which it offers commercially to other companies) more often flags as 'toxic' comments which identify the speaker as Black or gay.[216]

The inability to assess content in context also makes it particularly likely that content from marginalised users will falsely be classed as hate speech. For example, content is often classed as aggressive or hateful based on the use of certain keywords. This is problematic because many marginalised communities (such as drag performers and other LGBTQ+ subcultures) commonly use slurs or insults in a positive, reclaimed sense, and are thus disproportionately likely to be mistakenly censored.[217] Similarly, researchers and journalists have documented many cases in which marginalised users who attempt to discuss their experiences of racism or sexism are censored, probably because content which openly discusses hateful ideas or behaviour is prone to being falsely classified as hate speech.[218] Box 2 provides an overview of auditing techniques as a means of identifying and tackling bias in automated moderation, and suggests how they could fit into the EU framework for social media governance.

---

[214] Monea, A., Digital Closet: How The Internet Became Straight, MIT Press, S.L., 2023.

[215] Park et al., 'Reducing Gender Bias in Abusive Language Detection', Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2799–2804; Sap et al., 'The Risk of Racial Bias in Hate Speech Detection', Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1668–1678; Duarte et al., 'Mixed Messages? The Limits of Automated Social Media Content Analysis', Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 2018, available at: https://cdt.org/wp-content/uploads/2017/12/FAT-conference-draft-2018.pdf; Talat, Z., '"It ain't all good:" Machinic abuse detection and marginalisation in machine learning', PhD thesis, University of Sheffield, 2021; Batorsky, B., 'New Technology, Old Problems: The Missing Voices in Natural Language Processing', The Gradient, April 9, 2022, available at: https://thegradient.pub/nlp-new-old/.

[216] Kayser-Bril, N., 'Automated Moderation Tool from Google Rates People of Color and Gays as "Toxic', AlgorithmWatch, 2020,available at: https://algorithmwatch.org/en/automated-moderation-perspective-bias/; Talat, Z., "It ain't all good:" Machinic abuse detection and marginalisation in machine learning', PhD thesis, University of Sheffield, 2021, available at: https://etheses.whiterose.ac.uk/30950/.

[217] Lux, D., and Lil Miss Hot Mess, 'Facebook's Hate Speech Policies Censor Marginalized Users', Wired, 2017, available at: https://www.wired.com/story/facebooks-hate-speech-policies-censor-marginalized-users/; Haimson et al., 'Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas', Proceedings of the ACM on Human-Computer Interaction, Vol. 5, No. CSCW2, October 13, 2021, pp. 1–35; Dias Oliva et al., 'Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online', Sexuality & Culture, Vol. 25, No. 2, April 2021, pp. 700–732; Grison, T., and V. Julliard, 'Les Enjeux de La Modération Automatisée Sur Les Réseaux Sociaux Numériques : Les Mobilisations LGBT Contre La Loi Avia', Communication, Technologies et Développement, No. 10, May 20, 2021.

[218] Guynn, J., 'Facebook While Black: Users Call It Getting 'Zucked,' Say Talking about Racism Is Censored as Hate Speech', USA TODAY, 2019, available at: https://www.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/; Nurik, C., ''Men Are Scum': Self-Regulation, Hate Speech, and Gender-Based Censorship on Facebook', International journal of communication, June 2019; Gray, K.L., and K. Stein, "We 'Said Her Name' and Got Zucked": Black Women Calling-out the Carceral Logics of Digital Platforms', Gender & Society, Vol. 35, No. 4, August 2021, pp. 538–545.

Box 2: Auditing algorithmic moderation

A recent study by the US-based NGO Algorithmic Justice League finds that although consensus exists amongst auditors and other stakeholders (i.e., regulators, independent researchers, etc.) as to the need for increased auditing and regulation to prevent algorithmic bias and discrimination, debates remain as to how audits and regulations should be administered. In particular, there are disagreements as to auditing methodologies, the appropriate degree of transparency vis-à-vis the public, and whether (and how) auditing techniques should be standardised and overseen by regulators. The DSA requires platforms to take mitigation measures against systemic risks to fundamental rights including non-discrimination and have these measures independently audited. The vague language of certain provisions creates uncertainty, but can also enable regulation to evolve and become more detailed via judicial decisions and supplementary legal action by the Commission. For auditing under the DSA to be effective, clearer and more specific auditing standards will need to be developed.

A main focus of the discussion questions the limitations of the current tech-focused approach to AI audits. Most auditing of algorithmic moderation (AI auditing) 'focus[es] on technical implementation of principles' , and is as such largely quantitative. This approach has two downsides. First of all, by failing to examine how the AI system functions within its socioeconomic context, audits fail to acknowledge and address real-life discriminatory impacts. This, of course, diminishes the effectiveness of the audit process as a whole. This is of particular relevance within the social media context, due to the harm engendered by automated moderation to the freedom of speech of users from marginalised communities. Secondly, by adopting a narrow approach focused on technical debiasing measures, auditors (and the legislation that regulates them) risk 'divert[ing] important political questions into the realm of the technical', and thereby concentrating regulation of matters such as hate-speech and discrimination in the hands of technology companies as opposed to legislators.

This leads to a second major point of discussion, which calls for a greater reliance on real-life qualitative inputs in the AI auditing process. In particular, this could involve greater participation by affected communities in auditing algorithmic moderation, also known as stakeholder involvement. Stakeholder involvement is particularly important because it can rectify knowledge deficiencies by providing information that auditors would otherwise lack. For instance, an auditing platform devised by graduate fellows of Stanford University's Human-Centred Artificial Intelligence found that stakeholders were able to address issues such as the overflagging of terms that were originally used as slurs, which have now been reappropriated by marginalised communities. Unfortunately, due to cost-related and client-confidentiality (transparency) issues, few audits currently employ stakeholder involvement. This is something the Commission, as well as the Board and national regulators, could look into promoting and incentivising, and potentially mandating through supplementary rules on the performance of audits (as permitted under Article 37(7) DSA).

In short, though the DSA addresses the need for mandated, compulsory audits by very large online platforms and search engines, in its current state it falls short of developing clear and specific standards for the performance of audits and auditing methodologies. This absence of clear standards risks perpetuating the discrimination and inconsistency in algorithmic moderation that we see today.

Sources: Costanza-Chock et al., 'Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem', *2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, Association for Computing Machinery, New York, NY, USA, 2022, pp. 1571–1583; Balayne, A. and Gürses, S., 'If AI Is the Problem, Is Debiasing the Solution?',

European Digital Rights (EDRi), 2021 , available at: EDRi 2-21; Miller, K., 'A DIY Approach to Algorithmic Audits', Stanford HAI, 2022, available at: https://hai.stanford.edu/news/diy-approach-algorithmic-audits.

3.4.3. Limited protection for victims

Where hate content is not identified through automated moderation software, the primary way it can come to the attention of human moderators is through user reporting. The option to report content to the platform is important and will be useful for some people who see or are targeted by hate speech. However, it will be inadequate or unhelpful for many others, and is limited as a way of addressing hate speech and abusive behaviour.

First, many instances of hate speech will inevitably go unreported. This could be because the individuals who are targeted do not want to report it, for example because they find the process laborious or emotionally draining; because it does not target any particular individual; or because it is primarily seen by sympathetic audiences, which is particularly likely where hate speech is shared within closed groups or by channels/accounts primarily followed by people who share their views.[219]

Moreover, research shows that hate speech, abuse and harassment on social media often involve coordinated, networked activity from many users.[220] These situations will be particularly harmful and threatening for victims, and are particularly likely to target marginalised users.[221] However, they are difficult to address through reporting: reporting each individual piece of content is impractical and emotionally burdensome, and moderators reviewing posts individually may not identify them as hate speech, since the full context will not be apparent.[222]

Box 3 discusses the use of behavioural prompts as an alternative means of preventing hate speech and harassment, which does not rely on user reporting.

[219] Crawford, K., and T. Gillespie, 'What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint', *New Media & Society*, Vol. 18, No. 3, March 2016, pp. 410–428.

[220] Jeong, S., 'The Internet of Garbage by Sarah Jeong', *The Verge*, August 28, 2018, available at: https://www.theverge.com/2018/8/28/17777330/internet-of-garbage-book-sarah-jeong-online-harassment; Khoo, C., *Deplatforming Misogyny: Report on Platform Liability for Technology-Facilitated Gender-Based Violence*, Women's Legal Education and Action Fund (LEAF), 2021, available at: https://www.leaf.ca/wp-content/uploads/2021/04/Full-Report-Deplatforming-Misogyny.pdf; Marwick, A.E., 'Morally Motivated Networked Harassment as Normative Reinforcement', *Social Media + Society*, Vol. 7, No. 2, April 2021.

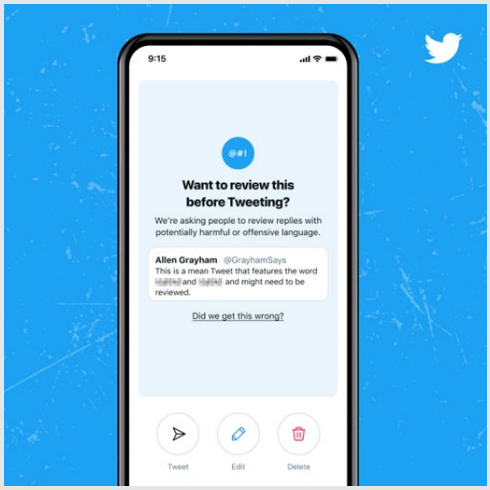
[221] Chemaly, S., 'Demographics, Design, and Free Speech: How Demographics Have Produced Social Media Optimized for Abuse and the Silencing of Marginalized Voices', by S. Chemaly, Free Speech in the Digital Age, *Oxford University Press*, 2019, pp. 150–169, available at: https://academic.oup.com/book/27505/chapter-abstract/197448652?redirectedFrom=fulltext

[222] Duguay et al., 'Queer Women's Experiences of Patchwork Platform Governance on Tinder, Instagram, and Vine', *Convergence: The International Journal of Research into New Media Technologies*, Vol. 26, No. 2, April 2020, pp. 237–252; Griffin, R., 'New School Speech Regulation as a Regulatory Strategy against Hate Speech on Social Media: The Case of Germany's NetzDG', *Telecommunications Policy*, Vol. 46, No. 9, October 2022.

Box 3: Behavioural prompts and nudges

One design intervention which appears relatively effective in influencing user behaviour, without significantly restricting their freedom of expression, is the use of behavioural prompts. These were first introduced by Twitter in 2021, with a system that would automatically scan tweets for language which could be seen as aggressive. Instead of blocking or deleting these tweets, the system would just show users a prompt asking them if they wanted to reconsider before posting. According to peer-reviewed research by Twitter's internal researchers, these prompts lead around a third of users to delete or rewrite their tweet, mostly (though not always) to make it less offensive. Since then, Instagram has introduced a similar feature. Notably, Instagram's prompts target not only comments flagged as offensive, but replies to such comments – an attempt to address networked abuse, where users may participate in or encourage harassing behaviour without using explicitly hateful or abusive language themselves.

Figure 1: Twitter prompt



Source: Katsaros, M. et al., 'Reconsidering Tweets: Intervening During Tweet Creation Decreases Offensive Content', International AAAI Conference on Web and Social Media, 2022, https://ojs.aaai.org/index.php/ICWSM/article/view/19308/19080

Twitter originally only rolled out these prompts for English-language users, later expanding them to Portuguese. Since Elon Musk's takeover of the platform in late 2022, the company's entire AI ethics team and many of its security and moderation staff have been fired. As a result, although safety features like these appear promising, it seems doubtful that they will continue to be developed and rolled out in more languages and markets. EU policymakers should aim to create stronger regulatory requirements or incentives, to ensure that major platforms systematically develop, test and share research on design interventions like these and invest in promising safety measures in all markets and languages. Potential regulatory levers are discussed further in Section 3.5 and Box 5.

Sources: Katsaros, M. et al., 'Reconsidering Tweets: Intervening During Tweet Creation Decreases Offensive Content', *International AAAI Conference on Web and Social Media*, 2022, https://ojs.aaai.org/index.php/ICWSM/article/view/19308/19080; Ghaffary, S., 'Instagram's Surprising Strategy for Bullies: Tell Them to Be Nice', *Vox*, October 20, 2022. https://www.vox.com/recode/2022/10/20/23413581/instagram-nudging-meta-creators-wellbeing-bullying-harassment; Butler, A., and A. Parella, 'Tweeting with Consideration', Twitter, 2021. https://blog.twitter.com/en_us/topics/product/2021/tweeting-with-consideration; Newton C., and Z. Schiffer, 'Twitter, cut in half', *Platformer News*, 2022.

3.5. The EU legal framework

The primary EU instruments addressing online hate speech – the 2016 Code of Conduct on Hate Speech, the TCR, and the DSA – unfortunately do not adequately engage with the three key problems discussed above in Section 3.4. The developing EU legal regime effectively encourages platforms to increase reliance on automated moderation, which will exacerbate problems of unreliability and bias. The DSA's provisions regulating content moderation – particularly Article 16, which requires platforms to implement user reporting systems and promptly review reported content – also entrenches an approach to moderation which is focused on reviewing individual pieces of content that break the rules, rather than addressing hate speech and harmful content in its context.[223] Such an approach fails to protect victims from the most harmful forms of networked harassment, and does not incentivise platforms to work on more proactive, design-based interventions to discourage hate speech and harmful behaviour.

Given the limitations and bias of automated moderation, improving the quality of moderation by increasing the number of moderators and improving their training and working conditions should be a priority. However, moderators should still be supported by automated tools: for example, to identify priority content for human review, and to reduce the need for human moderators to view the most harmful or upsetting content.[224] Thus, platforms should also prioritise addressing algorithmic bias and language disparities in such tools.

The DSA does establish some very general requirements about the staffing and resources of moderation teams. For example, Article 20 requires challenges to moderation decisions to be reviewed by 'appropriately qualified staff'; appropriate qualifications would presumably include, for example, speaking whatever languages are involved. However, this only applies to reviews, not initial moderation decisions. The DSA does not establish any concrete or detailed obligations regarding staffing, resources and working conditions. This is a missed opportunity, since mandating better conditions for moderators, increased headcounts, and improved training and technical resources would simultaneously protect workers' rights and improve the effectiveness and reliability of moderation of hate speech (and other types of illegal content). Box 4 discusses Germany's Network Enforcement Act as an example of how legislators can regulate such issues.

[223] Douek, E., 'Content Moderation as Systems Thinking', *Harvard Law Review*, Vol. 136, No. 2, December 2022, pp. 526–607.

[224] Gillespie, T., 'Content Moderation, AI, and the Question of Scale', *Big Data & Society*, Vol. 7, No. 2, July 2020.

Box 4: Staffing and resources for content moderation under Germany's NetzDG

In 2017, Germany passed the Network Enforcement Act (NetzDG), a law which aimed to concretise and strengthen the enforcement of major platforms' notice and takedown obligations under the ECD. It supplemented the liability obligations in the ECD with additional obligations which did not require the removal of individual pieces of content, but instead focused on the organisation, staffing and overall standards of platforms' internal content moderation systems. The influence of this approach can be seen in the DSA's system of due diligence obligations. However, in some respects NetzDG created stronger and more specific obligations than the DSA.

NetzDG can be criticised for its narrow focus on reactive reporting and removal of individual pieces of illegal content. It largely ignores more systemic issues and alternative interventions, like those discussed in Section 3.4. How much it has achieved in terms of reducing online hate speech and protecting victims remains debated. However, it has undoubtedly had an impact in some areas – notably in relation to the staffing and training of moderation teams. Article 3 NetzDG sets out procedures platforms must follow when users report potentially-illegal content. Besides mandating them to promptly examine and respond to complaints (generally within 24 hours), it also requires them to ensure that complaints are reviewed and handled by staff who have appropriate training, German language skills and subject-matter expertise, including in German law. After NetzDG came into effect, leading platforms significantly increased their numbers of moderation staff in Germany, compared to other countries.

Independent research has shown that moderators in Germany still face economic insecurity, intense time pressure and highly routinised work processes which do not allow careful consideration of content or flexible responses. However, the German experience shows that regulating staffing and training can achieve concrete improvements in platforms' moderation operations, without simply demanding stricter policies or more censorship. This suggests that it would be possible for the DSA to establish more demanding requirements regarding moderators' numbers, training and working conditions – for example, as a new category of due diligence obligation for very large online platforms. These should go further than NetzDG and place more emphasis on ensuring that moderators have adequate support and technical resources, and improving their working conditions – which should in turn lead to better-quality and more context-sensitive decisions. Member States should consider similar legislation to regulate the staffing and working conditions of content moderation staff based in the relevant Member State and/or moderating content from that Member State.

Sources: Helberger, N., 'The Political Power of Platforms: How Current Attempts to Regulate Misinformation Amplify Opinion Power', *Digital Journalism, Vol. 8, No. 6*, 2020, pp. 842–854; Griffin, R., 'New School Speech Regulation as a Regulatory Strategy against Hate Speech on Social Media: The Case of Germany's NetzDG', *Telecommunications Policy, Vol. 46, No. 9*, October 2022; Oltermann, P., 'Tough New German Law Puts Tech Firms and Free Speech in Spotlight', *The Guardian*, January 5, 2018, sec. World news.; Heldt, A., 'Reading between the Lines and the Numbers: An Analysis of the First NetzDG Reports', *Internet Policy Review, Vol. 8, No. 2*, June 12, 2019; Ahmad, S., and M. Greb, 'Automating Social Media Content Moderation: Implications for Governance and Labour Discretion', *Work in the Global Economy, Vol. 2, No. 2*, November 2022, pp. 176–198.

Finally, given the inherent difficulties of effectively moderating hate speech content at scale without excessive censorship, greater attention should be given to 'human rights by design' approaches, which aim to design online spaces in ways that discourage hate speech and abuse, while providing support for affected users.[225] Scholars have linked platforms' current failings on hate speech to the perspectives

[225] Suzor et al., 'Human Rights by Design: The Responsibilities of Social Media Platforms to Address Gender-Based Violence Online: Gender-Based Violence Online', Policy & Internet, Vol. 11, No. 1, March 2019, pp. 84–103; Chemaly, S., 'Demographics, Design, and Free Speech: How Demographics Have Produced Social Media Optimized for Abuse and the

of those designing and running them, suggesting that predominantly straight white male executives and engineers fail to take account of the experiences of groups most vulnerable to hate speech.[226] Thus, proactive design interventions against hate speech and abuse would particularly benefit from drawing on ideas of 'design justice', which hold that technologies should be designed with the participation of marginalised groups to serve their specific needs.[227]

As will be highlighted in Box 8, some design interventions that platforms have already deployed to tackle hate speech seem promising. However, such measures are unlikely to be pursued and tested consistently and systematically if platforms do not have concrete legal incentives to do so. This is well illustrated by the case of Twitter, where the company's entire ethical AI team, which had been responsible for developing new safety and accountability measures, as well as many of its trust and safety experts and moderators, were recently fired after Elon Musk's takeover of the company.[228] Voluntary efforts to research and develop new safety measures for online public spaces cannot be relied on if they are always subject to a change in business strategy. Concrete legal obligations are thus essential.

Unfortunately, proactive interventions to discourage hate speech and improve user safety are largely overlooked in the DSA. While they could in principle be part of VLOPs' risk mitigation measures, required by Article 35, platforms have a lot of freedom to determine how they interpret these obligations and what measures they take. As even complying with the minimal requirements of the DSA will be quite resource-intensive,[229] they are unlikely to invest significant additional resources in measures like hiring and training staff to support victims of hate speech or researching and implementing product changes.

However, the Commission has the opportunity to influence platforms' responses through its oversight role, and in particular through the industry Codes of Conduct, which will be established under Article 45 to further concretise VLOPs' risk mitigation obligations.[230] A positive first step to strengthen the protection of fundamental rights on social media would be ensuring that these codes clearly mandate platforms to consider hate speech, abuse and other risks to marginalised groups in their product design and policy processes; to proactively develop and test interventions to reduce these risks; and to involve users and stakeholder groups from affected communities in these processes. As Box 5 discusses, the Code of Practice on Disinformation provides an example of how this can be achieved. Regulators should make it a priority to hire staff with expertise in UX/UI design (the design of technical interfaces and features which shape user experiences on a platform), as well as digital justice and the specific risks

Silencing of Marginalized Voices', by S. Chemaly, Free Speech in the Digital Age, Oxford University Press, 2019, pp. 150–169, available at: https://academic.oup.com/book/27505/chapter-abstract/197448652?redirectedFrom=fulltext.

[226] Chemaly, S., 'Demographics, Design, and Free Speech: How Demographics Have Produced Social Media Optimized for Abuse and the Silencing of Marginalized Voices', by S. Chemaly, Free Speech in the Digital Age, Oxford University Press, 2019, pp. 150–169,
available at: https://academic.oup.com/book/27505/chapter-abstract/197448652?redirectedFrom=fulltext.

[227] Costanza-Chock, S., Design Justice: Community-Led Practices to Build the Worlds We Need, Information Policy, The MIT Press, Cambridge, Massachusetts, 2020.

[228] Newton C., and Z. Schiffer, 'Twitter, cut in half', Platformer News, 2022.

[229] Keller, D., 'Facebook Filters, Fundamental Rights, and the CJEU's Glawischnig-Piesczek Ruling', *GRUR International*, Vol. 69, No. 6, June 1, 2020, pp. 616–623.

[230] Vander Maelen, C., 'Hardly Law or Hard Law? Investigating the Dimensions of Functionality and Legalisation of Codes of Conduct in Recent EU Legislation and the Normative Repercussions Thereof', *European Law Review*, Vol. 47, No. 6, 2022, pp. 752–772.

faced by marginalised users of online services, in order to effectively enforce these obligations and develop concrete best practices and industry standards.[231]

Box 5: 'Safe design' in the Code of Practice on Disinformation

The 2022 updated CoP on Disinformation contains numerous commitments intended to strengthen transparency and accountability in relation to advertising, security and other aspects of disinformation policy, discussed in detail in Chapter 4 on disinformation. However, one notable aspect could serve as a best practice in relation to other policy areas, such as hate speech.

Commitment 18 requires platforms to 'minimise the risks of viral propagation of Disinformation by adopting safe design practices as they develop their systems, policies, and features' – for example, by adapting their interfaces and recommendation systems and by pre-testing new products and features. In itself, this largely reiterates the risk assessment and mitigation obligations that very large online platforms already face under Articles 34-35 DSA. However, Measure 18.3 provides additional detail which could strengthen these obligations. Specifically, signatories commit to:

- invest and participate in researching safe design practices in relation to disinformation

- publish their findings and report on them to the industry taskforce which will oversee the implementation of the Code

- explain to the taskforce how they are using or plan to use these findings to improve the design of their platforms

- where possible, publish the amount of their financial investments in these research activities

Overall, all of these obligations should make platforms' risk mitigation obligations much more demanding. In theory, companies cannot just publish risk assessments and highlight some superficial risk mitigation measures as a formality – they will have to show that they are actively investing in researching safe design practices. This research will be open to scrutiny and improvement from the broader research community, and platforms will be under pressure to show that they are adapting their services accordingly.

For example, regulators could question why platforms are not investing more into researching, expanding and improving behavioural prompts like those described in Box 8; why Twitter has shared so little information publicly about their functioning and success metrics; and why – if they have proved effective – they have only been rolled out in a few markets. Failure to improve in these areas could be classed as a failure to comply with Twitter's commitments under the Code, and ultimately with its risk mitigation obligations under Article 35 DSA (since the Code serves as guidance on the interpretation of these obligations).

Besides taking this opportunity to push for better design practices which could address systemic factors contributing to disinformation, EU policymakers should follow this approach in other areas. Further co-regulatory codes of conduct should be developed to concretise platforms' risk mitigation obligations in relation to hate speech, harassment, privacy violations and other abusive behaviours which tend to exclude marginalised users. These should draw on principles of design justice, requiring platforms not only to invest more resources in researching how to create safer

[231] Pershan, C., and C. Sinders, 'Why Europe's Digital Services Act Regulators Need Design Expertise', Tech Policy Press, December 12, 2022, available at: https://techpolicy.press/why-europes-digital-services-act-regulators-need-design-expertise/.

environments for marginalised groups, but also to actively involve users and advocacy organisations from these communities in researching and developing new safe design practices.

Sources: Authors' own elaboration, based on the Strengthened Code of Practice on Disinformation 2022.

3.6. Recommendations

a. A new Code of Conduct on Online Hate Speech

- In order to strengthen and concretise very large online platforms' obligations to mitigate systemic risks under the DSA, the Commission should take the lead on establishing multistakeholder discussions to update and expand the 2016 Code of Conduct on Hate Speech and Harassment.
- These discussions should include a diverse range of independent researchers and civil society organisations from all over Europe. Representing marginalised communities such as Roma people, LGBTQ+ people and migrants should be a top priority in convening these discussions. Funding should be available to support participation by organisations who may otherwise lack the resources.

  As a starting point, the new Code of conduct should:

  - Establish a broader definition of online hate speech as incitement to hatred or violence based on any characteristic protected by Article 21 of the Charter of Fundamental Rights. To recognise intersectional forms of marginalisation, this should also extend to combinations of characteristics where any one of those characteristics is protected by the Charter.
  - Broaden the scope of platforms' obligations beyond hate speech. Platforms should additionally commit to tackle all forms of threats, harassment and privacy violations which target a person or group based on a protected characteristic.
  - Require platforms to establish adequate moderation staff and technical resources for all languages which are widely spoken in markets where they operate, and to publish detailed reports on their moderation capabilities in all such languages.
  - Establish clear and specific commitments from platforms to investigate, develop and test proactive measures (including design changes) to discourage hate speech and support affected users. Platforms should also commit to ongoing consultation and participation from stakeholder groups representing affected communities as part of these processes.
  - Establish clear and specific standards on the working conditions (e.g. pay, training, performance quotas, working hours, psychological support) of all platform staff working on content moderation. These should also apply to staff working on behalf of a platform via outsourcing companies, and staff based outside the EU.

b. DSA enforcement

- The Commission and national digital services coordinators (DSCs) should issue guidance stating that, in accordance with the ECJ decision in *Poland v Parliament and Council* [2022], platforms' obligations to have due regard to fundamental rights (under Article 14(4) DSA and Article 5 TCR) and very large online platforms' obligations to address systemic risks to fundamental rights (under Articles 34-35 DSA) preclude the use of automated moderation tools which are indiscriminate (make high proportions of errors) or clearly discriminatory (disproportionately censor users from marginalised groups). The guidance should further state that platforms must clearly document the design, use, performance and outcomes of such

tools, including industry-standard accuracy and bias metrics, to establish regulatory compliance.

- The Commission and national DSCs should also issue guidance stating that the obligation for platforms to enforce their content policies in a diligent, objective and proportionate manner under Article 14(4) DSA requires adequate moderation capacities in all languages widely spoken by their users, including adequate investment in competent moderation staff. All relevant moderation processes should be clearly and publicly documented to establish compliance.
- In overseeing and enforcing very large platforms' systemic risk mitigation obligations under Articles 34-35 DSA, the Commission should place significant weight on design changes and other interventions which aim to proactively discourage and prevent the occurrence of online hate speech, harassment and other systemic risks, as opposed to moderating or removing content retroactively. Risk assessments and audit reports which indicate that platforms are not investing in such proactive risk mitigation measures should not be regarded as compliant.
- The Commission and national regulators should ensure that they have sufficient staff with relevant technical and UX/UI design expertise to effectively assess compliance with these obligations. This would also be aided by effective procedures for collaboration, co-investigations and knowledge sharing between different regulatory agencies.

c. Legislative reform

- The Commission should consider and consult with Member States, civil society and other relevant stakeholders on proposing EU-level legislation to regulate the staffing and operation of platforms' content moderation teams. This could include:

  - Minimum thresholds for numbers of staff with relevant language and market expertise for each EU country in which a platform operates;
  - Regulation of the working conditions (e.g. training, performance quotas, working hours, psychological support) of content moderation staff.

4. DISINFORMATION

4.1. Introduction

This chapter addresses the implications of mis- and disinformation on social media for the rule of law, fundamental rights, and democracy. The chapter aims to highlight the most important findings and implications for fundamental rights and democracy, and to analyse recent EU regulatory reforms and other recent developments. It also refers to several relevant literature reviews and scoping papers which offer useful resources for further investigation.

The chapter proceeds as follows. Section 4.2 provides some necessary context and background: it first defines the concepts of dis- and misinformation and the scope of this chapter, then briefly discusses relevant human rights principles and case law. Section 4.3 offers a brief overview of relevant empirical research around disinformation on social media. On this basis, it identifies normative and policy concerns that EU disinformation policy should seek to address, focusing in particular on specific types of content with the capacity to cause direct harm (e.g., dangerous health misinformation), disinformation targeting marginalised social groups, and broader implications for trust in media and political institutions. Section 4.4 outlines major social media platforms' principal responses to disinformation to date, most importantly content moderation, fact-checking partnerships, and transparency initiatives. Section 4.5 discusses regulatory responses to disinformation in Europe: notably the DSA, the CoP on Disinformation, and the proposed Political Advertising Regulation. It assesses the effectiveness of these existing and proposed initiatives and highlights the concerns they raise with regard to democracy and fundamental rights. Finally, Section 4.6 concludes with a discussion of potential paths forward and policy recommendations.

4.2. Background

4.2.1. Definitions

A widely-accepted definition of disinformation has been provided by an EU High-Level Expert Group (HLEG) on Fake News and Online Disinformation report: 'disinformation…includes all forms of false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit'[232]. This definition has since been widely used in academic and policy literature.[233] Disinformation as defined by the HLEG excludes forms of speech which are already illegal, such as hate speech (see Chapter 3), defamation, and incitement to violence. Instead, it focuses on legal speech

[232] De Cock Bunning, M., A Multi-Dimensional Approach to Disinformation: Report of the Independent High Level Group on Fake News and Online Disinformation, Publications Office of the European Union, 2018, available at: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50271.

[233] Freelon, D., and Wells C., 'Disinformation as Political Communication', Political Communication, Vol. 37, No. 2, March 3, 2020, pp. 145–156; Alaphilippe, A., et al., Automated Tackling of Disinformation: Major Challenges Ahead, European Parliament, Brussels, 2019, available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624278/EPRS_STU(2019)624278_EN.pdf; Strand, C., et. al, Disinformation Campaigns about LGBTI+ People in the EU and Foreign Influence: Briefing, European Parliament, Brussels, 2021, available at: https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/653644/EXPO_BRI(2021)653644_EN.pdf ; Szakàcs J., Bognàr E., European Parliament. Directorate General for External Policies of the Union., The Impact of Disinformation Campaigns about Migrants and Minority Groups in the EU: In Depth Analysis., Publications Office, Brussels, 2021, available at: https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/INGE/DV/2021/07-12/IDADisinformation_migrant_minorities_EN.pdf.

which can nonetheless be harmful.[234] It effectively establishes three criteria to identify disinformation: false or misleading information, potential for public harm or profit, and the intention to cause public harm or profit.[235]

The third criterion distinguishes disinformation from misinformation, as defined in a widely cited typology by Wardle and Derakhshan: disinformation is spread intentionally to cause harm or for profit, while misinformation is spread without this intention.[236] However, these distinct concepts may be closely linked or overlap in practice, for example, when disinformation is subsequently disseminated further by people who believe it (which would typically be expected, if a disinformation operation is successful). This chapter will thus discuss both mis- and disinformation, as a successful policy response must address both.

The HLEG's definition of disinformation has been criticised for being overly broad and vague, raising fundamental rights concerns due to the possibility of excessive censorship of information deemed false by platforms or political figures.[237] The salience of these criticisms depends on how and in what context the definition is used. Importantly, the HLEG explains that disinformation is 'a problem that must be understood in the wider context of how information is produced, how it is distributed, and how people engage with it in the public sphere',[238] and explicitly recommended against censorship of disinformation. Instead, it emphasised the need for broader policy interventions aiming to create a healthy media environment and strengthen trust in reliable media institutions. Along these lines, researchers from the University of Amsterdam's Institute for Information Law have argued that disinformation as defined by the HLEG is 'not fit to function as a legal category' but should rather be used to indicate a policy area.[239]

In popular culture and political discourse, the term 'fake news' is widely used as an umbrella term to indicate any form of dis- or misinformation. It has been criticised by academics and policy researchers for its imprecision, as it includes low-risk forms of speech such as honest mistakes, parodies and partisan political discourse, as well as malicious fabrications, amongst others. It has also been widely used by political figures to discredit opponents and journalists.[240] It will accordingly not be used in this study.

---

[234] De Cock Bunning, M., A Multi-Dimensional Approach to Disinformation: Report of the Independent High Level Group on Fake News and Online Disinformation, Publications Office of the European Union, 2018. p. 12, available at: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50271.

[235] Freelon, D., and Wells, C., 'Disinformation as Political Communication', Political Communication, Vol. 37, No. 2, March 3, 2020, pp. 145–156.

[236] Wardle, C., and Derakhshan, H., Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making, Council of Europe, Strasbourg, 2017, available at: https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c.

[237] Ó Fathaigh, R., et. al., 'The Perils of Legally Defining Disinformation', Internet Policy Review, Vol. 10, No. 4, November 4, 2021.

[238] De Cock Bunning, M., A Multi-Dimensional Approach to Disinformation: Report of the Independent High Level Group on Fake News and Online Disinformation, Publications Office of the European Union, 2018. p. 13, available at: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50271.

[239] Ó Fathaigh, R., et. al., 'The Perils of Legally Defining Disinformation', Internet Policy Review, Vol. 10, No. 4, November 4, 2021.

[240] Alaphilippe, A., et. al., Automated Tackling of Disinformation: Major Challenges Ahead, European Parliament, Brussels, 2019,
available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624278/EPRS_STU(2019)624278_EN.pdf; Freelon, D., and Wells, C., 'Disinformation as Political Communication', Political Communication, Vol. 37, No. 2, March 3, 2020, pp. 145–156.

4.2.2. Disinformation and fundamental rights

As a form of speech, dis- and misinformation directly implicate the right to freedom of speech and information, protected by Article 11 of the European Charter of Fundamental Rights and recognised as an essential foundation of democratic society. In addition, restrictions on dissemination of disinformation raise broader concerns around the rule of law and democratic debate, as they can allow governments to suppress criticism and dissent.

First, as with hate speech, discussed in detail in Chapter 3, available tools to moderate and remove disinformation content are inevitably imperfect and overinclusive. This means that harmless speech and legitimate contributions to political and social debates will also often be censored. In addition, moderation systems also often exhibit bias against minority groups. Thus, overinclusive censorship may disproportionately affect such groups, which not only affects individuals' fundamental rights but also raises broader concerns around equal participation in democratic processes.

Second, regulatory regimes empowering states and platforms to censor content deemed to be false and harmful – both vague and easily manipulated criteria – create obvious potential for politically motivated abuse and biased application, which could skew democratic debate and restrict media freedom. As Section 4.5 will discuss, this is especially relevant given the avenues for informal state intervention created by the DSA's notice and takedown regime. Even well-intentioned disinformation policies can be highly concerning in this regard, if they allow powerful actors - whether governments or platforms - to enforce their own definition of the 'truth' in relation to political and social issues which are often contested and uncertain.

In-depth discussions of the relevant case law on freedom of expression can be found in the 2021 study on disinformation and freedom of expression,[241] and in recent academic studies.[242] It is here relevant to note that (although the European and international case law has not yet addressed legal questions specifically related to online disinformation) false and misleading information is clearly protected by international human rights law and European fundamental rights standards on freedom of expression and information. Indeed, though the ECJ's media jurisprudence is underdeveloped and often relies on ECtHR case law for more detailed guidance, the ECJ has recognised the importance of the right to freedom of expression, including for expressions that 'offend, shock, or disturb.' Thus, any limitations to this right must be interpreted restrictively, and require particular consideration.[243] The ECtHR tends to interpret freedom of expression broadly. For example, in *Salov v. Ukraine*, the ECtHR held that: 'Article 10 of the Convention as such does not prohibit discussion or dissemination of information received even if it is strongly suspected that this information might not be truthful. To suggest otherwise would deprive persons of the right to express their views and opinions about statements made in the mass media, and would thus place an unreasonable restriction on the freedom of expression set forth in Article 10 of the Convention.'[244]

Like all fundamental rights, however, freedom of expression and information may be restricted by the state, so long as the restrictions are provided by law, proportionate, and necessary either to meet

---

[241] Bayer, J., et. al., *The Fight against Disinformation and the Right to Freedom of Expression*, European Parliament, Policy Department for Citizens' Rights and Constitutional Affairs, Brussels, 2021, available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/695445/IPOL_STU(2021)695445_EN.pdf.

[242] Ó Fathaigh, R. et. al., 'The Perils of Legally Defining Disinformation', *Internet Policy Review*, Vol. 10, No. 4, November 4, 2021; Cavaliere, P., 'The Truth in Fake News: How Disinformation Laws Are Reframing the Concepts of Truth and Accuracy on Digital Platforms', *European Convention on Human Rights Law Review*, Vol. 3, No. 4, October 11, 2022, pp. 481–523.

[243] Judgment of 6 March 2001, *Connolly v. Commission*, C-274/99 P, ECLI:EU:C:2001:127.

[244] Judgement of 27 April 2004 of the European Court of Human Rights, *Salov v. Ukraine*, application no. 65518/01.

general interest objectives recognised by the EU, such as public health or security, or to protect the rights and freedoms of others. For example, in *Google Spain*, the ECJ held that the rights of a private individual to privacy and data protection, protected by Articles 7 and 8 of the Charter, override the interests of internet users in accessing that information. Even this, however, 'depends on specific cases, on the nature of the information (…) and on the interest of the public in having that information.'[245] Permissible limitations of freedom of expression are interpreted restrictively by the ECJ, and commentators have noted that the ECJ would most likely be reluctant to uphold responses to mis- and disinformation, beyond illegal speech, that could have chilling effects, such as censorship or online surveillance.[246] Broad prohibitions or penalties based on vague criteria cannot be regarded as permissible under ECtHR case law.[247]

Despite all this, as Sections 4.4 and 4.5 will describe, prohibiting and removing content identified as disinformation is now playing an increasing role in platforms' content moderation mechanisms and policymakers' responses. As well as ordering the removal of illegal content, states may attempt to indirectly influence the suppression of online speech by exercising pressure on private actors. This can involve imposing legal liability for their users' speech, but can also involve informal cooperation between state institutions and platforms, or coercing platforms into self-regulation by threatening harder regulation.[248] Platforms' policies on terrorist content have historically been strongly influenced by state law enforcement institutions,[249] and since the onset of the Covid-19 pandemic, policymakers have also openly encouraged platforms to take more action against disinformation.[250] This is problematic from a fundamental rights perspective, as European human rights law also mandates states to refrain from measures which indirectly target freedom of expression by incentivising platforms to censor users.[251] In general, alternative interventions which aim to reduce the spread of disinformation without deleting it entirely will be more likely to be proportionate under human rights law.

4.3. Empirical research on disinformation

As disinformation on social media has become an increasingly prominent policy concern in recent years, the volume of research in this area has exploded.[252] Nonetheless, key issues and basic factual

---

[245] Judgment of the Court of Justice of 13 May 2014 Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González, C-131/12, ECLI:EU:C:2014:317, paragraph 81.

[246] Bayer, J., et. al., The Fight against Disinformation and the Right to Freedom of Expression, European Parliament, Brussels, 2021, page 24,
available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/695445/IPOL_STU(2021)695445_EN.pdf.

[247] Ó Fathaigh, R., et. al., 'The Perils of Legally Defining Disinformation', Internet Policy Review, Vol. 10, No. 4, November 4, 2021.

[248] Land, M.K., 'Against Privatized Censorship: Proposals for Responsible Delegation', Virginia Journal of International Law, 2019.

[249] Bellanova, R., and De Goede, M., 'Co-Producing Security: Platform Content Moderation and European Security Integration', JCMS: Journal of Common Market Studies, Vol. 60, No. 5, September 2022, pp. 1316–1334.; Bloch-Wehba, H., 'Content Moderation as Surveillance', Berkeley Technology Law Journal, Vol. 36, Iss. 3, 2022, pp. 1297-1340.

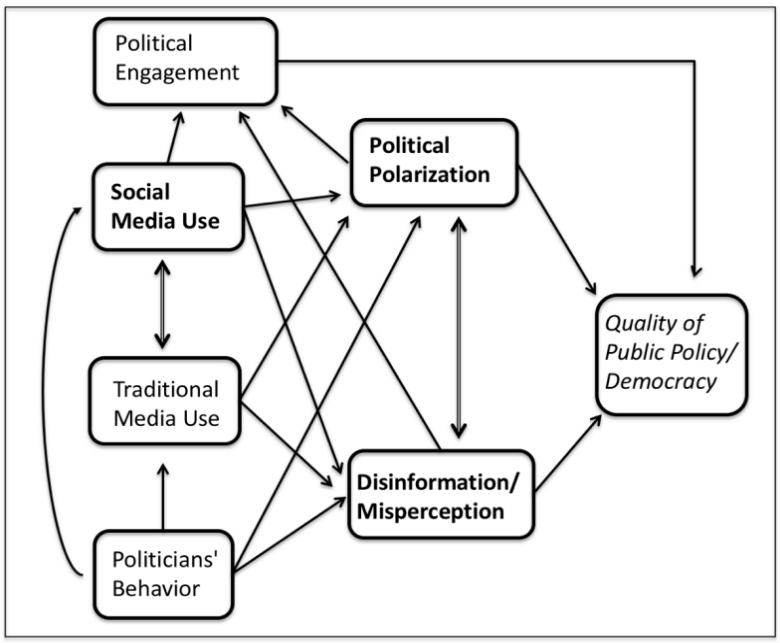
[250] Jourová, V., 'Speech "From pandemic to infodemic"', European Commission, Brussels, 2020, available at:
https://ec.europa.eu/commission/presscorner/detail/en/speech_20_1000.

[251] Land, M.K., 'Against Privatized Censorship: Proposals for Responsible Delegation', Virginia Journal of International Law, 2019.; Bayer, J., et. al., The Fight against Disinformation and the Right to Freedom of Expression, European Parliament, Brussels, 2021,
available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/695445/IPOL_STU(2021)695445_EN.pdf.

[252] Freelon, D., and Wells, C., 'Disinformation as Political Communication', *Political Communication*, Vol. 37, No. 2, March 3, 2020, pp. 145–156; Bernstein, J., 'Bad News: Selling the Story of Disinformation', *Harper's Magazine*, Vol. September 2021,

questions – in particular, regarding the causal effects of disinformation – remain highly uncertain and debated.[253] A useful literature review by Joshua Tucker and colleagues suggests some of this complexity with the below diagram, presented in Figure 2, which identifies just some of the possible causal relationships between social media platforms, other media and institutions, and political outcomes. The authors note that as well as being complex, most of these relationships are as yet hypothetical/unconfirmed and could be either positive or negative. This illustrates how complex the online media and political environment is and cautions against making broad generalisations about how social media and online disinformation affect society.

Figure 2: Social Media, Political Polarisation, Misperception and Democratic Quality



Source: Tucker, J., et. al. [254]

There are several reasons for this uncertainty. One is the inherent complexity of the sociotechnical systems involved, in which millions or billions of users interact with technologically complex media environments involving many platforms, interfaces and publishers. In addition, it is very difficult to disentangle the effects of online disinformation on democracy and political debate from other trends in the political and media environment. A recent literature review suggests that many existing studies are methodologically flawed and likely to overstate how many people believe in disinformation and

August 9, 2021, available at: https://harpers.org/archive/2021/09/bad-news-selling-the-story-of-disinformation/ ; Lenoir, T., and Anderson, C., 'Introduction Essay: What Comes After Disinformation Studies', *Bulletin of Technology and Public Life,* 2023.

[253] Tucker, J., et. al., 'Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature', *SSRN Electronic Journal*, 2018; Keller, C.I., 'Don't Shoot the Message: Regulating Disinformation Beyond Content', *Direito Público*, Vol. 18, No. 99, 2021, pp. 486–515; Haidt, J., & Bail, C. 'Social media and political dysfunction: A collaborative review', New York, 2022, available at: https://docs.google.com/document/d/1vVAtMCQnz8WVxtSNQev_e1cGmY9rnY96ecYuAj6C548/edit.

[254] Tucker, J., et. al., 'Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature', *SSRN Electronic Journal*, 2018, p. 5.

how much it impacts their behaviour, for example because they rely on survey designs which are known to encourage people to endorse beliefs they do not necessarily hold.[255]

Finally, it is hard to generalise about the prevalence and impacts of disinformation, as there is little reason to expect the highly complex and recursive dynamics between users, platforms, media publishers and the broader political and media environment to operate in the same way in different sociopolitical contexts.[256] Existing research on disinformation disproportionately focuses on the US, and there is generally a lack of understanding about how and when these findings might generalise to other contexts.[257] More research on social media, disinformation and democracy in different European contexts - and in other contexts around the world - would be useful.

4.3.1. Causes and effects of online disinformation

The causal effects of exposure to disinformation content on individuals are unclear and disputed.[258] They are likely very different for different individuals (for example, age, political leanings, and education have been identified as factors that affect the likelihood of believing disinformation[259]) and in different social contexts. However, since the start of the Covid-19 pandemic, numerous studies have suggested that, for a minority of people, exposure to online misinformation negatively affects vaccination rates and other safety measures such as mask-wearing.[260] One German study has linked political misinformation to voting behaviour.[261] Misinformation targeting racial minorities during the Covid-19 pandemic has also been linked to violence and institutional discrimination in various European countries.[262] Given the potential seriousness of such consequences, it is important to recognise that even if disinformation does not have such causal effects on the majority of people, effects on a small minority can still have important impacts on the rule of law, democracy, and fundamental rights.

Another prominent debate is whether social media promote belief in false information by creating 'filter bubbles'[263] in which social networks and personalised recommendations only expose individuals

---

[255] Altay, S., et. al, 'Misinformation on Misinformation: Conceptual and Methodological Challenges', *Social Media + Society*, 2023.

[256] Lorenz-Spreen, P., et al., 'A Systematic Review of Worldwide Causal and Correlational Evidence on Digital Media and Democracy', *Nature Human Behaviour*, November 7, 2022.

[257] Tucker, J., et. al., 'Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature', *SSRN Electronic Journal*, 2018.

[258] Altay, S., et al., 'Misinformation on Misinformation: Conceptual and Methodological Challenges', *Social Media + Society*, 2023.

[259] Roozenbeek, J., et al., 'Susceptibility to Misinformation about COVID-19 around the World', *Royal Society Open Science*, Vol. 7, No. 10, October 2020, p. 201199.

[260] Roozenbeek, J., et al., 'Susceptibility to Misinformation about COVID-19 around the World', *Royal Society Open Science*, Vol. 7, No. 10, October 2020, p. 201199; Borges do Nascimento, I.J., et. al., 'Infodemics and Health Misinformation: A Systematic Review of Reviews', *Bulletin of the World Health Organization*, Vol. 100, No. 9, September 1, 2022, pp. 544–561; Joseph, A.M., et al., 'COVID-19 Misinformation on Social Media: A Scoping Review', *Cureus*, April 29, 2022. In their literature review, Borges do Nascimento et al (footnote n. 266) suggest that the quality of many of these studies is not very high. At present, such findings are suggestive of the effects of Covid-related misinformation, but it is to be hoped that further research will clarify the situation over time.

[261] Zimmermann, F., and Kohring M., 'Mistrust, Disinforming News, and Vote Choice: A Panel Survey on the Origins and Consequences of Believing Disinformation in the 2017 German Parliamentary Election', *Political Communication*, Vol. 37, No. 2, March 3, 2020, pp. 215–237.

[262] Szakàcs J., Bognàr E., European Parliament. Directorate General for External Policies of the Union.,
The Impact of Disinformation Campaigns about Migrants and Minority Groups in the EU: In Depth Analysis.,
Publications Office, LU, 2021 in *the EU: In Depth Analysis.*, Publications Office, Brussels, 2021. https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/INGE/DV/2021/07-12/IDADisinformation_migrant_minorities_EN.pdf.

[263] Pariser, E., The Filter Bubble: What the Internet Is Hiding from You, Penguin books, London, 2012.

to information that reinforces their existing beliefs and perspectives, thus encouraging belief in unreliable information as people do not encounter opposing or critical views. Although this phenomenon has been much discussed in academia and the media, researchers have generally found little evidence that it is a widespread or consistent effect of social media.[264] On the contrary, social media use has repeatedly been linked to exposure to more diverse news sources and political perspectives.[265]

However, many of these key studies have been criticised for making overly broad claims that filter bubbles do not exist, based on investigation of rather narrow causal mechanisms, and for overlooking contextual variations and more nuanced causal explanations (for example, the unpredictable effects that platform recommendations may produce when they interact with particular individual behaviours and social networks[266]). Importantly, the claim that filter bubbles are not a widespread and consistent phenomenon does not mean that they do not exist for some individuals in some contexts.[267] As noted above, effects on a small minority of individuals may still have significant policy implications.

Moreover, personalised curation of social media content may be relevant in ways other than its direct impacts on individuals. For example, media theorist Mark Andrejevic has suggested that it may be linked to decreased trust in politics and increased polarisation and social division, as people have less of a sense of sharing a common political and media environment with people who are different from them, weakening the 'imagined communities' created by traditional mass media.[268] Along these lines, one empirical study suggests that a key factor driving the spread of Covid-19 misinformation on Facebook is that the platform's architecture is designed to encourage users to form communities with other like-minded users. These communities offer channels to seek out and disseminate alternative narratives around vaccinations and the pandemic.[269]

Relatedly, platform recommender algorithms – which are typically optimised for some version of 'engagement', i.e. the probability that a user will interact with content and keep using the platform[270] – have frequently been accused of promoting dis- and misinformation, as well as other sensationalist, extreme and/or polarising content, because it is most likely to get a reaction from users.[271] Given the

---

[264] Zuiderveen Borgesius, F.J., et. al., 'Should We Worry about Filter Bubbles?', *Internet Policy Review*, Vol. 5, No. 1, March 31, 2016; Tucker, J., et. al., 'Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature', *SSRN Electronic Journal*, 2018; Arguedas, A. R., 'Echo Chambers, Filter Bubbles, and Polarisation: A Literature Review', *Reuters Institute for the Study of Journalism*, Oxford, 2022. https://reutersinstitute.politics.ox.ac.uk/echo-chambers-filter-bubbles-and-polarisation-literature-review.

[265] Tucker, J., et. al, 'Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature', *SSRN Electronic Journal*, 2018; Scharkow, M., et al., 'How Social Network Sites and Other Online Intermediaries Increase Exposure to News', *Proceedings of the National Academy of Sciences*, Vol. 117, No. 6, February 11, 2020, pp. 2761–2763.

[266] Narayanan, A., 'Is There a Filter Bubble on Social Media? A Call for Epistemic Humility', Princeton University Media Central, 2021, https://mediacentral.princeton.edu/media/1_45q6h2g0.

[267] Tucker, J., et. al., 'Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature', *SSRN Electronic Journal*, 2018.

[268] Andrejevic, M., *Automated Media*, Routledge, London ; New York, NY, 2020.

[269] Broniatowski, D.A., et. al., 'Facebook's Architecture Undermines Vaccine Misinformation Removal Efforts', ArXiv, 2022. https://arxiv.org/abs/2202.02172.

[270] Leerssen, P., 'The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems', *European Journal of Law and Technology*, Vol. 11, No 2 2020.

[271] Vaidhyanathan, S., *Antisocial Media: How Facebook Disconnects US and Undermines Democracy*, Oxford University Press, New York, NY, United States of America, 2018.; Bennett, L., et. al., 'Treating Root Causes, not Symptoms: Regulating Problems of Surveillance and Personal Targeting in the Information Technology Industries', *G20 Insights*, 2021 https://www.g20-insights.org/policy_briefs/treating-root-causes-not-symptoms-regulating-problems-of-surveillance-and-personal-targeting-in-the-information-technology-industries/; Barrett, P., and Hendrix, J., 'Research Highlights A Platform 'Weaponized': How YouTube Spreads Harmful Content—And What Can Be Done About It, NYU Stern', 2022

diversity among platforms and user experiences, it is difficult to claim or disprove that platforms generally and consistently promote such content.[272] However, there is abundant evidence of harmful content, including misinformation and extremist content, being promoted by specific platforms in specific contexts.[273] Some evidence also suggests that platforms such as Twitter and TikTok, which make it easy for users to immediately reshare information they engage with to their followers, are particularly prone to amplifying dis- and misinformation.[274]

That said, evidence suggests that classically 'viral' spread of disinformation, reliant on many individual users resharing information with their close contacts, is relatively rare. Rather, when misleading content becomes widely popular, this is typically because it is amplified by highly visible elite actors such as politicians, celebrities and/or legacy media.[275] This points to the importance of the broader media environment and political culture: holding powerful actors accountable for spreading disinformation is more important than censoring it everywhere it appears online. Research and policy on disinformation would benefit from paying more attention to how individuals engage with reliable and unreliable news and information in their overall information diets, including traditional media and other information sources, rather than assuming that social media platforms are solely responsible for promoting disinformation.[276]

4.3.2. The bigger picture

Research on dis- and misinformation repeatedly highlights that it cannot sensibly be considered as an isolated or technical problem caused by social media, but is rather an aspect of broader social and political trends.[277] These include declining trust in mainstream media, politicians and political institutions; the economic decline of traditional news media; rising authoritarianism; and increasing political polarisation, especially 'affective polarisation' (which refers to emotional antipathy for opposing political sides and identification with one's own side, as opposed to substantive

https://www.stern.nyu.edu/experience-stern/faculty-research/platform-weaponized-how-youtube-spreads-harmful-content-and-what-can-be-done-about-it ;Tufekci, Z., 'Opinion | We Pay an Ugly Cost for Ads on Twitter', *The New York Times*, November 4, 2022, sec. Opinion https://www.nytimes.com/2022/11/04/opinion/elon-musk-twitter-free.html ; Davy, J., 'Amicus Brief for Gonzalez v Google', Integrity Institute, 2022 https://integrityinstitute.org/amicus-brief-for-gonzalez-v-google.

272 Silverman, B., Twitter thread, 2022 https://twitter.com/brandonsilverm/status/1534527964796186625.

273 Kaiser, J., and Rauchfleisch, A., 'Unite the Right? How YouTube's Recommendation Algorithm Connects The U.S. Far-Right', *D&S Media Manipulation: Dispatches from the Field*, April 11, 2018 https://medium.com/@MediaManipulation/unite-the-right-how-youtubes-recommendation-algorithm-connects-the-u-s-far-right-9f1387ccfabd ; Seetharaman, J.H., and D., 'Facebook Executives Shut Down Efforts to Make the Site Less Divisive', *WSJ*, 2020 https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499; 'Malgorithm', Center for Countering Digital Hate, 2021 https://www.counterhate.com/malgorithm ; Zadrozny, B., ''Carol's Journey': What Facebook knew about how it radicalized users', NBC News, 2021 https://www.nbcnews.com/tech/tech-news/facebook-knew-radicalized-users-rcna3581 ; Merrill, J. B., and Oremus, W., 'Five points for anger, one for a 'like': How Facebook's formula fostered rage and misinformation', Washington Post, 2021 https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/.

274 'Misinformation Amplification Analysis and Tracking Dashboard', Integrity Institute, 2022 https://integrityinstitute.org/our-ideas/hear-from-our-fellows/misinformation-amplification-tracking-dashboard.

275 Tucker, J.et. al. 'Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature', *SSRN Electronic Journal*, 2018; Bruns, A., et. al, ''Corona? 5G? Or Both?': The Dynamics of COVID-19/5G Conspiracy Theories on Facebook', *Media International Australia*, Vol. 177, No. 1, November 2020, pp. 12–29.

276 Altay, S., et. al., 'Misinformation on Misinformation: Conceptual and Methodological Challenges', *Social Media + Society*, 2023.

277 Altay, S., et. al., 'Misinformation on Misinformation: Conceptual and Methodological Challenges', *Social Media + Society*, 2023.

disagreement[278]). These are structural economic and political problems and cannot simply be attributed to technological change. Addressing them requires broader political and institutional changes which are outside the scope of this report.

However, while their causal influence should not be overstated, social media may play a role in intensifying certain existing trends. In particular, the rise to dominance of online platforms in the media and advertising industries is widely accepted as one of multiple causal factors behind the decline of traditional news media, in particular local journalism.[279] At the same time, social media have enabled people to establish alternative media networks, distribution channels, and communities for information which is more politically partisan and/or less reliable than traditional media outlets.[280]

Social media affordances which expose people to diverse political views while also encouraging them to share their identity and interests have also been linked to increased affective polarisation.[281] This is a particularly significant point in the context of disinformation, as research suggests that many people consume and share false information not because they rationally believe it to be true, but rather because it is congruent with the identity they wish to signal to others.[282] Affective polarisation increases people's incentives to share information that discredits opposing political sides, even if they do not see it as truthful or reliable.

4.3.3. Recent developments

Several recent trends in the online information environment are relevant to mention here. First, recent reports highlight that organised and professional disinformation operations are on the rise.[283] Organised campaigns to spread disinformation may be undertaken by or on behalf of states, for political purposes – with Russia's operations during US and European elections being the most

---

278 Freelon, D., and Wells, C., 'Disinformation as Political Communication', *Political Communication*, Vol. 37, No. 2, March 3, 2020, pp. 145–156.

279 Pickard, V., *Democracy without Journalism?: Confronting the Misinformation Society*, 1st ed., Oxford University Press, 2020; for more details see chapter 5 of this study on media pluralism .

280 Lewis, R., 'Alternative influence', 2018 https://datasociety.net/library/alternative-influence/ ; Barrett P., and Hendrix J., 'Research Highlights | A Platform 'Weaponized': How YouTube Spreads Harmful Content—And What Can Be Done About It, NYU Stern', 2022 https://www.stern.nyu.edu/experience-stern/faculty-research/platform-weaponized-how-youtube-spreads-harmful-content-and-what-can-be-done-about-it; Broniatowski, D.A., et. al., 'Facebook's Architecture Undermines Vaccine Misinformation Removal Efforts', 2022 https://arxiv.org/abs/2202.02172 .

281 Freelon, D., and Wells, C., 'Disinformation as Political Communication', *Political Communication*, Vol. 37, No. 2, March 3, 2020, pp. 145–156; Törnberg, P., 'How Digital Media Drive Affective Polarization through Partisan Sorting', *Proceedings of the National Academy of Sciences*, Vol. 119, No. 42, October 18, 2022, p. e2207159119 .

282 Polletta, F., and Callahan, J., 'Deep Stories, Nostalgia Narratives, and Fake News: Storytelling in the Trump Era', *American Journal of Cultural Sociology*, Vol. 5, No. 3, October 2017, pp. 392–408; Boyd, D., 'You Think You Want Media Literacy… Do You?', *Data and Society: Points*, 2018 https://points.datasociety.net/you-think-you-want-media-literacy-do-you-7cad6af18ec2 ; Hagood, M., 'Emotional Rescue', *Real Life Magazine*, 2021 https://reallifemag.com/emotional-rescue/; Altay, S., et. al., 'Misinformation on Misinformation: Conceptual and Methodological Challenges', *Social Media + Society*, 2023.

283 Bradshaw, S., et al., 'Industrialized Disinformation: 2020 Global Inventory of Organised Social Media Manipulation. Working Paper 2021.1',
Project on Computational Propaganda, Oxford, 2021 https://demtech.oii.ox.ac.uk/research/posts/industrialized-disinformation/ ; Bayer, J., et. al., European Parliament. Directorate General for External Policies of the Union., *Disinformation and Propaganda: Impact on the Functioning of the Rule of Law in the EU and Its Member States : 2021 Update.*, Publications Office, LU, 2021 https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653633/EXPO_STU(2021)653633_EN.pdf;
Woolley, S., Manufacturing Consensus: Understanding Propaganda in the Era of Automation and Anonymity, Yale University Press, 2023.

notorious example – or for profit. There are increasing numbers of businesses offering 'computational propaganda as a service' on behalf of other actors.[284]

These trends are obviously concerning for democracy and the rule of law, given the potential of large-scale, strategic political disinformation campaigns to weaken trust in politics and increase divisions and polarisation. Importantly, just creating the perception that foreign operators are interfering in the political process and that media information cannot be trusted may be enough to produce these effects, even if disinformation content is not actually widely shared or believed.[285] Previous investigations for the European Parliament have shown that foreign disinformation campaigns have often targeted minority groups in Europe, such as migrants and LGBTQ+ people, successfully stoking prejudice and mobilising political activism against these groups.[286] In a recent example, Meta's trust and safety team have been attempting to tackle a sustained disinformation campaign originating from Russian state operatives, which aims to spread false information and resentment against Ukrainian refugees in Europe.[287] In light of these particular impacts on political cohesion, trust and equality, identifying and targeting large-scale, coordinated disinformation operations should be a priority.

Organised disinformation operations may also be aided by technological developments which will increasingly facilitate large-scale automated generation and dissemination of disinformation content.[288] AI programmes which can generate convincing text and facial images are now widely available. As well as enabling cheaper and more efficient production of disinformation content, these tools will create new challenges in detecting disinformation, as researchers and trust and safety teams will no longer be able to search for repeated patterns in text or perform reverse image searches on photos.[289]

Capacities to create AI-generated 'deepfake' videos of well-known figures are also rapidly advancing, and becoming harder to automatically detect.[290] While their potential use for political disinformation has been much discussed, the vast majority of deepfake videos currently published online involve non-consensual pornography, gendered abuse, and privacy violations.[291] As Chapter 3 discussed, these

---

[284] Bradshaw, S., et al., 'Industrialized Disinformation: 2020 Global Inventory of Organised Social Media Manipulation. Working Paper 2021.1', *Project on Computational Propaganda*, Oxford, 2021 https://demtech.oii.ox.ac.uk/research/posts/industrialized-disinformation/.

[285] Vaccari, C., and Chadwick, A., 'Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News', *Social Media + Society*, Vol. 6, No. 1, January 2020, p. 205630512090340.

[286] Strand, C., et. al, Disinformation Campaigns about LGBTI+ People in the EU and Foreign Influence: Briefing, European Parliament, Brussels, 2021 https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/653644/EXPO_BRI(2021)653644_EN.pdf ; Szakàcs, J., Bognàr, E., European Parliament. Directorate General for External Policies of the Union., *The Impact of Disinformation Campaigns about Migrants and Minority Groups in the EU: In Depth Analysis*., Publications Office, Brussels, 2021 https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/INGE/DV/2021/07-12/IDADisinformation_migrant_minorities_EN.pdf.

[287] Morris, L., and Oremus, W., 'Russian Disinformation Is Demonizing Ukrainian Refugees', *Washington Post*, December 8, 2022 https://www.washingtonpost.com/technology/2022/12/08/russian-disinfo-ukrainian-refugees-germany/.

[288] Goldstein, J.A., et. al., 'Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations', *Stanford Internet Observatory*, 2023 https://cdn.openai.com/papers/forecasting-misuse.pdf .

[289] DiResta, R., 'The Supply of Disinformation Will Soon Be Infinite', *The Atlantic*, September 20, 2020 https://www.theatlantic.com/ideas/archive/2020/09/future-propaganda-will-be-computer-generated/616400/.

[290] Farid, H., 'Creating, Using, Misusing, and Detecting Deep Fakes', *Journal of Online Trust and Safety*, Vol. 1, No. 4, September 20, 2022.

[291] Toparlak, R. T., 'Criminalising Pornographic Deep Fakes: A Gender-Specific Inspection of Image-Based Sexual Abuse', *Sciences Po Law School 10th Graduate Conference on Law & Technology,* Sciences Po, Paris, June 16, 2022 https://www.sciencespo.fr/public/chaire-numerique/wp-content/uploads/2022/06/3a-Toparlak_Criminalising-Pornographic-Deep-Fakes.pdf.

forms of gendered harassment are themselves a pressing issue for fundamental rights and equal participation in online public spaces. However, recent incidents have also illustrated the potential for deepfake videos to impact political debate. Early in the Ukraine war, a deepfake video was disseminated of President Volodymyr Zelensky appearing to announce Ukraine's surrender to the Russian invasion. In this case, the video was not widely believed, largely because Zelensky had already publicly warned that such content would be produced.[292] In future, even if deepfakes are easily identified, they could effectively be released at strategic moments when they can have a significant impact in a short time before being debunked; they could provide an excuse or cover for escalatory political responses even if they are not widely believed. In addition, the mere possibility and presence of deepfake videos on social media may impact accountability and trust in politics, as it will be possible to dismiss any video or photo as fake. Experimental evidence suggests that exposure to deepfake videos reduces overall trust in news on social media.[293]

Other relevant developments in the broader social media landscape include the growing popularity of TikTok, a platform which centres short-form (up to 60-second) videos distributed through personalised algorithmic recommendations.[294] Its popularity has in turn significantly influenced product design and business strategies at other major platforms, marking a general shift towards increased emphasis on video content and algorithmic recommendations of content from accounts the user does not follow. This poses new challenges for disinformation policy for two reasons. First, there is evidence that audiovisual messages tend to be more convincing and have a greater impact on users.[295] Second, they are also much more difficult to moderate or fact-check. Automated analysis of video content is technically even more difficult than analysing text or images, while manual analysis is very time-consuming.[296] In addition, specific features of TikTok (and of copycat products, e.g., Instagram Reels), such as the presentation of full-screen videos with little contextual information, appear to exacerbate credibility issues.[297]

The relevance of design and business factors in the spread of mis - and disinformation is further highlighted by recent events following the acquisition of Twitter by Elon Musk. At the time of writing, Musk has laid off most of Twitter's contracted content moderators, as well as large numbers of senior

---

[292] Simonite, T., 'A Zelensky Deepfake Was Quickly Defeated. The Next One Might Not Be', *Wired*, 2022 https://www.wired.com/story/zelensky-deepfake-facebook-twitter-playbook/.

[293] Vaccari, C., and Chadwick A., 'Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News', *Social Media + Society*, Vol. 6, No. 1, January 2020, p. 205630512090340 .

[294] Riemer, K., and Peter, S., 'Algorithmic Audiencing: Why We Need to Rethink Free Speech on Social Media', *Journal of Information Technology*, Vol. 36, No. 4, December 2021, pp. 409–426.

[295] Tucker, J., et. al., 'Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature', *SSRN Electronic Journal*, 2018; Hameleers, M., et. al., 'A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated via Social Media', *Political Communication*, Vol. 37, No. 2, March 3, 2020, pp. 281–301; Vaccari, C., and Chadwick, A., 'Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News', *Social Media + Society*, Vol. 6, No. 1, January 2020, p. 205630512090340; Weikmann, T. and Lecheler, S., 'Visual disinformation in a digital age: A literature synthesis and research agenda', *New Media + Society*, 2022.

[296] Chowdhury, N., 'Automated Content Moderation: A Primer', *Freeman Spogli Institute of International Studies at Stanford University*, 2022 https://fsi.stanford.edu/news/automated-content-moderation-primer.

[297] Nilsen, J.,et. al., 'Tiktok, the war on Ukraine, and 10 features that make the app vulnerable to misinformation', *The media manipulation casebook,* 2022 https://mediamanipulation.org/research/tiktok-war-ukraine-and-10-features-make-app-vulnerable-misinformation; Hsu, T., 'Worries Grow That TikTok Is New Home for Manipulated Video and Photos', *The New York Times*, November 4, 2022, sec. Technology https://www.nytimes.com/2022/11/04/technology/tiktok-deepfakes-disinformation.html?action=click&module=RelatedLinks&pgtype=Article; Stokel-Walker, C., 'Why Do People Believe Everything They Watch on TikTok?', *VICE*, February 6, 2023, https://www.vice.com/en/article/z34a95/why-do-people-believe-everything-on-tiktok.

executives and other employees working on AI ethics, security, accessibility, and content policy issues.[298] He has also overseen several major design changes which were heavily criticised by experts in social media trust and safety,[299] notably a short-lived system in which any user could pay 8 USD for their account to be marked 'verified'. This status previously signified that a user's identity had been confirmed by Twitter staff; it also gives tweets from verified accounts higher prominence in search and recommendations. Paid verification therefore offered an easy way to amplify disinformation.[300] The new verification scheme was paused after a wave of parody accounts imitating well-known brands, [301] but, at the time of writing, levels of disinformation and hate speech on Twitter still appear to be elevated, likely because staff who would previously have been addressing these issues have been fired.[302]

Future developments at Twitter remain highly uncertain, but these events should be taken to illustrate that design, operational, and staffing decisions – not only content policies and moderation - are of central importance in addressing disinformation. Maintaining safe and trustworthy information environments requires significant and ongoing investments of resources and personnel from platforms. It also requires security and integrity issues to be considered in all product design decisions. Similarly, the rapid changes underway at Twitter raise questions about the enforcement of EU regulatory frameworks like the DSA and Code of Practice on Disinformation, and their ability to ensure that private companies make these necessary investments and commitments. It is clear that effective oversight and intervention by the Commission and national digital services coordinators will require high levels of resources and expertise, and an active regulatory strategy.[303]

4.3.4. Implications for democracy, fundamental rights and the rule of law

Based on the brief literature review in Section 4.3.1, it is possible to reach some general conclusions about the implications of mis- and disinformation for democracy, fundamental rights, and the rule of law which will form the basis for Chapter 4's legal analysis and policy recommendations.

While many specific dynamics and causal effects of online disinformation remain disputed and uncertain – and may not be the same everywhere in Europe – there is evidence that disinformation can directly cause harm in at least some specific contexts (for example in relation to Covid-19 vaccination rates). More generally, beyond considering specific and direct causal effects on individual social media users, or on particular political events such as elections, widespread online disinformation is a policy concern because it contributes to and exacerbates general 'second order problems' for democracy and

---

[298] Knight, W., 'Elon Musk Has Fired Twitter's 'Ethical AI' Team', *Wired*, 2022 https://www.wired.com/story/twitter-ethical-ai-team/?utm_brand=wired-science&mbid=social_tw_sci&utm_source=twitter&utm_social-type=owned&utm_medium=social;

Newton, C., and Schiffer, Z., 'Twitter, cut in half', *Platformer News*, 2022 https://www.platformer.news/p/twitter-cut-in-half?utm_campaign=post ; Delcker, J., 'Twitter's sacking of content moderators raises concerns', DW, 2022 https://www.dw.com/en/twitters-sacking-of-content-moderators-will-backfire-experts-warn/a-63778330.

[299] Swisher, K., and Roth Y., 'The Crisis at Twitter'(video), Knight Foundation, 2022 https://vimeo.com/776426548.

[300] Schreiber, M., ''Verified' Anti-Vax Accounts Proliferate as Twitter Struggles to Police Content', *The Guardian*, November 21, 2022, sec. Technology https://www.theguardian.com/technology/2022/nov/21/twitter-anti-vax-health-misinformation.

[301] Lerman, R., and Zakrzewski, C., 'Elon Musk's first big Twitter product paused after fake accounts spread', *Washington Post*, 2022 https://www.washingtonpost.com/technology/2022/11/11/twitter-fake-verified-accounts/.

[302] Knight, W., 'Here's Proof Hate Speech Is More Viral on Elon Musk's Twitter', *Wired*, 2022 https://www.wired.com/story/heres-proof-hate-speech-is-more-viral-on-elon-musks-twitter/; Milman, O., '#ClimateScam: Denialism Claims Flooding Twitter Have Scientists Worried', The Guardian, December 2, 2022, sec. Technology https://www.theguardian.com/technology/2022/dec/02/climate-change-denialism-flooding-twitter-scientists.

[303] Fahy, R., et. al.,'The EU's regulatory push against disinformation', Verfassunblog, 2022 https://verfassungsblog.de/voluntary-disinfo/.

the rule of law:[304] for example, by weakening trust in politics and media institutions, and making it harder to hold politicians and public figures accountable for their actions.

On this basis, while identifying and censoring all online information which is deemed by platforms or public authorities to be false and/or harmful is neither technically feasible nor normatively desirable,[305] it is possible to identify certain phenomena which are particularly harmful and justify a more targeted response.

First, as compared to inadvertently disseminated misinformation and disinformation shared by individual accounts or small communities, organised and professional disinformation operations appear particularly harmful. This is because of their scale; because they figure prominently in public debate and thus generally undermine trust in politics; and because they often aim to exploit and strengthen existing social divisions and prejudices. In addition, organised disinformation operations often target marginalised social groups directly,[306] or target majority groups with messages designed to stigmatise minorities, such as LGBTQ+ people or migrants.[307] Disinformation which targets marginalised social groups undermines equality and non-discrimination, as well as equal access to information, political processes, and public debate.[308]

Second, platforms' operational and design decisions play an important role in enabling or checking the spread of disinformation. Regulating these decision-making processes and ensuring that they prioritise trust and safety considerations should thus be a policy priority.[309] Third, relatedly, certain types of false information, such as health-related claims and claims targeting minority groups, have particular potential to directly cause harm, for example by encouraging dangerous behaviour or inciting violence.[310] Accordingly, where content-level interventions like content moderation and fact-checking

---

[304] Spies, S., *On Digital Disinformation and Democratic Myths*, MediaWell, Social Science Research Council, December 10, 2019 https://mediawell.ssrc.org/expert-reflections/on-digital-disinformation-and-democratic-myths/; Keller, I. C., 'Don't Shoot the Message: Regulating Disinformation Beyond Content', *Direito Público*, Vol. 18, No. 99, 2021, pp. 486–515.

[305] Marsden, C., et al., 'Platform Values and Democratic Elections: How Can the Law Regulate Digital Disinformation?', *Computer Law & Security Review*, Vol. 36, April 2020, p. 105373; Keller, I. C., 'Don't Shoot the Message: Regulating Disinformation Beyond Content', *Direito Público*, Vol. 18, No. 99, 2021, pp. 486–515.

[306] Woolley, S., 'In Many Democracies, Disinformation Targets the Most Vulnerable', *Centre for International Governance Innovation*, 2022 https://www.cigionline.org/articles/in-many-democracies-disinformation-targets-the-most-vulnerable/; Takur, D., and DeVan, L. H., 'Facts and their Discontents:A Research Agenda for Online Disinformation, Race, and Gender', *Center for Democracy and Technology,* 2021 https://cdt.org/wp-content/uploads/2021/02/2021-02-10-CDT-Research-Report-on-Disinfo-Race-and-Gender-FINAL.pdf ; Bhatia, A., 'Election Disinformation in Different Languages is a Big Problem in the U.S.', *Center for Democracy and Technology,* 2022 https://cdt.org/insights/election-disinformation-in-different-languages-is-a-big-problem-in-the-u-s/ .

[307] Strand C., et al., Disinformation Campaigns about LGBTI+ People in the EU and Foreign Influence: Briefing, European Parliament, Brussels, 2021, https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/653644/EXPO_BRI(2021)653644_EN.pdf.

[308] Keller, C.I., 'Don't Shoot the Message: Regulating Disinformation Beyond Content', *Direito Público*, Vol. 18, No. 99, 2021, pp. 486–515; Szakàcs, J., and Bognàr, E., European Parliament. Directorate General for External Policies of the Union., *The Impact of Disinformation Campaigns about Migrants and Minority Groups in the EU: In Depth Analysis*., Publications Office, Brussels, 2021 https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/INGE/DV/2021/07-12/IDADisinformation_migrant_minorities_EN.pdf.

[309] Bennett, O., 'The Promise of Financial Services Regulatory Theory to Address Disinformation in Content Recommender Systems', *Internet Policy Review*, Vol. 10, No. 2, May 11, 2021.

[310] Szakàcs, J., and Bognàr, E., European Parliament. Directorate General for External Policies of the Union., *The Impact of Disinformation Campaigns about Migrants and Minority Groups in the EU: In Depth Analysis*., Publications Office, Brussels, 2021 https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/INGE/DV/2021/07-12/IDADisinformation_migrant_minorities_EN.pdf.

are used, they should focus on clearly defined categories like these, where the direct risk to others can justify the restriction of freedom of expression and information.

4.4. Platform responses

4.4.1. Content moderation

Given that disinformation is generally not illegal speech, disinformation on social media is primarily governed via platforms' contractual community standards. As private actors, they moderate content they consider harmful or not useful to their users – usually through filtering and blocking information,[311] though they also make increasing use of alternative interventions like 'shadowbanning' (hiding content from other users in certain feeds or interfaces) or demoting content (making it less prominent in algorithmic ranking systems).[312]

Disinformation as such is not banned under major platforms' content policies, and is therefore not by default subject to content moderation. However, some types of false information (e.g. Holocaust denial) are banned and moderated because they are illegal in particular countries and/or because they overlap with other banned categories, such as hate speech.[313] In addition, major platforms typically ban specific activities linked to organised disinformation operations, such as the coordinated use of fake accounts and bots to spread information.[314] Finally, due to widespread concerns about dangerous health-related misinformation during the pandemic, most major platforms updated their policies to ban specific categories of Covid-19-related misinformation (for example, claims already debunked by public health authorities).[315]

However, independent investigations have suggested that the effectiveness of such policies in reducing the visibility and engagement of misinformation content is limited.[316] Policy enforcement is very incomplete in practice, with some studies finding that the majority of prohibited misinformation content identified by researchers was not moderated by platforms.[317] Even where content is detected and moderated, this may come too late, after it has already been widely disseminated.[318]

In recent years, whistleblowers from within Meta and Twitter have leaked documents showing that both companies' trust and safety divisions are extremely thinly stretched, with a lack of staff and

---

311 Klonick, K., 'The New Governors: The People, Rules, and Processes Governing Online Speech', *Harvard Law Review*, vol. 131, no. 6, 2018.

312 Gillespie, T., 'Do Not Recommend? Reduction as a Form of Content Moderation', *Social Media + Society*, Vol. 8, No. 3, July 2022, p. 205630512211175.

313 Guynn, J., 'Facebook still has Holocaust denial content three months after Mark Zuckerberg pledged to remove it', *USA Today*, 2021 https://eu.usatoday.com/story/tech/2021/01/27/facebook-holocaust-denial-zuckerberg-twitter-youtube-twitch-reddit/4269035001/ .

314 Keller, T., et al., ''Coordinated Inauthentic Behaviour' And Other Online Influence Operations In Social Media Spaces', *AoIR Selected Papers of Internet Research*, October 5, 2020.

315 Krishnan, N., et al., 'Research Note: Examining How Various Social Media Platforms Have Responded to COVID-19 Misinformation', *Harvard Kennedy School Misinformation Review*, December 15, 2021.

316 Broniatowski, D.A., et al., 'Facebook's Architecture Undermines Vaccine Misinformation Removal Efforts', 2022 https://arxiv.org/abs/2202.02172 ; Darius, P., and Urquhart, M., 'Disinformed Social Movements: A Large-Scale Mapping of Conspiracy Narratives as Online Harms during the COVID-19 Pandemic', *Online Social Networks and Media*, Vol. 26, November 2021, p. 100174.

317 Krishnan, N., et al., 'Research Note: Examining How Various Social Media Platforms Have Responded to COVID-19 Misinformation', *Harvard Kennedy School Misinformation Review*, December 15, 2021.

318 Krishnan, N., et al., 'Research Note: Examining How Various Social Media Platforms Have Responded to COVID-19 Misinformation', *Harvard Kennedy School Misinformation Review*, December 15, 2021.

resources preventing effective enforcement of content policies.[319] All of these leaks indicated that enforcement in less wealthy markets and in languages other than English is a low priority, with basic resources (such as staff who speak the relevant languages) lacking. This raises obvious concerns in the European context given Europe's linguistic, cultural, and political diversity (though safety issues are even more acute in poorer countries around the world[320]).

This raises the question of whether content moderation could be a more effective response, if platforms invested more in moderation staff and tools, and gave more data access and support to independent researchers to identify threats and possible solutions. Some types of content, such as organised electoral disinformation campaigns, pose obvious and direct threats to public safety and democracy. There is a lot of room to dedicate more staff, training, and resources to identifying and countering organised disinformation campaigns around the world, which would undoubtedly be helpful.[321] As Section 4.3.3 suggested, focusing enforcement efforts on coordinated, strategic behaviour has advantages from a practical and fundamental rights perspective. Given the particular harms associated with large-scale strategic disinformation operations, this approach focuses platforms' resources where interventions are most necessary and proportionate. In addition, organised disinformation campaigns can be identified and countered by platforms' trust and safety teams based on behavioural signals (e.g. coordinated posting, use of fake accounts),[322] as opposed to censoring content based only on assessments of its truth or falsity, which raises greater concerns for freedom of expression and democratic debate.[323] Box 6 provides a case study of how Meta implements such anti-disinformation measures.

However, disinformation policies should not assume that faster and more comprehensive content moderation is a sufficient solution. Given the scale of contemporary social media, it is not possible for them to monitor all user content and evaluate it for potentially harmful disinformation.[324] Nor would this be desirable, given the obvious risks to fundamental rights that would come with subjecting all online expression to governments' or platforms' ideas of what is true. And in any case, this is a fundamentally reactive approach which does not address the underlying structural causes of disinformation, ranging from platform design features which encourage the viral spread of sensationalist content to broader issues around trust and reliability in the media. Content moderation efforts and resources have a role to play, but should be narrowly focused on identifying clearly harmful behaviour, such as organised political disinformation campaigns.

---

[319] Dixit, C.S., et al., "I Have Blood On My Hands': A Whistleblower Says Facebook Ignored Global Political Manipulation', *BuzzFeed News* 2020 https://www.buzzfeednews.com/article/craigsilverman/facebook-ignore-political-manipulation-whistleblower-memo; Horwitz, J.S., et al., 'Facebook Employees Flag Drug Cartels and Human Traffickers. The Company's Response Is Weak, Documents Show.', *Wall Street Journal*, September 16, 2021, sec. Tech. https://www.wsj.com/articles/facebook-drug-cartels-human-traffickers-response-is-weak-documents-11631812953?mod=article_inline; Dwoskin, E., Menn, J., and Zakrzewski, C.,'Twitter can't afford to be one of the world's most influential websites', *Washington Post*, 4 September 2022 https://www.washingtonpost.com/technology/2022/09/04/twitter-mudge-alethea-resources/.

[320] Horwitz, J.S., et al., 'Facebook Employees Flag Drug Cartels and Human Traffickers. The Company's Response Is Weak, Documents Show.', *Wall Street Journal*, September 16, 2021, sec. Tech https://www.washingtonpost.com/technology/2022/09/04/twitter-mudge-alethea-resources/.

[321] Hao, K., 'How Facebook got addicted to spreading misinformation', *MIT Technology Review*, 11 March 2021.

[322] Schoch, D., et al., 'Coordination patterns reveal online political astroturfing across the world', *Scientific Reports,* Vol. 12, Article 4572, 2022.

[323] Shadmy, T., 'Content Traffic Regulation: A Democratic Framework for Addressing Misinformation', *Jurimetrics,* Vol. 63, No. 1, 2022.

[324] Keller, D., 'The DSA's Industrial Model for Content Moderation', *Verfassungsblog*, 2022 https://verfassungsblog.de/dsa-industrial-model/.

Box 6: Meta's approach to coordinated inauthentic behaviour

The world's largest social media platform, Facebook, has policies banning what it terms 'coordinated inauthentic behaviour', defined as the coordinated use of fake accounts to mislead people and influence public debate. Since 2018, Facebook's parent company, Meta, has regularly shared relatively detailed reports on its efforts to detect and remove large-scale disinformation operations under this policy. For example, in September 2022 it shared details of how it had detected and removed accounts associated with two large-scale organised disinformation campaigns, one originating from China and primarily targeting the US and Czechia, and one originating from Russia and targeting several European countries with messages relating to the Ukraine war. According to the report, Meta detected some of these campaigns' activities using automated tools which identify behavioural signals associated with the use of fake accounts. It also collaborates with trust and safety staff at other tech companies, as well as with independent researchers, investigative journalists, and governments, in order to identify organised disinformation operations. In turn, it shared details of the messages, websites, and strategies involved with outside researchers to enable further independent research.

Meta's policies and practices remain open to criticism. Meta's reporting on its anti-disinformation operations has been criticised for being insufficiently detailed and specific. Commentators have noted that its 'coordinated inauthentic behaviour' policy remains ambiguous, meaning it can be enforced in an arbitrary and selective way. In addition, whistleblower Sophie Zhang revealed in 2020 that Meta's efforts to counter organised disinformation campaigns are extremely patchy at the global level, with a disproportionate focus on threats targeting the US and other wealthy countries (and originating from their geopolitical rivals, China and Russia) and very little resources invested in tackling disinformation campaigns in poorer and less wealthy countries.

Nonetheless, these forms of cooperation between tech platforms and independent security researchers represent a promising path forward to counter the most serious, coordinated forms of disinformation, establishing practices which can be further built on and improved. Focusing on behavioural signals, such as the coordinated use of fake accounts, allowed Meta to efficiently identify organised disinformation operations, using automated tools, without indiscriminately censoring content. Collaboration between platform companies and independent researchers appears to be a good way to achieve the flexible, adaptable response required to deal with the constantly evolving nature of these kinds of strategic disinformation operations. Transparency and data sharing can not only aid the security community in understanding and responding to disinformation operations, but can also enable critical journalism which counters and contextualises disinformation narratives for the wider public (the *Washington Post* article on Russian disinformation, cited below, offers a good example). In this context, regulators should focus on establishing and developing stronger best practices: for example, requiring platforms to share more granular information with researchers, and to invest more resources in trust and safety efforts in markets which are currently not prioritised. These commitments could be incorporated into the CoP on Disinformation and the evaluation of platforms' risk mitigation obligations under Articles 34-35 DSA (discussed in more detail in Section 4.5.2).

Sources: 'Coordinated Inauthentic Behavior', Meta, n.d. https://about.fb.com/news/tag/coordinated-inauthentic-behavior/; Nimmo, B., and Torrey, M., 'Taking down coordinated inauthentic behavior from Russia and China', Meta, 2022 https://about.fb.com/wp-content/uploads/2022/11/CIB-Report_-China-Russia-Sept-2022.pdf; Lomas, N., 'Meta Reports Takedowns of Influence Ops Targeting US Midterms, Ukraine War', *TechCrunch*, September 27, 2022 Meta reports takedowns of influence; Douek, E., 'What Does "Coordinated Inauthentic Behavior" Actually Mean?', *Slate*, July 2, 2020 https://slate.com/technology/2020/07/coordinated-inauthentic-behavior-facebook-twitter.html; Dixit, C.S., et al., ''I Have Blood On My Hands': A Whistleblower Says Facebook Ignored Global Political Manipulation', *BuzzFeed News*,, 2020

https://www.buzzfeednews.com/article/craigsilverman/facebook-ignore-political-manipulation-whistleblower-memo; Morris, L., and Oremus W., 'Russian disinformation is demonizing Ukrainian refugees', *Washington Post*, December 8, 2022 https://www.washingtonpost.com/technology/2022/12/08/russian-disinfo-ukrainian-refugees-germany/.

4.4.2. Fact-checking partnerships

During the pandemic, as well as deleting content and user accounts, many platforms introduced 'soft moderation' measures, such as accompanying content with warning labels or links to reliable information.[325] In implementing these measures against mis- and disinformation, major platforms have formed partnerships with independent fact-checking organisations, such as leading press agencies. These partnerships have been pioneered by Meta since 2018[326] and have since been adopted by other leading platforms.[327] Fact-checkers can investigate potential mis- and disinformation; supply reliable information to be added through warning labels; and advise on other interventions, such as removing the content entirely or demoting it in recommendations. Platform companies have provided extensive funding for such activities, as well as providing partners with access to additional data, such as how fast content is spreading on the platform.[328]

Notably, Twitter has also in recent years been testing crowdsourced fact-checking, in which participating users can add contextual information to posts. In Twitter's system, to ensure the publication of reliable fact-checks, these notes are only widely displayed once they have gained approval from a diverse range of other participating users. While Twitter claims that the programme has received good feedback from users and effectively discourages belief in misinformation,[329] it is so far only available in the US,[330] and independent research on its effectiveness is lacking. While it is doubtful that crowdsourced information could substitute for professional fact-checking or address all the relevant factors driving the spread of misinformation, it could be one element of a successful response and would be useful for regulators to promote more testing and research on such interventions in Europe.

Some studies suggest that fact-checking can be quite effective at the individual user level, i.e. seeing warning labels and additional information often (though not always) discourages people from believing or sharing misinformation.[331] However, successfully reaching users with these interventions faces many of the same difficulties as content moderation: locating relevant content is difficult and

---

[325] Krishnan, N., et al., 'Research Note: Examining How Various Social Media Platforms Have Responded to COVID-19 Misinformation', *Harvard Kennedy School Misinformation Review*, December 15, 2021.

[326] Ananny, M., 'Making Up Political People: How Social Media Create The Ideals, Definitions, And Probabilities Of Political Speech', *Georgetown Law Technology Review*, Vol. 1, no. 4, 2020.

[327] Mantas, H., 'Twitter Finally Turns to the Experts on Fact-Checking', *Poynter*, August 5, 2021, available at: https://www.poynter.org/fact-checking/2021/twitter-finally-turns-to-the-experts-on-fact-checking/ ; 'Safety Partners', TikTok, n.d. https://www.tiktok.com/safety/en/safety-partners/ ; Ma, O., and Feldman, B., 'How Google and YouTube are investing in fact-checking', *Google News Initiative,* 2022, available at: https://blog.google/outreach-initiatives/google-news-initiative/how-google-and-youtube-are-investing-in-fact-checking/.

[328] Ananny, M., 'Making Up Political People: How Social Media Create The Ideals, Definitions, And Probabilities Of Political Speech', *Georgetown Law Technology Review*, Vol. 4, no. 1, 2020; Ma, O., and Feldman, B., 'How Google and YouTube are investing in fact-checking', *Google News Initiative,* 2022.

[329] 'Birdwatch is getting a new onboarding process and more visible notes', Twitter, 2022, available at: https://blog.twitter.com/en_us/topics/product/2022/birdwatch-getting-new-onboarding-process-more-visible-notes.

[330] Bobrowsky, M., 'Elon Musk Champions Twitter Fact-Checking Feature That Corrects Him', *WSJ*, 2022, available at: https://www.wsj.com/articles/elon-musk-champions-twitter-fact-checking-feature-that-corrects-him-11669436937.

[331] Hameleers, M. et al., 'A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated via Social Media', *Political Communication*, Vol. 37, No. 2, March 3, 2020, pp. 281–301.; Singhal, M., et al., 'SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice', *arXiv*, October 27, available at: 2022 https://arxiv.org/pdf/2206.14855.pdf.

investigating it takes time. Manual fact-checking is resource-intensive and slow, while automated fact-checking tools have only very basic capabilities and cannot be relied on to provide trustworthy information.[332] Research by the Integrity Institute suggests that misinformation primarily spreads and is viewed within 24 hours of being posted, while fact checks are typically added only after this period.[333]

Considering the second-order effects of disinformation on democracy suggests that independent fact-checking has an important role to play. Even if they appear too late to prevent the spread of individual pieces of disinformation, the visible presence on social media of fact-checking labels and information resources from independent media institutions can promote media literacy and general awareness of disinformation. In addition, the prospect of being fact-checked can discourage politicians and other high-profile figures from making false claims, or hold them accountable when they do[334] – which is particularly important, since these prominent network members have the most impact on the spread of disinformation.[335] Importantly, some studies suggest that users are more receptive to fact-checking in countries with strong public and non-partisan media institutions.[336] Fact-checking should thus not be considered in isolation, but as part of a broader ecosystem of accountability and independent media, as discussed in Chapter 5 on media pluralism.

However, scaling up fact-checking enough to address all harmful misinformation does not appear feasible. Even if this were not the case, research suggests that users' emotional motivations to engage with and share disinformation content – like signalling their dislike for a particular politician, or affiliation with a particular group – will often not be addressed by just providing them with more information.[337] Nor can fact-checking address structural factors, such as platform design features which promote the spread of misinformation online, as discussed in Section 4.3.3.

4.5. The EU legal framework on disinformation

This section provides a brief overview and evaluation of key existing and proposed regulatory measures aimed at tackling mis- and disinformation in the EU, complementing the background legal framework presented in Chapter 2. The regulatory regime governing disinformation on social media has four main elements. First, some types of disinformation qualify as 'illegal content' within the notice-and-takedown framework established by the 2000 ECD and refined in the DSA. This means platforms can be notified of such content by users or public authorities and required to remove it. Second, the DSA additionally creates tiered due diligence obligations regarding transparency and safety measures, which require platforms to take action not only against illegal content, but also against disinformation regarded as legal but potentially harmful. Third, the CoP on Disinformation creates additional self-

[332] Alaphilippe, A.,et al., *Automated Tackling of Disinformation: Major Challenges Ahead*, European Parliament, Brussels, 2019, available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624278/EPRS_STU(2019)624278_EN.pdf.

[333] 'Misinformation Amplification Analysis and Tracking Dashboard', Integrity Institute, 2022, available at: https://integrityinstitute.org/our-ideas/hear-from-our-fellows/misinformation-amplification-tracking-dashboard.

[334] Nyhan, B., and Reifler, J., 'The Effect of Fact-Checking on Elites: A Field Experiment on U.S. State Legislators: The Effect Of Fact-Checking On Elites', *American Journal of Political Science*, Vol. 59, No. 3, July 2015, pp. 628–640; Kyriakidou, M., et al., 'Questioning Fact-Checking in the Fight Against Disinformation: An Audience Perspective', *Journalism Practice*, July 7, 2022, pp. 1–17.

[335] Tucker, J., A et al., 'Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature', *SSRN Electronic Journal*, 2018.

[336] Kyriakidou, M., et al., 'Questioning Fact-Checking in the Fight Against Disinformation: An Audience Perspective', *Journalism Practice*, July 7, 2022, pp. 1–17.

[337] Boyd, D., 'You Think You Want Media Literacy… Do You?', *Data and Society: Points*, 2018, available at: https://points.datasociety.net/you-think-you-want-media-literacy-do-you-7cad6af18ec2 ; Hagood, M., 'Emotional Rescue', *Real Life Magazine*, 2021 https://reallifemag.com/emotional-rescue/.

regulatory commitments and further concretises platforms' due diligence obligations. Finally, the EU has created additional transparency and due diligence obligations in the specific area of political advertising.

4.5.1. Notice and takedown obligations

Disinformation policy and research have often started from the assumption that 'disinformation is not per se illegal, even if it can be harmful.'[338] This implies that, under the intermediary liability framework outlined in Chapter 2, platforms are not legally obliged to remove such content. However, they may still do so because such content violates their contractual terms of service, or as a means of implementing best practices and co-regulatory commitments like those in the CoP on Disinformation.

In fact, however, many types of disinformation are illegal in certain EU Member States. The HLEG's definition – false information intentionally disseminated to cause harm or for profit – overlaps with several existing legal categories, such as defamation, false advertising or Holocaust denial. However, as detailed in a survey by researchers from the University of Amsterdam, many EU Member States have broader laws in place criminalising the dissemination of false information as such, if it meets certain additional conditions.[339] These may be quite broad: for example, potentially threatening public order (in France) or the conduct of elections (in Poland). Additionally, Member States including Spain and Hungary introduced broad new criminal prohibitions on spreading disinformation during the Covid-19 pandemic.[340] Given their vague and easily manipulable criteria for speech to be criminalised, allowing for broad and arbitrary censorship as well as intensified surveillance of online speech, these laws are generally questionable from a fundamental rights perspective.[341] This has been made clear by multiple ECtHR judgments relating to Poland's electoral disinformation laws.[342] Moreover, the criminalisation of disinformation in European democracies has global implications, as it can provide legitimation for authoritarian governments to implement similar laws.[343]

While a full examination and fundamental rights analysis of national speech laws is outside the scope of this study, highlighting the presence of such laws is essential because the DSA notice and takedown framework delegates the definition of illegal content to national law (see Article 3(h) DSA). This means that where platforms have been notified of the presence of such content, they lose their intermediary liability immunity under Article 6 and can face criminal or civil liability for hosting it, unless they remove it expeditiously. National disinformation laws thus enable not only law enforcement authorities but 'any individual or entity' (who can report content to platforms under Article 16) to notify platforms of content they consider to be illegal. This creates a risk of liability for the platform and thus a powerful incentive to remove it.

Importantly, this will still be the case – at least to some extent – even if the national laws in question are rarely enforced. Platforms face few consequences for removing legal content, but face uncertain regulatory and liability risks for failing to remove it. As a result, their incentives will generally weigh in

---

338 Van Hoboken, J., and Ó Fathaigh, R., 'Regulating Disinformation in Europe: Implications for Speech and Privacy', *UC Irvine Journal of International, Transnational, and Comparative Law*, Vol. 6, No. 1, May 27, 2021, p. 9.

339 Ó Fathaigh, R., et al., 'The Perils of Legally Defining Disinformation', *Internet Policy Review*, Vol. 10, No. 4, November 4, 2021.

340 Bayer, J., et al., *The Fight against Disinformation and the Right to Freedom of Expression*, European Parliament, Brussels, 2021 https://www.europarl.europa.eu/RegData/etudes/STUD/2021/695445/IPOL_STU(2021)695445_EN.pdf.

341 Ó Fathaigh, R., et al., 'The Perils of Legally Defining Disinformation', *Internet Policy Review*, Vol. 10, No. 4, November 4, 2021.

342 Judgment of 8 October 2008 of the European Court of Human Rights, *Kita v. Poland*, application no. 57659/00; Judgment of 25 July 2019 of the European Court of Human Rights, *Brzeziński v. Poland*, application no. 47542/07).

343 Canaan, I., 'NetzDG and the German Precedent for Authoritarian Creep and Authoritarian Learning', *Columbia Journal of European Law*, Vol. 28, pp. 101-133.

favour of removing potentially illegal content – even where it is unclear whether it falls within the scope of the legal prohibition, or where it is unlikely that the user posting the content would face charges.

This is illustrated by the operations of 'Internet Referral Units' (IRUs) which operate within Europol and many national police forces. It has become common practice for IRUs to report content to platforms as illegal or incompatible with their contractual content policies via informal channels, rather than issuing formal removal orders.[344] While IRUs have historically focused primarily on terrorism-related content, recent research suggests that at least in the UK, since the beginning of the Covid-19 pandemic, multiple government agencies now have the remit of monitoring pandemic-related disinformation content, or other content considered to undermine the government's Covid-19 policies, and flagging it to platforms for removal.[345] Although platforms will typically not have any legal obligation to remove such content, or these obligations may be unclear, independent research and a recent decision by the Meta Oversight Board indicate that informal requests like these are often highly effective in having content removed.[346] Where vague national disinformation laws enable law enforcement authorities to make a plausible case that content could be illegal, it is likely that platforms will remove it on request to avoid any risk of liability. Even where the content is clearly not illegal, as with much Covid-related disinformation, platforms' close relationships with IRUs - which often have 'trusted flagger' status - make it likely that they will voluntarily remove content flagged under their community standards.[347] At the same time, informally flagging content through the same channels available to any other user allows state institutions to circumvent the legal safeguards associated with more formal removal orders.

Removal of disinformation content which directly threatens democratic processes or the rights of others will sometimes be justified from a fundamental rights perspective, as discussed in Section 4.2.2. However, censoring online content purely on the basis that it is false, or on the basis of other vague criteria such as potential threats to public order, raises serious concerns about freedom of expression, information, and democratic debate. By deferring to national law to define 'illegal content', while also allowing national authorities to use informal back-channels to request that platforms remove legal content, the ECD/DSA regime creates high risks of arbitrary and unaccountable censorship of content deemed harmful or threatening to public order by state authorities. This might often include protest, activism or political dissent, which enjoy particularly strong protection under international law on freedom of expression.

Strengthening fundamental rights safeguards in the DSA's notice and takedown system should be a priority for EU legislators. This should involve narrowing the definition of illegal content (for example to provide that platforms' intermediary liability immunity can still apply to known illegal content, if it does not directly threaten the rights of others or other important public interests) and requiring content removal requests from state institutions to use formal channels (such as Article 9 DSA) which are subject to legal safeguards.

---

[344] Chang, B., 'From Internet Referral Units to International Agreements: Censorship of the Internet by the UK and EU', *Columbia Human Rights Law Review*, NY, October 31, 2017; Bloch-Wehba, H., 'Content Moderation as Surveillance,' *Berkeley Technology Law Journal, Vol. 36, Iss. 3,* 2022, pp. 1297-1340.

[345] Big Brother Watch, *Ministry of Truth: The secretive government units spying on your speech*, n.d. 2023 https://bigbrotherwatch.org.uk/wp-content/uploads/2023/01/Ministry-of-Truth-Big-Brother-Watch-290123.pdf.

[346] Meta Oversight Board, UK drill music 2022-007-IG-MR.

[347] Appelman, N., and Leerrssen, P., 'On "Trusted" Flaggers', *Yale-Wikimedia Initiative on Intermediaries & Information*, July 12, 2022, https://law.yale.edu/sites/default/files/area/center/isp/documents/trustedflaggers_ispessayseries_2022.pdf ; Big Brother Watch, *Ministry of Truth: The secretive government units spying on your speech*, n.d. 2023, https://bigbrotherwatch.org.uk/wp-content/uploads/2023/01/Ministry-of-Truth-Big-Brother-Watch-290123.pdf .

4.5.2. Due diligence obligations under the DSA

Beyond moderation of specific types of content, the DSA creates due diligence obligations relating to how platform companies run their moderation systems and other technical and operational processes.[348] Most significantly, very large online platforms (with over 45 million EU users) will now be under wide-ranging obligations to assess and mitigate 'systemic risks' to various public interests specified in Article 34 DSA: these include the dissemination of illegal content, fundamental rights, civic discourse and electoral processes, security and public health. In these respects, disinformation regarding politics, public health, minority social groups, and other issues discussed in Section 4.3 is obviously highly relevant. Platforms will thus be obliged to consider how their services could create risks by facilitating the spread of mis- and disinformation; formally assess these risks at least once a year, and before deploying new technical functions (Article 34(1)); have their risk assessment and mitigation measures independently audited (Article 37); and submit the resulting reports to the Commission (Article 42(4)).

These due diligence obligations will be an important element of the EU policy response to disinformation, in particular because they go beyond a narrow focus on individual pieces of harmful content. They require platforms to more holistically assess how their business, technical, and design decisions contribute to the spread of disinformation, and will enable the Commission to oversee their actions in this regard.[349] Existing experiences have shown that design changes can be an effective way of addressing disinformation, but that platforms have often lacked incentives to pursue them, especially in a systematic or consistent way. For example, journalistic investigations have shown that Meta teams repeatedly proposed and tested changes to Facebook's recommender algorithms which performed well in decreasing exposure to disinformation and extreme content, but which were then rejected by company executives because they were seen as disadvantageous from a business perspective.[350] Even where companies have pursued design interventions to address disinformation, they have often ignored academic research and best practices from other industries, such as cybersecurity, and continued pursuing solutions which are known to be ineffective or suboptimal.[351] Under the DSA, however, the Commission could threaten platforms with fines for ignoring known risks or refusing to implement risk mitigation measures known to be effective.

However, researchers have also pointed out potential weaknesses in the DSA's systemic risk framework, in particular the vague and flexible nature of the obligations it creates. Since there are many possible ways to interpret the relevant systemic risks and appropriate risk mitigation measures, platform companies may be able to perform compliance by implementing only superficial changes in their decision-making procedures.[352] Independent audits may also provide only superficial accountability, as there are no established substantive standards that auditors could apply to check whether platforms

---

[348] Husovec, M., and Roche Laguna, I., 'Digital Services Act: A Short Primer', *Principles of the Digital Services Act*, Oxford, Oxford University, 2023, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4153796.

[349] Jaursch, J., 'Strengthening EU proposals on deceptive platform design', *Stiftung Neue Verantwortung - Policy Briefs,* 2022.

[350] Hao, K., 'How Facebook got addicted to spreading misinformation', *MIT Technology Review,* March 11, 2021, https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/ Merril, J., and Oremus, W., 'Five Points for Anger, One for a 'Like': How Facebook's Formula Fostered Rage and Misinformation', *The Washington Post,* Oct. 26, 2021, https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/.

[351] Kaiser B., et. al., 'Warnings That Work: Combating Misinformation Without Deplatforming', *Lawfare*, Friday, July 23, 2021, https://www.lawfareblog.com/warnings-work-combating-misinformation-without-deplatforming.

[352] Griffin, R., 'The Sanitised Platform', *JIPITEC*, Vol. 3, No.1, 2022, https://www.jipitec.eu/issues/jipitec-13-1-2022/5514/citation.

are taking adequate action to mitigate risks, beyond checking the basic accuracy of their reports.[353] Finally, since platforms have the discretion to choose how they mitigate risks, there is no guarantee that they will choose the most effective measures or those that best balance competing fundamental rights: for example, they might choose to simply roll out more automated moderation, because it is cheap and easily scalable, instead of making more fundamental design changes.[354] The Commission could help counteract these risks through an active oversight strategy. For example, it could make clear to platforms that they must assess and consider risk mitigation measures including design and operational changes in order to be compliant with Article 35. Through developing codes of conduct and industry best practices, it can also develop more concrete and substantive standards that auditors and platforms can apply.[355]

Article 36 DSA also creates a new 'crisis protocol' which can be activated by the Commission 'where extraordinary circumstances lead to a serious threat to public security or public health in the Union or in significant parts of it'. In such circumstances, very large platforms can be required to take additional measures to mitigate such threats; the choice of measures is in the first instance up to the platform, but the Commission can oversee them and decide whether they are adequate. This provision was introduced during the Ukraine war and, although it does not only relate to disinformation, disinformation relating to future pandemics or conflict situations are obvious contexts in which it might be used. Nonetheless, the potential scope of the provision is very broad and it has been heavily criticised by civil society, since it could enable the Commission to implement far-reaching restrictions of freedom of expression, while evading public scrutiny and legal accountability, as these restrictions would be implemented by platforms rather than by public authorities.[356]

In addition, as Section 4.4.1 described, stepping up content moderation efforts has proved limited as a way of dealing with previous crises such as the Covid-19 pandemic, and comes at a significant cost for freedom of expression. Disinformation policy should rather focus on establishing effective platform integrity and security teams which can counter systematic and organised disinformation campaigns, tackling platform design features which encourage or facilitate the spread of disinformation, and strengthening trustworthy and independent media more broadly – all of which require sustained, longer-term efforts, rather than emergency responses. The Commission's enforcement efforts should focus on using regulatory measures like Article 34 to compel platforms to implement better safety measures and design practices over the long term, rather than on instant responses to crises which could severely compromise fundamental rights.

4.5.3. The Code of Practice on Disinformation

One way of specifying and strengthening platforms' systemic risk obligations is through the establishment of industry codes of conduct, which can be taken into account under Article 45 DSA when determining whether very large online platforms have appropriately identified and mitigated systemic risks. The Commission encouraged major platforms to agree to a self-regulatory Code of

---

[353] Laux, J., et. al., 'Taming the few: Platforms regulation, independent audits, and the risks of capture created by the DMAA and DSA,' *Computer Law & Security Review, Vol. 43,* 2021, https://doi.org/10.1016/j.clsr.2021.105613.

[354] Barata, J., 'The Digital Services Act and Its Impact on the Right to Freedom of Expression: Special Focus on Risk Mitigation Obligations', *DSA Observatory*, 2021, https://dsa-observatory.eu/2021/07/27/the-digital-services-act-and-its-impact-on-the-right-to-freedom-of-expression-special-focus-on-risk-mitigation-obligations/.

[355] Vander Maelen, C., "Hardly Law or Hard Law? Investigating the Dimensions of Functionality and Legalisation of Codes of Conduct in Recent EU Legislation and the Normative Repercussions Thereof." *European Law Review*, vol. 47, no. 6, Sweet & Maxwell, 2022, pp. 752–72.

[356] European Digital Rights, '*On New Crisis Response Mechanism and Other Last Minute Additions to the DSA*,' EDRi, April 12, 2022, https://edri.org/wp-content/uploads/2022/04/EDRi-statement-on-CRM.pdf.

Practice on Disinformation in 2018. A significantly expanded and updated version was agreed in 2022 and will be incorporated into the DSA under Article 45.[357] This means that failure to comply with its commitments could be an indicator of non-compliance with the risk mitigation obligations in Article 35 DSA, while following these commitments could be used to demonstrate compliance. As a result, while the 2018 Code was widely regarded as relatively ineffective due to its lack of concrete enforcement,[358] the 2022 Code should have more regulatory force.[359]

The updated Code places heavy emphasis on increasing accountability through transparency: for example, platforms commit to publicly reporting how they assess disinformation in paid adverts (Commitment 2), and how they enable users to identify manipulated media such as deepfakes (Commitment 15). These commitments appear aimed at incentivising platforms to take more voluntary action to address disinformation by enabling more public scrutiny, allowing independent researchers and civil society to criticise platforms whose safety measures appear inadequate. In this respect, they will be complemented by platforms' commitments under the Code to provide enhanced access to data for researchers (Chapter VI) and the DSA's new obligations to provide requested internal data to vetted researchers (Article 40).

However, the Code also regulates platform design and decision-making processes more directly. Commitment 18 requires them to adopt 'safe design' practices, such as adapting their recommender systems to downrank disinformation and pre-testing new features. Platforms commit to investing, participating in and publishing ongoing research into safe design practices, and to reporting on how they are adapting and improving their services in light of such research findings. An illustrative example of this kind of design intervention is presented in Box 8. Failure to do so could in turn be grounds for a finding of non-compliance with Article 35 DSA, which should provide strong regulatory incentives for platforms to strengthen internal oversight and trust and safety efforts – provided that these obligations are actively enforced by the Commission.

In establishing and evaluating compliance with industry best practices, EU policy should give more attention and support to the emerging professionalisation of 'trust and safety' for online platforms. Responding to rapidly evolving disinformation trends and tactics requires a flexible and adaptable response which cannot be wholly prescribed by regulators but must be led by the industry itself. At the same time, experiences such as Twitter's abrupt rollback of its AI ethics and safety programmes,[360] and Meta executives' refusal to implement anti-disinformation measures which could compromise revenue,[361] suggest that leaving this to companies' voluntary security and corporate social responsibility programmes produces a response which is at best patchy and selective. Developing

---

[357] Strengthened Code of Practice on Disinformation 2022, Chapter I(i).

[358] European Commission, SWD (2020)180 Final - Assessment of the Code of Practice on Disinformation, 2020, available at: https://digital-strategy.ec.europa.eu/en/library/assessment-code-practice-disinformation-achievements-and-areas-further-improvement ; Sander, B., 'Democratic Disruption in the Age of Social Media: Between Marketized and Structural Conceptions of Human Rights Law,' *European Journal of International Law*, Vol. 32, Issue 1, 2021, p. 159–193, https://doi.org/10.1093/ejil/chab022.

[359] Vander Maelen, C., "Hardly Law or Hard Law? Investigating the Dimensions of Functionality and Legalisation of Codes of Conduct in Recent EU Legislation and the Normative Repercussions Thereof." *European Law Review*, vol. 47, no. 6, Sweet & Maxwell, 2022, pp. 752–72.

[360] Knight, W., 'Elon Musk Has Fired Twitter's 'Ethical AI' Team', *Wired*, 2022, available at: https://www.wired.com/story/twitter-ethical-ai-team/?utm_brand=wired-science&mbid=social_tw_sci&utm_source=twitter&utm_social-type=owned&utm_medium=social; Newton, C., and Schiffer, Z., 'Twitter, cut in half', Platformer News, 2022, available at: https://www.platformer.news/p/twitter-cut-in-half?utm_campaign=post; Delcker J., ' Twitter's sacking of content moderators raises concerns', DW, 2022 https://www.dw.com/en/twitters-sacking-of-content-moderators-will-backfire-experts-warn/a-63778330.

[361] Hao, K., 'How Facebook got addicted to spreading misinformation', *MIT Technology Review*, 11 March 2021.

professional standards and networks which are founded on the expertise and experience of people working directly in the industry, but which span multiple platforms and policy areas, offers an alternative way to coordinate responses and establish best practices which are less beholden to the business interests of any particular company. Some recent developments in the professionalisation of trust and safety work are presented in Box 7.

Such organisations should be supported to develop best practices, ethical standards and granular research expertise which can guide platforms' responses to disinformation and help regulators assess compliance with risk mitigation obligations. They should also be prominently involved in consultations and stakeholder dialogues in relation to the CoP on Disinformation and other co-regulatory measures under the DSA.

Box 7: Trust and safety professional associations

'Trust and safety professional' is a catch-all term for a new category of professionals whose role is to ensure users' safety online. It encompasses not only the numerous content moderators who review and/or remove material which does not comply with the platform's policies, but also various other specialists working on issues such as fraud and supporting law enforcement. A more recent and ongoing development is the increasing professionalisation and institutionalisation of the trust and safety profession. There are now several independent associations and think tanks bringing together people working in trust and safety at different platforms in order to develop strategies for tackling evolving challenges and promoting online safety.In 2018, the Trust & Safety Professional Association and the Centre for Humane Technology were founded. The former was the first global association for professionals working on safety and integrity at online platforms, encompassing content moderation, design and engineering, security and policy staff. The latter is a nonprofit organisation established by former 'big tech' employees to educate the public, advise legislators, and train technologists to address risks in the digital world. In 2021, ex-Meta staff founded the Integrity Institute, a network of current and former trust and safety professionals which operates an independent think tank and research institute. Also in 2021, leading tech companies including Meta, Google, Microsoft and LinkedIn founded the Digital Trust & Safety Partnership, an industry-led initiative for the development of best practices in trust and safety.

While the development of best practices by tech companies themselves will be an important component of the DSA framework, independent industry-wide professional associations for current and former tech workers offer a promising way to develop and share knowledge, best practices, and research which is informed by industry experience but more insulated from companies' economic incentives. For example, the Integrity Institute publishes regular research on issues such as the amplification of misinformation on different platforms, and recently contributed an amicus brief to the US Supreme Court. In the longer term, professional associations could also serve as a forum for the development of ethical standards and safe design practices for trust and safety professionals. This could strengthen the ability of platform staff to advocate for safe and ethical practices within companies. Supporting the development and growth of such associations, and facilitating collaboration between professional associations, civil society and regulators, could be a promising way for the EU to strengthen accountability and knowledge sharing in the implementation of the DSA and other tech regulations.

As platforms, products, and threats are evolving, professional associations can contribute both theoretical and practical expertise. For example, taking a more theoretical approach, the Center for Humane Technology has set out a series of principles which they suggest should guide policymakers in drafting digital regulation so as to ensure that future technology will respect people's attention, improve their wellbeing, and strengthen communities. Importantly, they suggest a precautionary

approach to technology that should involve a pre-emptive assessment of the risks a product or feature could pose to different social groups. In this context, expertise from current and former industry professionals could aid in developing and implementing regulatory safeguards such as AI audits (discussed in Box 2) and 'safe design practices' (discussed in Box 5). The Center also advises governments to find sustainable solutions that can adapt to technological evolution over time, as well as taking into account the complexity of the environment in which social media are used. To achieve this, regulators would benefit from drawing on the experience and insights of professional associations, for example through stakeholder consultations.

On a more practical note, the Integrity Institute has published a set of best practices for responsible algorithm and platform design. For example, drawing on established practices in areas like search engines, they provide some guidance on how recommendation algorithms could be designed to prioritise content quality over engagement, reducing the promotion of harmful or untrustworthy content. They also highlight established design mechanisms to prevent malicious users from exploiting ranking systems and sharing harmful content: for example, limiting new users' ability to reach large audiences, limiting posting of the same or very similar content across many spaces, and creating barriers to the use of multiple accounts. They also argue that platforms should be more transparent about their quality metrics and integrity measures, including sharing protocols for changes during special events (e.g. elections). Best practices and guidelines like these could, for example, provide a basis for guidance on platforms' due diligence obligations under the DSA and their commitments under the Code of Practice on Disinformation. This will allow regulators to provide concrete detail on due diligence and safe design practices which is informed by industry expertise and will be more flexible and easier to update than prescriptive regulation.

Sources: Trust & Safety Professional Association, 'About Us', Trust & Safety Professional Association, https://www.tspa.org/about-tspa/; Lapowsky, I., 'Jeff Allen, Sahar Massachi Launch Integrity Institute', *Protocol*, 2021 https://www.protocol.com/policy/integrity-institute; 'Misinformation Amplification Analysis and Tracking Dashboard', Integrity Institute, 2022, https://integrityinstitute.org/our-ideas/hear-from-our-fellows/misinformation-amplification-tracking-dashboard; Davy, J., 'Amicus Brief for Gonzalez v Google', Integrity Institute, 2022 https://integrityinstitute.org/amicus-brief-for-gonzalez-v-google; Center for Humane Technology, 'Who we are', n.d. https://www.humanetech.com/who-we-are#team; Center for Humane Technology, 'Policy Principles', n.d. https://www.humanetech.com/policy-principles.

The Code of Practice also includes numerous commitments aimed at tackling commercial disinformation campaigns in two key aspects. First, it aims to prevent platforms from carrying adverts that themselves spread disinformation. Second, it aims to prevent them from running adverts alongside disinformation content or on websites spreading disinformation, which would enable publishers to profit from such content. To this end, platforms commit to developing policies to enhance oversight of ad content and placements. In addition, the Code includes commitments from key advertising industry bodies and other intermediaries, such as adtech companies, to strengthen 'brand safety' efforts through which they endeavour to exclude disinformation content from ad placements.

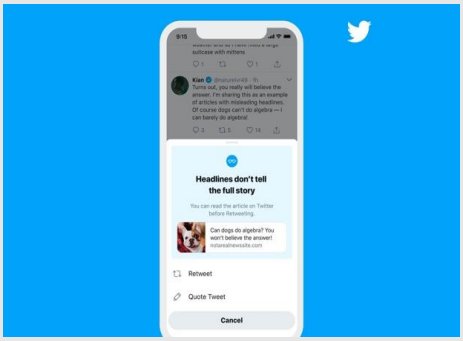
While addressing commercial incentives to spread disinformation may sound appealing, it should be recognised that these efforts will face many of the same limitations and drawbacks as content moderation. Evaluating all adverts and all content running with adverts to identify disinformation content, at the scale of contemporary social media platforms, is a huge challenge. Automated tools to analyse and classify content – which are already widely used by major platforms and advertising companies to restrict ad placements – remain limited, biassed and unreliable. Delegating assessments of what content is 'safe' to advertisers and adtech companies also creates an obvious possibility that their evaluations of disinformation will be more shaped by their own commercial incentives and assessments of what is good for their brand than by the public interest. Existing brand safety tools have

been shown to frequently demonetise content related to LGBTQ+ and political issues, which many advertisers consider to reflect negatively on their brands.[362] As such, encouraging more use of demonetisation and delegating assessments of safety or risk to advertising industry actors threatens freedom of expression, media pluralism and non-discrimination rights, while bringing uncertain benefits in terms of tackling disinformation.

Box 8: Behavioural prompts and friction on Twitter

In 2020, Twitter began testing behavioural prompts – similar to those described in Box 3 – which aimed to discourage users from sharing articles without reading them. Where a user attempts to retweet a tweet containing a link, without first clicking on the link, they are first shown a prompt asking if they want to read the article first (though this does not prevent them from retweeting without reading, if they choose to).

In light of the empirical research about susceptibility to misinformation reviewed in Section 4.3.1, this appears a promising intervention. People will often share information without much concern for whether it is true or false, as a way of signalling their identity or political sympathies to others. However, adding 'friction' through prompts like these can encourage more conscious consideration of whether the information is reliable. Research suggests that this type of friction, which forces users to actively make a decision – even if that just requires clicking through a prompt – are more effective in changing user behaviour and encouraging them to seek more information than just adding contextual warning labels. At the same time, these interventions do not need to target any particular type of content or make assessments of truth or falsity, and the interference with users' freedom of expression is minimal.



Source: Twitter, 2020

At the same time, there is currently insufficient evidence available to make firm judgments about the utility of such measures. Twitter has claimed that, in its initial tests, 33% more people read articles before retweeting them, and an unspecified number of people chose not to retweet them at all. However, the company has not made any more information available on how the feature has performed since then, or how it has affected the spread of mis- and disinformation. In contrast to the offensive comment prompts discussed in Box 3, the company has not published detailed research results, nor shared data with independent researchers.

Following the introduction of the data sharing obligations in the DSA and the Code of Practice on Disinformation, as well as the commitments to research and implement safe design practices in the Code, there should be more independent scrutiny of design interventions like these. Regulators

[362] Kumar, S., 'The Algorithmic Dance: YouTube's Adpocalypse and the Gatekeeping of Cultural Content on Digital Platforms', *Internet Policy Review*, Vol. 8, No. 2, June 30, 2019.; Cunningham, S., and Craig, D., 'Creator Governance in Social Media Entertainment', *Social Media + Society*, Vol. 5, No. 4, October 2019, p. 205630511988342.

should be able to demand that very large platforms implement and expand such interventions as part of their risk mitigation obligations under Article 35 DSA.

Sources: Vincent, J., 'Twitter is bringing its 'read before you retweet' prompt to all users / Don't tl;dr that article', The Verge, 2020, , available at: https://www.theverge.com/2020/9/25/21455635/twitter-read-before-you-tweet-article-prompt-rolling-out-globally-soon.

4.5.4. Political advertising

Another key element of the EU's policy response to disinformation has been intensified regulation of political advertising on social media.[363] Online political advertising raises specific concerns because of the possibility of targeting small and precisely defined groups of people, based on personal and behavioural data, rather than larger audiences as with traditional advertising methods. Although targeted political advertising does not necessarily overlap with disinformation, it can in some cases be used to spread disinformation,[364] and the Commission has made this a key focus of EU disinformation policy.[365]

Additionally, even where targeted political adverts are not used to spread false information, they still raise many of the same normative concerns as disinformation content. How much they directly influence voters' behaviour remains uncertain and debated: the empirical evidence is mixed and gives only an incomplete picture, as Box 9 outlines in detail. Crucially, however, political advertising raises wider normative concerns, beyond the direct manipulation of electoral outcomes. Since adverts are only shown to narrowly-defined audience segments, it is difficult to see the whole picture of who is paying to contact voters and what messages political figures are promoting to whom.[366] This undermines accountability in political processes. By targeting specific social groups, politicians can play on social divisions and exclude those who might lose out from their policies from seeing their messages, undermining equal participation in political debate.[367] Since social media platforms determine which people within the potential audience will see an advert based on algorithmically predicting who is most likely to engage with it, these effects can arise even where political actors placing ads do not target narrowly-defined population segments - the subset of the target audience who actually sees the ad might nonetheless be highly unrepresentative.[368] Finally, the fragmentation of political messages across different sectors of society could also create 'second-order effects' such as weakening trust in politicians and solidarity across the electorate.[369]

---

[363] Nenadić, I., 'Unpacking the "European approach" to tackling challenges of disinformation and political manipulation,' *Internet Policy Review* Vol. 8. Issue 4., 2019, available at: https://policyreview.info/articles/analysis/unpacking-european-approach-tackling-challenges-disinformation-and-political.

[364] NYU Tandon School of Engineering, Cybersecurity for Democracy and Global Witness, 'TikTok and Facebook Fail to Detect Election DIsinformation in the US, While YouTube Succeeds', *Cybersecurity for Democracy*, October 2022, available at: https://cybersecurityfordemocracy.cdn.prismic.io/cybersecurityfordemocracy/390e0f2e-2818-4210-92fc-61922140e8f9_Election+disinformation+on+social+media+in+the+midterms+-+Global+Witness_C4D_Oct22.pdf.

[365] European Commission, 'Guidance on Strengthening the Code of Practice on Disinformation', Brussels, 2021, available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52021DC0262&qid.

[366] Dobber, T., et. al., 'The regulation of online political micro-targeting in Europe', *Internet Policy Review* Vol. 8., Issue 4., 2019.

[367] Dobber, T., et. al., 'The regulation of online political micro-targeting in Europe', *Internet Policy Review* Vol. 8., Issue 4., 2019; Keller, C.I., 'Don't Shoot the Message: Regulating Disinformation Beyond Content', *Direito Público*, Vol. 18, No. 99, 2021, pp. 486–515.

[368] Foronda, F.H., and Iwańska, K., 'A thousand Facebook-Cambridge Analytica scandals every day', *Euractiv*, February 2, 2023, available at: https://www.euractiv.com/section/digital/opinion/a-thousand-facebook-cambridge-analytica-scandals-every-day/.

[369] Keller, C.I., 'Don't Shoot the Message: Regulating Disinformation Beyond Content', *Direito Público*, Vol. 18, No. 99, 2021, pp. 486–515.

Box 9: How political advertising affects voter turnout

Social media users are constantly bombarded with information, whether it is advertising, news, or interactions with other users. However, questions about how far this information actually affects people's beliefs and behaviour remain unresolved. One of the most debated questions - with implications for EU policies such as the proposed Political Advertising Regulation - has been the impact of political advertising on voter behaviour.

In the early 2000s, multiple US studies suggested that exposure to campaign advertising increases the likelihood that people would vote (see e.g. Freedman et al. and Hillygus below). More recently, many scholars have investigated the impacts of social media adverts on voting behaviour - although the topic has not been researched in the EU as extensively as in the US. Notably, one recent study (Tappin et al., below) found that political adverts on social media which were targeted according to users' individual characteristics were significantly more effective in influencing political adverts than ads targeted randomly within broad audiences. However, other studies have found little evidence for impacts on voting behaviour. One study focusing on Texas voters during the 2018 US midterm elections (Thomsen, below) found that micro-targeted Facebook ads on four issues - abortion rights, health care, immigration, and gun control - had a slight impact on voting behaviour. However, the impact of these ads varied depending on the salience of the debated issue. Another recent study (Aggarwal et al., below) involved two million moderate- and low-information persuadable voters in five battleground states during the 2020 US presidential election. During eight months, the cohorts were exposed to social media advertising designed to persuade them into voting for Joe Biden and against Donald Trump. The authors found no evidence that this advertising campaign increased or decreased average turnout.

This research offers various insights. First, the salience of the debated issue, and the election in question, could affect the potential influence of political advertising. Second, voters who are better informed about the campaign and the candidates are less likely to be influenced. The type of advert, the targeted audience, and the political issue involved could all affect the capacity of targeted adverts to influence voters. This suggests that more research is needed into their effects in different local and political contexts, including in Europe.

Sources: Freedman, P., et al., 'Campaign Advertising and Democratic Citizenship', *American Journal of Political Science, Vol. 48, No. 4*, October 2004, pp. 723–741; Hillygus, D.S., 'Campaign Effects and the Dynamics of Turnout Intention in Election 2000', *The Journal of Politics, Vol. 67, No. 1*, February 2005, pp. 50–68; Tappin, B.M., et al., 'Quantifying the Persuasive Returns to Political Microtargeting', *PsyArXiv Preprints*, 2022 , available at: https://psyarxiv.com/dhg6k/; Thomsen, I., 'Do Facebook Ads Win Elections? It's Complicated.', *Northeastern Global News*, March 8, 2022 , available at: https://news.northeastern.edu/2022/03/08/facebook-ad-campaigns-voter-turnout; Aggarwal, M., et al., 'A 2 Million-Person, Campaign-Wide Field Experiment Shows How Digital Advertising Affects Voter Turnout', *Nature Human Behaviour*, January 12, 2023.

Additionally, even where targeted political adverts are not used to spread false information, they still raise many of the same normative concerns as disinformation content. How much they directly influence voters' behaviour remains uncertain and debated: the empirical evidence is mixed and gives only an incomplete picture, as Box 9 outlines in detail. Crucially, however, political advertising raises wider normative concerns, beyond the direct manipulation of electoral outcomes. Since adverts are only shown to narrowly-defined audience segments, it is difficult to see the whole picture of who is paying to contact voters and what messages political figures are promoting to whom. This undermines accountability in political processes. By targeting specific social groups, politicians can play on social divisions and exclude those who might lose out from their policies from seeing their messages, undermining equal participation in political debate. Since social media platforms determine which people within the potential audience will see an advert based on algorithmically predicting who is

most likely to engage with it, these effects can arise even where political actors placing ads do not target narrowly-defined population segments - the subset of the target audience who actually sees the ad might nonetheless be highly unrepresentative. Finally, the fragmentation of political messages across different sectors of society could also create 'second-order effects' such as weakening trust in politicians and solidarity across the electorate.

For all these reasons, the Commission's 2022 proposal for a Regulation on the Transparency and Targeting of Political Advertising is a positive step. At the time of writing, however, negotiations are ongoing and many questions have not been resolved (an in-depth analysis can be found in a 2021 report published on behalf of the European Audiovisual Observatory[370]). However, as the title suggests, the regulation focuses on two main areas: transparency and targeting.

First, it aims to strengthen transparency in all stages of the value chain for political advertising services. Advertisers themselves and all service providers involved are required to document the nature, spending and targeting of political advertising campaigns, and make these records available to regulators and independent researchers. Users must also be clearly informed that they are seeing a political advert and on whose behalf it is placed. Second, it restricts the targeting of political adverts based on sensitive personal data (e.g., race, religion, sexual orientation). While the GDPR already limits the processing of such data, the Proposal would restrict its use for political advertising even further, permitting it only where there is explicit consent or where the processing is by a non-profit organisation and relates to former members or regular contacts. The Proposal would establish a broad definition of political advertising: content could qualify either because it is published by or behalf of one of the categories of political actor listed in Article 2(4), or because its content and context mean it is liable to influence voting processes or behaviour.

This is a promising effort to establish a holistic regulatory framework for targeted political advertising, which does not just regulate false information or particular types of content, but engages more broadly with the effects of narrowly targeted political messaging on democratic debate and participation. However, current proposals raise some questions with regard to fundamental rights and freedom of political debate. Notably, the definition of political advertising advocated by the Council is broad enough to extend to political messages disseminated through commercial platforms, even if the publisher does not pay to distribute them.[371] Civil society organisations have raised concerns that their non-commercial activities could be subject to onerous transparency requirements, hampering their ability to participate in public debate.[372]

In addition, the current restrictions on direct use of sensitive data in targeting political adverts appear too narrow to have much impact. This is because adverts can effectively – whether intentionally or unintentionally on the part of the advertiser – target or exclude certain groups based on race, political views, sexuality etc., without directly referring to that data, as there are many proxy values (e.g., neighbourhood, friend group, cultural tastes) which correlate strongly with these characteristics.[373]

---

[370] Cappello, M. (ed.), 'New actors and risks in online advertising,' *IRIS Special 2022-1*, European Audiovisual Observatory, Strasbourg, 2022.

[371] Council of the European Union, 'Transparency and targeting of political advertising: Council agrees its negotiating mandate',

Council of the European Union, 13 December 2022, available at: https://www.consilium.europa.eu/en/press/press-releases/2022/12/13/transparency-and-targeting-of-political-advertising-council-agrees-its-negotiating-mandate/.

[372] Stiftung Neue Verantwortung e. V. et. al, 'Open Letter: EU must protect fundamental freedoms for online political speech,' Algorithm watch, November 29, 2022, available at: https://algorithmwatch.org/en/open-letter-online-political-speech/.

[373] Griffin, R., 'Tackling Discrimination in Targeted Advertising: US regulators take very small steps in the right direction- but where is the EU?', *Verfassungsblog,* June 23, 2022, available at: https://verfassungsblog.de/targeted-ad/.

Thus, even if such sensitive data cannot be used for targeting, ads can still be targeted and delivered to precisely segmented groups, which may correspond closely to protected political or social groups. This has implications not only for those groups, but for equal and open participation in political debate by society as a whole. In light of this, more consideration should be given to current proposals to entirely ban personalised targeting of online political adverts, except based on certain broad characteristics like location.[374]

Alongside the proposed regulation, as noted in Section 4.5.3, advertising content and placements are already dealt with in the updated CoP on Disinformation. The Code generally aims to prevent the use of adverts to disseminate or monetise disinformation, but Section III of the Code also contains numerous commitments relating specifically to political advertising. Many of these overlap substantially with the Political Advertising Regulation, focusing on enhancing transparency towards users and regulators about who pays for adverts and how they are targeted. However, the commitments in the CoP are in some respects more concrete, with a greater focus on design practices and how platforms' obligations should be implemented and monitored in practice.

For example, platforms commit to sharing their labelling designs for adverts and researching the effectiveness of different labelling approaches, and to engaging with researchers to ensure the data and APIs (application programming interfaces – software tools which allow researchers to access platform data) they provide are presented in ways that are useful. Research suggests that tweaking the design and implementation of warning labels can significantly impact their effectiveness against disinformation, and that there is significant scope for platforms to improve their current practices based on empirical research.[375] Equally, when it comes to data sharing, the details of how data is presented significantly impacts its usefulness to journalists, civil society and other actors who can hold politicians accountable.[376] Establishing and enforcing specific, detailed best practices and success metrics will thus be an important way to strengthen the impact of the DSA and Political Advertising Regulation.[377] Regulators should prioritise recruiting staff with sufficiently detailed knowledge and experience of UX design to achieve this.[378]

4.6. Recommendations

a. DSA enforcement

- Building on the obligations and commitments already established in the DSA and Code of Practice, the Commission and national DSCs should issue guidance stating that safe design practices should be a primary line of defence against disinformation, and should be prioritised over content moderation except where disinformation directly endangers the public or threatens the rights of others.
- In overseeing and enforcing very large platforms' systemic risk mitigation obligations under Articles 34-35 DSA, the Commission should place significant weight on design changes and other interventions which aim to proactively discourage and prevent the occurrence of online

[374] Killen, M., 'Germany supports ban on personal data for political ads', *Euractiv*, September 7, 2022, available at: https://www.euractiv.com/section/digital/news/germany-supports-ban-on-personal-data-for-political-ads/.

[375] Kaiser, B., et. al., 'Warnings That Work: Combating Misinformation Without Deplatforming', *Lawfare*, Friday, July 23, 2021, available at: https://www.lawfareblog.com/warnings-work-combating-misinformation-without-deplatforming.

[376] Leerssen, P., et al., 'News from the Ad Archive: How Journalists Use the Facebook Ad Library to Hold Online Advertising Accountable', *Information, Communication & Society*, December 26, 2021, pp. 1–20.

[377] Jaursch, J., 'Strengthening EU proposals on deceptive platform design', *Stiftung Neue Verantwortung - Policy Briefs*, 2022.

[378] Pershan, C., and Sinders, C., 'Why Europe's Digital Services Act Regulators Need Design Expertise', *Tech Policy Press,* Dec. 22, 2022 https://techpolicy.press/why-europes-digital-services-act-regulators-need-design-expertise/.

hate speech, harassment and other systemic risks, as opposed to moderating or removing content retroactively. Risk assessments and audit reports which indicate that platforms are not investing in such proactive risk mitigation measures should not be regarded as compliant. The Commission should ensure that it has sufficient staff with expertise in UX/UI design to effectively assess compliance with these obligations.

- The Commission and national DSCs should also issue guidance stating that the obligation for platforms to enforce their content policies in a diligent, objective and proportionate manner under Article 14(4) DSA requires adequate moderation capacities in all languages widely spoken by their users, including adequate investment in competent moderation staff. All relevant moderation processes should be clearly and publicly documented to establish compliance.

b. Legislative reform

- The Commission should consider and consult on amending Articles 3 and 6 DSA to create a narrower and more fundamental-rights-compliant definition of 'illegal content' which can attract liability for platforms. For example, the amended DSA could specify that platforms retain their intermediary liability immunity even where they have knowledge of illegal content, except where that content creates a direct and specific threat to public safety or the fundamental rights of others.
- The Commission should positively consider proposals to entirely ban or very significantly restrict the personalised targeting of political advertising, recognising that microtargeting of political messaging has negative impacts for civic and political debate even where it does not infringe the rights of individual users.

c. Strengthening trust and safety

- Recognising that countering organised disinformation operations and other emerging threats requires flexible response capacities within the social media industry and civil society, EU policy should make it a priority to strengthen the online trust and safety profession. This should include support for professional associations of platform engineers and moderation staff and consultation with such organisations in the development of industry best practices and safety standards under the DSA.

d. Enhance media literacy, but with caution

- Through media campaigns, in schools, and in other civic spaces, the EU should promote and fund new and existing programmes which teach individuals about best practices to evaluate the reliability of online content, as well as identifying bots and strategically-promoted disinformation read and identify bots, potentially harmful and/or mis-informative content online.
- However, policymakers should not over-rely on media literacy as a solution. Not only does it emphasise individual agency and control over more consequential structural issues, research has also shown that it can backfire, as individuals may also learn to doubt trustworthy

content.[379] Media literacy education should be one component of a broader policy programme aimed at promoting a trustworthy information environment.

e. Promoting reliable independent media

- As detailed in Chapter 5 on media pluralism, the EU's disinformation policy should be part of a broader policy programme to strengthen independent journalism and trust in media, for example through funding programmes.
- Public media and independent journalism institutions across the EU should be supported to provide fact-checking services and to create easily shareable, accurate information on sensitive political topics (e.g. public health risks, conflict situations).

[379] Boyd, D. 'You think you want media literacy… do you?', *Data & Society: Points*, March 9, 2018, available at: https://points.datasociety.net/you-think-you-want-media-literacy-do-you-7cad6af18ec2.

5. MEDIA PLURALISM

5.1. Introduction

This chapter addresses the relationship between media pluralism and social media. As an significant aspect of freedom of expression and media freedom, media pluralism plays an important role in liberal democracies such as the EU. Indeed, public opinion is formed through constant confrontation and exposure to different points of view, which enables citizens to engage in public discussions and participate in the governance of their community. Consequently, access to diverse perspectives, editorial lines and analyses is essential for citizens to be able to discover and evaluate ideas, make informed choices, hold power to account and conduct their lives freely.

In the past two decades, the media ecosystem has evolved significantly, due in no small part to the internet and the rise of social media. As citizens increasingly use social media to access information and news, these changes affect many aspects of the news ecosystem. Important changes have been observed in the ways news is disseminated and consumed, but also in how it is produced. These changes create new opportunities, but also new risks including the spread and impact of disinformation and hate speech; the increasing influence of private technology companies over online communication; a lack of transparency in relation to how these companies algorithmically curate and moderate content; the polarisation of public debate; and the undermining of legacy news media and traditional journalism. These issues thus affect not only the production and dissemination of news, but the role of the media in the democratic process, in shaping public opinion, and in securing people's access to diverse information sources, with clear implications for the health of democracy.

To provide an overview of these challenges and how they are being addressed by the current European framework, this chapter proceeds as follows. Section 5.2 provides some necessary background on media pluralism and how the news ecosystem has changed in the context of the rise of social media. Section 5.3 outlines the challenges posed by social media to news business models, and how they have affected the dissemination and consumption of news and content. Section 5.4 analyses the influence of social media on news media markets, in particular as regards market concentration and challenges to local news. Section 5.5 outlines recent European developments and discusses the merits and limitations of the existing legal framework in relation to these issues, before Section 5.6 concludes with recommendations.

5.2. Background

5.2.1. Defining media pluralism

There is no single definition of media pluralism, but it generally refers to the importance of a media ecosystem that represents a variety of information, a diversity of opinions and different worldviews to inform public opinion.[380] Media pluralism also refers to a media ecosystem in which a variety of social actors and their needs and interests are represented and can influence public opinion on matters of public interest.[381] It is widely accepted that media pluralism is a democratic value, an enabler of other fundamental rights, and essential to the integrity of democratic discourses and procedures, as the

[380] Raeijmaekers, D. and Maeseele, P., 'Media pluralism and democracy: what's in a name?', *Media, Culture & Society,* Vol. 37, Issue 7, 2015 https://doi.org/10.1177/0163443715591.

[381] Raeijmaekers, D. and Maeseele, P., 'Media pluralism and democracy: what's in a name?', *Media, Culture & Society,* Vol. 37, Issue 7, 2015 https://doi.org/10.1177/0163443715591.

existence of a diverse and independent media is needed to hold policymakers and institutions accountable to the public. It is closely related to one of the core social purposes of freedom of expression: enabling citizens to engage in public discussion and thereby participate in the governance of their community.[382]

Media refers to the actors involved in the production and dissemination of content (information, analysis, opinion, entertainment, etc.), to potentially large numbers. It also includes the applications and infrastructures designed to facilitate the dissemination of news, when the providers of these services retain editorial or oversight control over the contents. As actors that participate in the dissemination and curation of content, social media platforms can thus be understood as media too (even if they have famously denied being media companies).

The EUI Media Pluralism Monitor defines media pluralism as comprising at least four areas: first, an area related to fundamental protections such as freedom of expression, and the right to seek, receive and impart information. Second, the protection and standardisation of the journalistic profession and access to traditional media and to the Internet. Third, market plurality, or the economical context in which market players operate. Market plurality elements include transparency of ownership, new media and platform concentration, media viability, and commercial and owner influence over editorial content. Fourth, social inclusiveness refers to access by minorities, local and regional communities, and women to the media, as well as media literacy and protection against illegal and harmful speech.[383] This chapter touches upon all of these areas as they are related to and affected by social media.

5.2.2. Media pluralism and digitisation: some background

No aspect of the media ecosystem has been unaffected by the wider adoption of the internet and the rise of social media. New telecommunications technology has always created moments of instability in the media environment – from the telegraph to the radio.[384] Social media played a part in changing the mass-media model that dominated the newspaper, radio and television industries by the late 20th century. This model was characterised by geographical industry concentration, and by a market-based model of production and dissemination of information, which relied mostly on property-like protections – such as state-issued licences or copyright protections - to incentivise, but also control the production of information.[385] Social media changed the incentives and methods for news production and dissemination for private or quasi-private actors. They created space for newcomers to reach wider audiences than would previously have been possible, but also shifted the economic incentives of news producers in ways that have arguably heightened concentration in the industry.

In Europe, the mass-media model was not only private but also involved a significant role for state-financed media and direct state intervention, especially in radio and television. Some of these elements remain today. As radio was first being commercialised, the British Broadcasting Company (BBC) was born private and wholly owned by Britain's radio manufacturers. It became a public corporation in the

---

[382] Bayer, J. et. al., *European Report, The Fight against disinformation and the right to freedom of speech,* European Parliament's Committee on Civil Liberties, Justice and Home Affairs, Brussels, 2021: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/695445/IPOL_STU(2021)695445_EN.pdf.

[383] Centre for Media Pluralism and Media Freedom, Monitoring media pluralism in the digital era: application of the Media Pluralism Monitor in the European Union, Albania, Montenegro, the Republic of North Macedonia, Serbia and Turkey in the year 2021 - Media Pluralism Monitor, European University Institute, San Domenico di Fiesole, 2022, p. 151, available at: http://hdl.handle.net/1814/74712.

[384] Starr, P., *The Creation of the Media: Political Origins of Modern Communications*, Diane Publishing Company, 2006.; Benkler, Y., *The Wealth of Networks*, Yale University Press, New Haven, US. p. 17.

[385] Benkler, Y., *The Wealth of Networks*, Yale University Press, New Haven, US. p. 17, p. 179.

1920s to be run as a public service.[386] Still, its structure retained an important degree of operational freedom and a mandate 'to act in the public interest, serving all audiences through the provision of impartial, high-quality and distinctive output and services which inform, educate and entertain.'[387] France awarded ten-year franchises to about a dozen private stations, but controlled all broadcasts concerning political and economic questions.[388] In Germany, all broadcasting was state-owned until the 1980s, at which point it was opened up to market competition, but subject to interventionist regulation, including oversight by media councils in each federal state and by media councils representing civil society perspectives within each broadcaster.[389] By the late 1990s, the resulting mass media ecosystem was rather concentrated – with a rather low number of media actors holding a significant amount of market power - and largely supported by advertising, as well as a certain level of public funding, though with variations between countries and industries.[390]

A key characteristic of the public sphere enabled by commercial mass media is that information and communication flows mostly from one small number of people – professional journalists and the corporate entities behind main outlets - to a much larger audience. In this chapter these kind of outlets - radio, TV, and printed newspapers - are referred to as 'legacy news'. The only limit to this predominantly one-way dissemination is the cost of dissemination (print copies in print media, and in the case of radio and television the constraints on physical reach). In this model, newsrooms within individual publishing companies played a key editorial and curatorial role.[391] Consequently, a second key characteristic of the model was that audiences had fewer avenues to offer active feedback. Third, the kind of content that is published or broadcast reflects publishers' loosely-defined understanding of their target audience. The mass media model gives editors the power to determine content based on their interpretation of what the loosely-defined audience prefers.[392] For example, as newspapers in the 18th and 19th century grew their audiences, their content shifted from being party-oriented, based in community interests and practice, to being more fact- and sensation-oriented, with content that required less embeddedness in local contexts and achieved broader circulation.[393] A similar pattern can be observed today in places where local news outlets are bought and owned by giant news companies, where consolidation of ownership and cost-cutting has led to decreased coverage of local communities and events. [394] Box 10 below illustrates how media concentration regulations have yet to adapt to the challenges associated with digital media.

---

386 Starr, P., The Creation of the Media: Political Origins of Modern Communications, Diane Publishing Company, 2006., p. 341.

387 Benkler, Y., The Wealth of Networks, Yale University Press, New Haven, US. p. 17 p. 201; 'Mission, Values and Public Purposes', n.d., available at: https://www.bbc.com/aboutthebbc/governance/bbc.com/aboutthebbc/governance/mission/.

388 Starr, P., The Creation of the Media: Political Origins of Modern Communications, Diane Publishing Company, 2006., p. 342.

389 Humphreys, P., 'Germany's 'Dual' Broadcasting System: Recipe for Pluralism in the Age of Multi-Channel Broadcasting?', New German Critique, No. 78, 1999, p. 23.

390 Benkler, Y., The Wealth of Networks, Yale University Press, New Haven, US. 40.

391 Benkler, Y., The Wealth of Networks, Yale University Press, New Haven, US. 198.

392 Benkler, Y., The Wealth of Networks, Yale University Press, New Haven, US. p. 17, p. 211.

393 Benkler, Y., The Wealth of Networks, Yale University Press, New Haven, US. 199.

394 Minow, M., Saving the News: Why the Constitution Calls for Government Action to Preserve Freedom of Speech, Oxford University Press, 2021, p. 2.

Box 10: Media concentration regulation patterns in the EU

Many EU Member States have rules that aim to limit media concentration. Typically, these rules include ex ante authorisation of media market transactions, limitations on the allocated numbers of broadcasting licenses and newspapers, restrictions related to foreign ownership, and audience and market share ceilings. These rules, however, are mostly geared towards traditional media, such as newspapers, commercial radio, broadcasting, and linear audiovisual media. They have not yet recognised changes related to the concentration of resources in the value chain, particularly the trends shifting advertising funding towards online media platforms.

A detailed review of national and/or regional legislation that governs media pluralism found that limitations on media reach exist in 21 Member States out of 27, and restrictions on market shares and audiences' shares exist in 15 Member States. However, these restrictions rarely cover online media, and if they do they primarily refer to on-demand video services.

Sources: Ranaivoson, H. et. al., "Chapter B1. Mapping of the measures and data gathering methods concerning the concentration of economic resources to ensure media plurality", European Commission Study on Media Plurality and Diversity Online, Publications Office of the European Union, 2022, , available at: https://data.europa.eu/doi/10.2759/529019.

The internet presented, and to an important degree delivered, the possibility of a reversal in the trend towards concentration, as it decentralised and democratised the affordances required to produce and distribute information, culture, and knowledge.[395] Specifically, it enabled two fundamental changes. First, the high capital costs that were a prerequisite to gather, produce and distribute information plummeted. Anyone with a connected personal computer could now produce content. Second, relatedly, information no longer flowed from corporate or professional actors to the public. Rather, a basic output of the new communication ecosystem became direct communication between individuals - or users.[396]

Social media platforms emerged in the mid-2000s out of the 'exquisite chaos' that came with the freedom the internet delivered. Like search engines, they became ways to organise, but also connect and distribute, the old and new content that was now becoming available online.[397] Indeed, the key functional feature and innovation of social media platforms was to organise, through a unified interface, content published by users – including other media outlets - and to effectively match that content with interested audiences through data collection and algorithmic decision-making. Social media platforms thus maximised and helped organise the decentralised peer-to-peer production and dissemination of content and knowledge that the internet had promised. They have also become important for the functioning of modern democracies and the exercise of freedom of speech, which has put them at the centre of policy, academic and political discussions.

The sheer growth of social media platforms turned them into an important locus of power in media environments and contributed to transforming the business model of traditional news media. This happened in at least three ways. First, social media platforms became key intermediaries in content dissemination. Their moderation and curation decisions - often mediated by AI - are now important factors shaping the distribution of media content and the plurality of public debate. Second, the advent

[395] Ranaivoson, H. et. al., "Chapter B1. Mapping of the measures and data gathering methods concerning the concentration of economic resources to ensure media plurality", *European Commission Study on Media Plurality and Diversity Online,* Publications Office of the European Union, 2022, p. 236, available at: https://data.europa.eu/doi/10.2759/529019.

[396] Benkler, Y., *The Wealth of Networks*, Yale University Press, New Haven, US. p. 17, p. 32.

[397] Gillespie, T., Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media, Yale University Press, 2018, p.12.

of the online advertising market changed the business model of many media companies, including non-social media companies. Third, and relatedly, the concentration of market power in a few of these digital intermediaries - especially Meta, Google-owned YouTube and Twitter[398] - also creates risks to market pluralism.[399] Challenges associated with the opacity of information dissemination practices and the quality of the information on social media were addressed in Chapter 4 on disinformation. The following section describes in further detail the challenges associated with news business models and market concentration.

5.3. Social media and the news business model

Before the internet emerged, news creation and distribution had already been affected by new technologies, such as radio and TV. Yet the internet collapsed the barriers to publishing and enabled an unprecedented explosion of available information. This forced news organisations to compete in a more intense way for attention and advertising revenue, and posed a challenge for the business model of traditional media, especially written media.[400]The growth of intermediaries like search engines and social media raises two particular issues for the news media. First, they are an important way in which people find news. By 2014-2016, a growing number of organisations across the world reported that about half of their online traffic came directly from search and social referrals.[401] Second, as will be explained in more detail below, many news media outlets rely on them to provide advertising services.

As with dis- and misinformation, research on media pluralism highlights that the effects of social media cannot be separated from broader social, economic and political trends. Indeed, and as will be explained in further detail in this section, the trend towards market concentration in the news media ecosystem preceded the impact of social media, especially after the 2008-09 global financial crisis affected the profitability of news. The role of social media in distributing news content should not be overstated. The Eurobarometer's News and Media Survey of 2022 identified that European citizens trust traditional broadcast and print media more than online news platforms. 75% of respondents' most commonly used media channel was television (TV), followed by online news platforms, such as the websites of legacy written news publishers or online-only news outlets (43%), and radio (39%).[402] Only

---

398 Twitter is much smaller than Meta and Google in terms of total user numbers and revenue, but is the dominant platform used by journalists and other public figures. *See* Jurkowitz, M., and J. Gottfried, 'Twitter Is the Go-to Social Media Site for U.S. Journalists, but Not for the Public', *Pew Research Center*, 2022, available at: https://www.pewresearch.org/fact-tank/2022/06/27/twitter-is-the-go-to-social-media-site-for-u-s-journalists-but-not-for-the-public/.

399 Centre for Media Pluralism and Media Freedom, Monitoring Media Pluralism in the Digital Era: Application of the Media Pluralism Monitor in the European Union, Albania, Montenegro, the Republic of North Macedonia, Serbia & Turkey in the Year 2021l, European University Institutel, San Domenico di Fiesole, 2022, p. 4, available at: https://cadmus.eui.eu/bitstream/handle/1814/74712/MPM2022-EN-N.pdf?sequence=1&isAllowed=y.

400 Pickard, V., and Williams, A.T., 'Salvation Or Folly?', *Digital Journalism*, Vol. 2, No. 2, April 3, 2014, pp. 195–213., available at: https://www-tandfonline-com.acces-distant.sciencespo.fr/doi/full/10.1080/21670811.2013.865967; *see also* Antheaume, A. et. al., *The Changing Business of Journalism and its Implications for Democracy*, Reuters Institute for the Study of Journalism, Department of Politics and International Relations, University of Oxford, 2010., available at: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/research/files/The%2520Changing%2520Business%2520of%2520Journalism%2520and%2520its%2520Implications%2520for%2520Democracy.pdf.

401 Pickard, V., and Williams, A.T., 'Salvation Or Folly?', *Digital Journalism*, Vol. 2, No. 2, April 3, 2014, pp. 195–213., available at: https://www-tandfonline-com.acces-distant.sciencespo.fr/doi/full/10.1080/21670811.2013.865967; see also Antheaume, A. et. al., *The Changing Business of Journalism and its Implications for Democracy*, Reuters Institute for the Study of Journalism, Department of Politics and International Relations, University of Oxford, 2010., available at: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/research/files/The%2520Changing%2520Business%2520of%2520Journalism%2520and%2520its%2520Implications%2520for%2520Democracy.pdf.

402 Ipsos European Public Affairs. 'News and Media Survey 2022', European Parliament, Strasbourg, 2022, p.12, 17-18, available at: https://europa.eu/eurobarometer/surveys/detail/2832.

26% of respondents reported that they primarily got their news from social media platforms. Lagging behind came the printed press, as the most commonly used news source for only 21% of respondents.[403]

Nevertheless, it is worth noting that results vary depending on the age of the respondents. Younger individuals (15- to 24-year-olds) are more likely to use social media platforms to access news than individuals over 50 (46% compared to 15%).[404] Similarly, although 85% of citizens aged 55 or older access news using TV, only 58% of 15- to 24-year-olds do so.[405] The same holds true for news access via radio: 42% of citizens aged 55 or older use it to inform themselves, compared to only 26% of those aged 15 to 24. Lastly, it is pertinent to point out that paying for online news content is less predominant: 70% of those who access news online only use free news content or news services online.[406]

5.3.1. The profitability of legacy news

The tipping point of the transformation of news business models is often traced back to the 2008-09 financial crisis. The rise of the internet and social media was already underway, but the financial crisis provided a shock that forced many media companies to make, and sustain, internal organisational changes. The financial crisis immediately affected commercial news organisations around the world, with revenue (from sales, advertising and other sources of income) dropping in most countries - up to 30% in the US and 21% in the UK, and to a lesser degree in Europe: 10% in Germany, 7% in Finland and only 4% in France.[407] News organisations struggled to structure and finance their news production as they had done in more profitable times.[408] At the same time, the financial crisis coincided with a general trend towards increased consolidation and concentration of the media industry, and declining print circulation revenue.[409] This is despite the fact that in Europe the prevalence of public service media organisations has reduced the relative systemic dependency of European countries on private newspapers.[410] Another relevant factor that accelerated the integration of news businesses and social media platforms was the consolidation of the use of smartphones around 2015. This turned social

---

[403] For further analysis of which news sources European citizens prefer and why, *see* 'How and Why Do Consumers Access News on Social Media?', *Reuters Institute for the Study of Journalism*, n.d., available at: https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021/how-and-why-do-consumers-access-news-social-media.

[404] Ipsos European Public Affairs. News and Media Survey 2022, European Parliament, Strasbourg, 2022., p.12, 17-18, https://europa.eu/eurobarometer/surveys/detail/2832.

[405] Ipsos European Public Affairs. News and Media Survey 2022, European Parliament, Strasbourg, 2022., p.13, available at: https://europa.eu/eurobarometer/surveys/detail/2832.

[406] Ipsos European Public Affairs. News and Media Survey 2022, European Parliament, Strasbourg, 2022., p.18, available at: https://europa.eu/eurobarometer/surveys/detail/2832.

[407] Antheaume, A. et. al., *The Changing Business of Journalism and its Implications for Democracy*, Reuters Institute for the Study of Journalism, Department of Politics and International Relations, University of Oxford, 2010., available at: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/research/files/The%2520Changing%2520Business%2520of%2520Journalism%2520and%2520its%2520Implications%2520for%2520Democracy.pdf.

[408] Antheaume, A. et. al., *The Changing Business of Journalism and its Implications for Democracy*, Reuters Institute for the Study of Journalism, Department of Politics and International Relations, University of Oxford, 2010., available at: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/research/files/The%2520Changing%2520Business%2520of%2520Journalism%2520and%2520its%2520Implications%2520for%2520Democracy.pdf.

[409] Bell, E. and Owen, T., 'The Platform Press: How Silicon Valley Reengineered Journalism', *Columbia Journalism Review*, Tow Center for Digital Journalism at Columbia's Graduate School of Journalism, New York, 2017, available at: https://www.cjr.org/tow_center_reports/platform-press-how-silicon-valley-reengineered-journalism.php/.

[410] Antheaume, A. et. al., The Changing Business of Journalism and its Implications for Democracy, Reuters Institute for the Study of Journalism, Department of Politics and International Relations, University of Oxford, 2010., available at: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/research/files/The%2520Changing%2520Business%2520of%2520Journalism%2520and%2520its%2520Implications%2520for%2520Democracy.pdf.

media into one of the main ways in which readers access news.[411] The challenge then was to grow digital revenue 'far and fast enough to offset the inevitable declines in print revenue.'[412] Below, Box 11, presents the patterns of the consumer use of different social media news to access news.

Box 11: How much do news consumers use social media to access news?

Audiences are less likely than before to access news through the homepage of a news brand and increasingly more likely to do so via a search engine, a social network, email, or the lock screen of a smartphone. This is especially true for younger audiences.

It is important, however, to see this trend in its due proportion and understand that these numbers vary per social media platform. For example, according to 2021 data from the Reuters Institute, Facebook is more often named as a network where people come across news, but it is less often a platform where people intentionally go to access the news. Twitter, on the other hand, is more often a primary destination for news, but its user base is smaller. For example, although 21% of people in the UK use Twitter for news because it's 'a good place to access the latest news', the fact that Twitter has a smaller user base means that only about 3% of the population of the UK uses Twitter to access the news. However, Twitter is particularly widely used by journalists and other public figures, which means that trends, debates, information and sources on Twitter are particularly likely to influence news content in other media.

Meanwhile, YouTube, Instagram, Snapchat, and TikTok are valued more for entertainment and less often as a way of accessing news. However, 26% of respondents to the Reuters survey reported that on YouTube they were able to access perspectives not available on mainstream media.

Sources: Andi, S., How and why do consumers access news on social media?, Reuters Institute Digital News Report 2021 , available at: https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021/how-and-why-do-consumers-access-news-social-media; McGregor, S.C. and Molyneux, L., 'Twitter's influence on news judgment: An experiment among journalists', *Journalism, Vol. 21. No. 5*, 2018.

The transition into these digital formats caused huge disruptions, especially for the print news industry.[413] In the traditional media business model, newspapers typically gain their revenue from a combination of newsstand sales, subscription, and advertising revenues. They thus operate a two-sided marketplace where they provide content to consumers, normally at a subsidised price, and sell consumers' attention to advertisers by selling advertising space.[414] The transition to digital formats led newspapers to rely far less on newsstand sales and subscriptions, and more on sources of revenue like online advertising.[415] Some outlets have also experimented with donations and other voluntary

[411] Bell, E. and Owen, T., 'The Platform Press: How Silicon Valley Reengineered Journalism', *Columbia Journalism Review,* Tow Center for Digital Journalism at Columbia's Graduate School of Journalism, New York, 2017, available at: https://www.cjr.org/tow_center_reports/platform-press-how-silicon-valley-reengineered-journalism.php/.

[412] Thompson, M., 'The Challenging New Economics of Journalism.', *Reuters Institute Digital News Report 2016*, Reuters Institute for the Study of Journalism, Department of Politics and International Relations, University of Oxford, 2016, p. 108, available Reuters Institute Digital News Report 2016.

[413] Thompson, M., 'The Challenging New Economics of Journalism.', *Reuters Institute Digital News Report 2016*, Reuters Institute for the Study of Journalism, Department of Politics and International Relations, University of Oxford, 2016, p. 108, available at: Reuters Institute Digital News Report 2016.

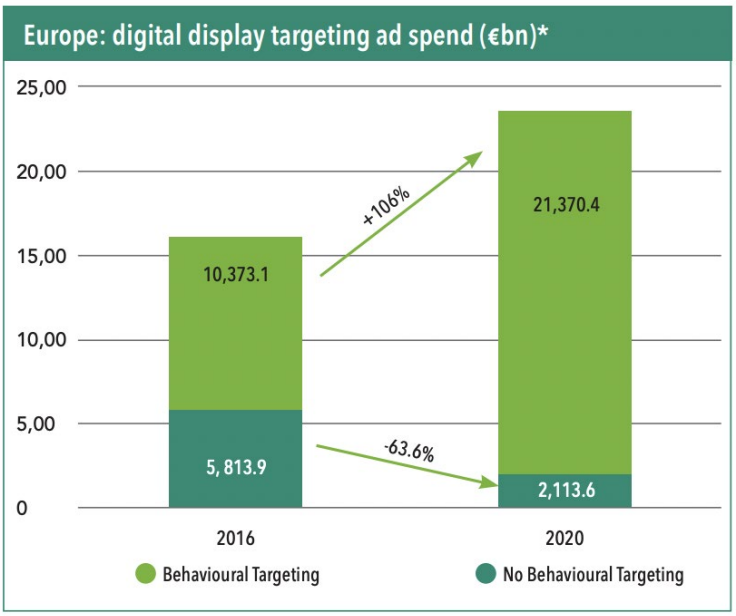
[414] Irion, K. et. al. 'Chapter B2. Overview of distribution of advertising revenues and access to and the intensity of use of consumer data' in Study on Media Plurality and Diversity Online, European Union, 2022, p. 250.

[415] Thompson, M., 'The Challenging New Economics of Journalism.', *Reuters Institute Digital News Report 2016*, Reuters Institute for the Study of Journalism, Department of Politics and International Relations, University of Oxford, 2016, p. 108, available at: Reuters Institute Digital News Report 2016.

contributions, philanthropy, selling other services, or linking to content produced elsewhere to attract traffic to their website.[416]

Importantly, reliance on online advertising has made media companies significantly dependent on technology companies, especially Meta and Google, both of which provide infrastructure to buy and sell online advertising and - until recently - held an effective duopoly over these intermediary services.[417] Websites, and news outlets, can 'sell' space on their websites to Google and/or Meta for them to show readers a targeted ad. In essence, targeted advertising involves leveraging data generated by consumers on these platforms, as well as data from other available sources (such as third-party apps, data brokers, etc.) to target individuals with ads and content that they are most likely to be interested in.[418] This type of data allows advertisements tailored to consumers, which makes it particularly effective for advertisers. Digital advertising generated €41.9 billion of annual revenue in Europe in 2016, and was then growing at a double-digit rate of 12.3%.[419] Figure 3 below shows the evolution of behavioural targeted advertising as a share of the broader advertising market between 2016 and 2020.

Figure 3: Behavioural Targeting Market Size



***Source:** IHS Markit Behavioural Impact Model. 2020 is forecast data assuming average annual growth rates of 10% between 2016 and 2020.*

Source: IHS Markit, 'The Economic Value of Behavioural Targeting in Digital Advertising', *Data Driven advertising* n.d., available at: https://datadrivenadvertising.eu/the-economic-value-of-behavioural-targeting-in-digital-advertising/.

[416] Antheaume, A. et. al., *The Changing Business of Journalism and its Implications for Democracy*, Reuters Institute for the Study of Journalism, Department of Politics and International Relations, University of Oxford, 2010., available at: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/research/files/The%2520Changing%2520Business%2520of%2520Journalism%2520and%2520its%2520Implications%2520for%2520Democracy.pdf.

[417] Open Markets Institute, 'America's Free Press and Monopoly: The Historical Role of Competition Policy in Protecting Independent Journalism in America', *Open Markets Institute*, n.d., Washington D.C., 2018, https://static1.squarespace.com/static/5e449c8c3ef68d752f3e70dc/t/5ea9fddc9d36302f95e273b8/1588198890990/Americas-Free-Press-and-Monopoly-PDF-1.pdf.

[418] Seufert, E., 'The App Tracking Transparency Recession', *Mobiledevmemo.com*, January 11, 2023. https://mobiledevmemo.com/the-att-recession/.

[419] *The Economic Value of Behavioural Targeting in Digital Advertising*, IHS Markit, London, 2017, p. 2, https://iabeurope.eu/wp-content/uploads/2019/08/BehaviouralTargeting_FINAL.pdf.

Besides integrating advertising into their online content by selling advertising space, media companies can also use these services to advertise their own content.[420] Media companies using these services, like anyone else using targeted advertising, are asked to create a targeted profile of the type of person they want to reach. Once the audience parameters are set, the content is targeted at the individuals who meet that criterion.[421] Because companies like Google and Facebook have an impressive amount of information about their users, and incredibly wide audiences they can offer ads that are more targeted, more precise, and which reach more people than those sold by other advertisers.[422]

Despite the opportunities digital advertising offers media companies, the rapid growth of the digital advertising market does not seem to have benefited publishers of original content much. Indeed, even where publishers' revenue from advertising has increased, it has not really compensated for the losses of the legacy model.[423] A recent study by digital rights consultancy AWO, commissioned by the EU Commission, provides some insights into the reasons for this. It shows that the market is complex, and up to 40-60% of ad spending goes into a complex network of intermediaries that is not transparent. This makes it hard for publishers to understand the efficiency of their ad spending, assess the performance of different ad models, potentially leading to inefficient spending but also strengthening the position of the players with stronger market power. [424]

The market for adtech services, which process user data and manage the automated placement and targeting of adverts on behalf of advertisers and publishers, exhibits huge economies of scale and strong network effects, which has led to significant market concentration.[425] This practice, as carried out by the largest tech companies, has also been widely criticised in the academic literature and by policymakers for its intrusiveness for privacy and autonomy.[426] For the purposes of this study, however, we focus on its effects for media pluralism. The complexity and opacity of supply chains for adtech services, and the dominance of platform companies like Meta, Google and Amazon at multiple stages of these supply chains, have negative implications for news organisations' pricing and bargaining power.[427] The oligopolistic structure of the advertising market, as well as its complexity and opacity, weaken the pricing and bargaining power of news organisations. The UK's Competition and Market

---

[420] Austin, S. and Newman, N., 'Attitudes to Advertising - Digital News Report 2015', *Reuters Institute Digital News Report*, Oxford, 2015., https://www.digitalnewsreport.org/essays/2015/attitudes-to-advertising/.

[421] Bell, E. and Owen, T., 'The Platform Press: How Silicon Valley Reengineered Journalism', *Columbia Journalism Review*, Tow Center for Digital Journalism at Columbia's Graduate School of Journalism, New York, 2017, available at: https://www.cjr.org/tow_center_reports/platform-press-how-silicon-valley-reengineered-journalism.php/.

[422] 'America's Free Press and Monopoly', *Open Markets Institute*, n.d., Washington D.C., 2018, p. 19, available at: https://static1.squarespace.com/static/5e449c8c3ef68d752f3e70dc/t/5ea9fddc9d36302f95e273b8/1588198890990/Americas-Free-Press-and-Monopoly-PDF-1.pdf.

[423] Irion, K. et.al., 'Introductory chapter. Outlining the value of safeguarding media pluralism and diversity to Member States, the EU and the relevant competences', Study on Media Plurality and Diversity Online (Centre for Media Pluralism and Media Freedom et. al.), 2022 p. 265.

[424] Armitage, C. et. al., Towards a more transparent, balanced and sustainable digital advertising ecosystem: Study on the impact of recent developments in digital advertising on privacy, publishers and advertisers, Publications Office of the European Union, January 31, 2023, available at: Study on the impact of recent developments in digital advertising on privacy, publishers and advertisers.

[425] Irion, K. et.al., 'Introductory chapter. Outlining the value of safeguarding media pluralism and diversity to Member States, the EU and the relevant competences', Study on Media Plurality and Diversity Online (Centre for Media Pluralism and Media Freedom et. al.), 2022 p. 274.

[426] See Zuboff, S., The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power, PublicAffairs, 2019; Richards, N., Why Privacy Matters, Oxford, Oxford University Press, 2021.

[427] Thompson, M., 'The Challenging New Economics of Journalism', Reuters Institute Digital News Report 2016, Reuters Institute for the Study of Journalism, Department of Politics and International Relations, University of Oxford, 2016, p. 108, available at: Reuters Institute Digital Report.

Authority has found that Google and Meta's market power effectively allows platforms to impose terms on publishers without needing to consult or negotiate with them.[428] Of the money spent by advertisers on digital advertising, only around 50% ultimately reaches publishers, with the rest going to various adtech intermediaries - often owned by Google or Meta.[429] Significant ad spending is also lost to ad fraud which simulates impressions and consumer behaviour,[430] and around 15% is simply untraceable.[431] The AWO study noted that over the last decade, the revenue of Europe's ten largest publishing businesses has remained basically flat, while Google and Meta's revenues increased by around 500%.[432] Representatives of publishers interviewed for the study described themselves as heavily dependent on these dominant intermediaries, in what felt like an 'abusive' or 'love/hate' relationship.[433]

At the same time as relying on Meta and Google for adtech services, news publishers are increasingly reliant on dominant social media platforms like Meta-owned Facebook and Twitter as a source of traffic - and thus ad revenue. Social media platforms' (generally algorithmically mediated) decisions about how to rank content for users, and how visible content and topics should be, influence the chances that news consumers will read a publisher's content and become aware of its brand.[434] This has two significant consequences.

First, platforms can change search algorithms in ways that affect website traffic, unexpectedly and with no explanations. This has direct financial consequences, as it represents lost website traffic and revenue for news organisations. Furthermore, unexplained and opaque algorithmic changes make planning and decision-making complicated, as understanding algorithms is important to optimise and prioritise content for visibility on social media.[435] For example, in the late 2010s, many news publishers laid off traditional journalists and invested significantly in video production and editing because Facebook had claimed that videos performed significantly better in terms of user engagement and traffic - only to later discover that Facebook's claims were inaccurate and based on inflated numbers.[436]

---

[428] Competition & Markets Authority, Online platforms and digital advertising: Market study final report, Competition & Markets Authority, London, 2019, p. 220, available at: Online platforms and digital advertising.

[429] ISBA, 'Programmatic Supply Chain Transparency Study', ISBA, May 6, 2020, available at: https://www.isba.org.uk/knowledge/executive-summary-programmatic-supply-chain-transparency-study.

[430] Armitage, C. et. al., Towards a more transparent, balanced and sustainable digital advertising ecosystem: Study on the impact of recent developments in digital advertising on privacy, publishers and advertisers, Publications Office of the European Union, January 31, 2023, available at: https://op.europa.eu/en/publication-detail/-/publication/8b950a43-a141-11ed-b508-01aa75ed71a1/language-en.

[431] ISBA, 'Programmatic Supply Chain Transparency Study', ISBA, May 6, 2020, available at: https://www.isba.org.uk/knowledge/executive-summary-programmatic-supply-chain-transparency-study.

[432] Armitage, C. et. al., Towards a more transparent, balanced and sustainable digital advertising ecosystem: Study on the impact of recent developments in digital advertising on privacy, publishers and advertisers, Publications Office of the European Union, January 31, 2023, available at: https://op.europa.eu/en/publication-detail/-/publication/8b950a43-a141-11ed-b508-01aa75ed71a1/language-en.

[433] Armitage, C. et. al., Towards a more transparent, balanced and sustainable digital advertising ecosystem: Study on the impact of recent developments in digital advertising on privacy, publishers and advertisers, Publications Office of the European Union, January 31, 2023, available at: https://op.europa.eu/en/publication-detail/-/publication/8b950a43-a141-11ed-b508-01aa75ed71a1/language-en .

[434] Competition & Markets Authority, *Online platforms and digital advertising: Market study final report,* Competition & Markets Authority, London, 2019, p. 220, available at: Online platforms and digital advertising: Market study final report.

[435] Competition & Markets Authority, *Online platforms and digital advertising: Market study final report,* Competition & Markets Authority, London, 2019, p. 220, available at: Online platforms and digital advertising: Market study final report.

[436] Owen, L.H., 'Facebook's pivot to video didn't just burn publishers. It didn't even work for Facebook', *Nieman Lab*, September 15, 2021, available at: https://www.niemanlab.org/2021/09/well-this-puts-a-nail-in-the-news-video-on-facebook-coffin/.

Second, dependence on dominant tech platforms prevents publishers from profiting directly from user data, as they instead rely on adtech intermediaries and analytics services for valuable insights into their audiences and the performance of their content. The data social media companies share with publishers is very aggregated and anonymised, partly because of privacy concerns. However, Meta and Google can profit from this ecosystem to develop their own services and advertising businesses, while publishers do not have access to the same level of data.[437] As Box 12 below illustrates, Meta and Google have ceased to hold most of the digital advertising market. How this will affect the landscape for publishers, however, is still unknown.

Box 12: The end of the digital advertising duopoly

In December 2022, the Financial Times reported that Meta and Alphabet (parent company of Google) had lost their joint majority of the digital advertising market, hit by fast-growing competition from rivals such as Amazon, TikTok, Microsoft and Apple. The article reported that the two companies' share of US ad revenue was projected to fall by 2.5 percentage points to 48.4% in 2022. This would be the first time the two groups will not hold a majority share of the market since 2014, and the fifth consecutive annual decline in their joint market share. Worldwide, Meta and Alphabet's share reportedly declined 1 percentage point to 49.5% in 2022.

These changes in the ad market are related to non-social media platforms such as Amazon and Apple leveraging their dominance in existing markets such as e-commerce and app stores to develop their own advertising businesses, as well as the growing popularity of TikTok. For example, Amazon has expanded its on-site ads business beyond its own site. Apple has launched efforts to 'redefine advertising' in a 'privacy-centric' way, making it harder for other companies to access data on its users for targeted advertising, while also expanding its own advertising business via its App Store.

Sources: McGee, P., 'Meta and Alphabet lose dominance over US digital ads market', *Financial Times*, December 23, 2022, available at: https://www.ft.com/content/4ff64604-a421-422c-9239-0ca8e5133042; Seufert, E.B., 'The Duopoly is over (because Everything is an Ad Network)', *Mobile Dev Memo*, December 21, 2022, available at: https://mobiledevmemo.com/the-duopoly-is-over-because-everything-is-an-ad-network/.

5.3.2. The dissemination and consumption of news

a. The transformation of the newsroom and the consumption of news

The possibility, and necessity, of reaching wider audiences online, and the reliance on online advertising, has changed the incentives for news productions for both editors and journalists alike. In Kleis Nielsen and Ganter's view, the relationship between social media companies and newsrooms is characterised by a tension between short-term operational pursuit of the opportunities offered by the possibility to reach more people online, and more long-term strategic worries about becoming too dependent on these new intermediaries. [438] This tension between long-term strategic and short-term operational needs has led to an acceleration of the news cycle; an apparent diminishment of investigative journalism and a proliferation of commentary and other forms of content that trigger strong emotional reactions; and a fragmentation of the public sphere.

[437] Armitage, C. et. al., Towards a more transparent, balanced and sustainable digital advertising ecosystem: Study on the impact of recent developments in digital advertising on privacy, publishers and advertisers, Publications Office of the European Union, January 31, 2023, available at: https://op.europa.eu/en/publication-detail/-/publication/8b950a43-a141-11ed-b508-01aa75ed71a1/language-en.

[438] Kleis Nielsen, R. and S.A. Ganter, 'Dealing with Digital Intermediaries: A Case Study of the Relations between Publishers and Platforms', New Media & Society, Vol. 20, No. 4, April 2018, pp. 1600–1617.

First, social media changed how and when news is produced and published. They have enabled 24/7 updates and commentary, so news content can rapidly become outdated, and have brought historically distinct and traditionally geographically separated media organisations into direct competition over the same potential users. [439] Newspapers now publish audiovisual content that resembles TV content as well as written articles, and the readership of newspapers like *Süddeutsche Zeitung* extends beyond southern Germany. This had already been identified by a 2010 OECD report which showed increased competition for attention and lack of resources led to sparser and lower-quality news coverage, as editors prioritised speed and interactivity over depth and quality. [440] More recent research echoes these early findings, suggesting that social media platforms' business model incentivises 'virality' - material people want to engage with and share - which does not necessarily correlate with journalistic quality. [441]

Reliance on social media to reach audiences, and the resulting pressures to produce viral content, have led to a diminishment of investigative journalism. Even well-established outlets rely more on outside news sources, news agencies and non-journalistic sources, without necessarily adding more original reporting. [442] Limited resources and the competition for audience's attention have favoured comment and opinion more than factual and investigative reporting, and 'softer' topics - such as lifestyle, celebrity content, etc. - which are cheaper to produce, appeal more to advertisers, and have more entertainment value than 'hard' news. These dynamics are not totally new: researchers have shown that in 'market-driven journalism' the success of a story as a product was already judged by the advertising revenues, and not necessarily its quality.[443] However, social media appear to have intensified this trend by making advertising revenue more directly dependent on the views and clicks that a story generates, and by creating a proliferation of new metrics for a story's success. Social media metrics - what attracts the most likes, comments and shares - influence which stories are written by journalists, which ones get promoted, and who succeeds at news workplaces. Relatedly, as they compete for attention on social media, individual journalists may be incentivised to show more extreme or partisan positions than the media organisations they work for, or to increase the degree of personalisation in the way they report news, which can increase politicisation and polarisation of the content they produce. [444]

Reliance on online advertising revenues appears to affect editorial decisions to some extent. Established outlets like the *New York Times* or *Le Monde* use social media to attract potential subscribers, using 'networked content' which leads back to the publisher's homepage. For some outlets like these, homepage traffic has been going up in recent years, after having seen a significant decrease around

---

[439] Kleis Nielsen, R. and S.A. Ganter, 'Dealing with Digital Intermediaries: A Case Study of the Relations between Publishers and Platforms', *New Media & Society*, Vol. 20, No. 4, April 2018, p. 1607.

[440] Wunsch-Vicent, S., *The Evolution of News and the Internet*, Organisation for Economic Co-operation and Development, France, 2010, p. 60.

[441] Wunsch-Vicent, S., *The Evolution of News and the Internet*, Organisation for Economic Co-operation and Development, France, 2010, p. 60.

Nielsen et. al. 2017, Bell, E. and Owen, T., 'The Platform Press: How Silicon Valley Reengineered Journalism', Columbia Journalism Review, Tow Center for Digital Journalism at Columbia's Graduate School of Journalism, New York, 2017, https://www.cjr.org/tow_center_reports/platform-press-how-silicon-valley-reengineered-journalism.php/.

[442] Wunsch-Vicent, S., *The Evolution of News and the Internet*, Organisation for Economic Co-operation and Development, France, 2010, p. 60.

[443] Wunsch-Vicent, S., *The Evolution of News and the Internet*, Organisation for Economic Co-operation and Development, France, 2010, p. 61; see also Boczkowski, P. *Digitising the News: Innovation in Online Newspapers*, The MIT Press. Cambridge, U.S. 2005.

[444] Barberá, P. et al., 'Social Media, Personalisation of News Reporting, and Media Systems' Polarisation in Europe', *Social Media and European Politics*, Palgrave Macmillan, United Kingdom, 2017, pp. 25-52.

2015. On the other hand, outlets that rely exclusively on advertising revenue often invest significantly in native content, meaning online networked content that is not intended to lead to a different website but rather to be read within the social media platform. Native posts can often be monetised using tools made available by social media companies, such as the sharing of further advertising revenues, but this means that control of audience data remains with the platform. [445] Whereas 98% of *Huffington Post's* total Facebook posts are native, just 16% of *New York Times* content is designed to be native while 84% is designed to drive audiences back to nytimes.com. [446] The *Wall Street Journal*, which relies on subscriptions, publishes only 3% native posts. [447]

As well as favouring certain content formats over others, dependence on advertising revenues could also influence which topics news outlets cover, since most advertisers now use 'brand safety' tools to avoid purchasing ad space on webpages with content which could reflect negatively on their brands. Concerningly, research has shown that in practice this often includes large swathes of content relating to politics, current affairs and LGBTQ+ themes[448] - meaning news publishers are disincentivised from covering such topics, as they will not bring in advertising revenue. How much this influences editorial decisions is a question that demands more research.

Box 13 below highlights emerging trends on news publishers use of TikTok to attract new audiences.

Box 13: News on TikTok

Recently, news publishers have slowly turned to TikTok to attract new audiences. A report published by the Reuters Institute for the Study of Journalism published in 2022 found that 49% of news publishers in 44 countries are regularly publishing content on TikTok, a large proportion of whom joined the platform in the last year. News on TikTok is still mostly generated by social media influencers, activists and ordinary people, but news organisations have been especially attracted by the fast-growing audience, the younger demographic and a desire to provide accurate information. This trend has gained strength in large European countries like France, Spain and the UK. News publishers such as *Le Monde* (FR), *BBC News* (UK), *El Mundo* (ES) and ARD *Tagesschau* (DE) have significant numbers of followers in the app.

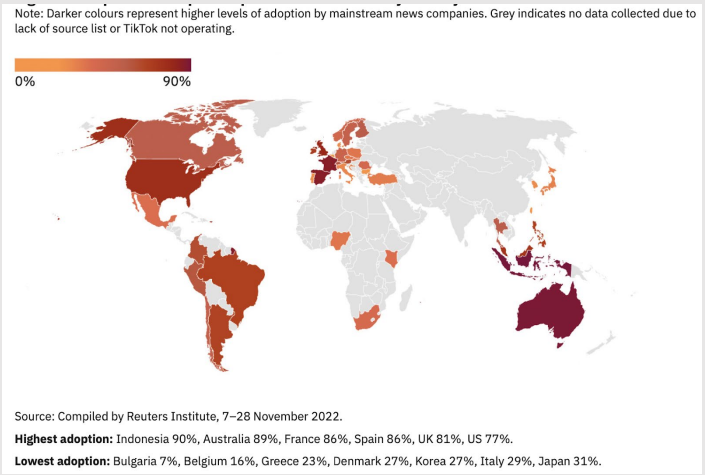
[445] Competition & Markets Authority, *Online platforms and digital advertising: Market study final report,* Competition & Markets Authority, London, 2019, Online platforms and digital advertising .

[446] Bell, E. and Owen, T., 'The Platform Press: How Silicon Valley Reengineered Journalism', *Columbia Journalism Review,* Tow Center for Digital Journalism at Columbia's Graduate School of Journalism, New York, 2017, https://www.cjr.org/tow_center_reports/platform-press-how-silicon-valley-reengineered-journalism.php/.

[447] Bell, E. and Owen, T., 'The Platform Press: How Silicon Valley Reengineered Journalism', *Columbia Journalism Review,* Tow Center for Digital Journalism at Columbia's Graduate School of Journalism, New York, 2017, https://www.cjr.org/tow_center_reports/platform-press-how-silicon-valley-reengineered-journalism.php/.

[448] CHEQ, 'Brand Safety's Technological Challenge: How Keyword Blacklists Are Killing Reach and Monetization. A Study by CHEQ's Department of Data Science' CHEQ, 2019; Parker, B., 'How advertisers defund crisis journalism', *The New Humanitarian*, January 27, 2021, https://www.thenewhumanitarian.org/analysis/2021/01/27/brand-safety-ad-tech-crisis-news.

Figure 4: Proportion of top news publishers on TikTok by country



Source: Newman, N., How Publishers are Learning to Create and Distribute News on TikTok, Reuters Institute for the Study of Journalism, University of Oxford, 2022, p. 28, available at: How Publishers are Learning to Create and Distribute News on TikTok.

Publishers seem to see it as a good opportunity to build a relationship with younger audiences, who tend to not go straight to news outlets, and develop more brand loyalty among these valuable audiences. Using features such as live-streaming, content creators are able to interact with their audiences, providing more personal or whimsical content that is more appealing for these audiences.

Nevertheless, there are also some downsides. TikTok centres around short-form videos (under 60 seconds) and a 'viral' model of content distribution where users see less content from accounts they follow, and more unknown content which is algorithmically recommended if it is attracting engagement from audiences. Thus, news publishers aiming to reach a wide audience on TikTok are incentivised to present content in simplified and sensationalist ways in order to quickly grab viewers' attention. TikTok's approach to content moderation could also disfavour content covering certain serious news topics, encouraging news publishers to focus on lighter topics such as lifestyle and celebrities instead. TikTok has also been found to censor content containing tags such as 'gay' or 'queer'. While the platform claims to have taken some steps to address this, concerns about opaque and arbitrary content moderation. The app also adds black screens over videos as a warning for violence, which can limit the diffusion of this content - potentially affecting coverage of important news topics. Another drawback is that there is still not a monetisation model to compensate for the value and content news publishers provide to the platform.

Sources: Newman, N., How Publishers are Learning to Create and Distribute News on TikTok, Reuters Institute for the Study of Journalism, University of Oxford, 2022, p. 8, available at: How Publishers are Learning to Create and Distribute News on TikTok; Nilsen, J. et. al., 'TikTok, the War on Ukraine, and 10 Features That Make the App Vulnerable to Misinformation', *Media Manipulation Casebook*, March 10, 2022., available at: https://mediamanipulation.org/research/tiktok-war-ukraine-and-10-features-make-app-vulnerable-misinformation.

b. Lack of transparency and changing trends

The landscape described above is not completely settled. Publishers often voice concern about the opacity of digital advertising, and transparency in the ad industry remains a key challenge for news outlets. Online advertisers retain all the information about readers, and the mechanisms through which publishers receive their remuneration is opaque.[449] Data access and clarity are a recurring concern for publishers.[450] This hinders news outlets from assessing the long-term value of reliance on social media, but also from obtaining valuable knowledge about their readers.[451] Certain disappointment with the reliability of ads metrics has led publishers to seek other revenue sources and strategies.

Consequently, many major news organisations have tried to strengthen their membership and subscription models of revenue. Recent developments such as the Covid-19 pandemic may have made some consumers more willing to pay for news, creating a window of opportunity - at least for outlets who can leverage their brand recognition to attract financial support. Disruptive events may be increasing the demand for high-quality news. In the early days of the Covid-19 pandemic, many publications experienced increases in digital subscriptions. In 2020, the Reuters Institute for the Study of Journalism reported an increase in the proportion of revenue coming from paid online news relative to the year before: the percentage of news consumers who paid for news increased by 19% in Finland, 17% in Denmark, 10% in Germany and France, and 8% in Norway.[452] The Institute's analysis emphasised that this phenomenon represented the value of trusted journalism during crises, but it also raised dilemmas around paywalls and access to information, as the journalism community felt that, especially at a moment of crisis, certain content (such as information about Covid-19) should be free.[453] Indeed, several news outlets, including *El Pais* and the *Financial Times,* partially or totally dropped their paywalls for a period of time during the pandemic.[454]

However, the pressures on media outlets to increase their reliance on subscriptions could undermine media pluralism, as larger national news outlets are better placed to safeguard their audience relationships, maintain brand awareness, and attract enough subscriptions to serve as a sustainable source of revenue, while this is more difficult for smaller, niche and local media publishers.[455] The

---

[449] Bell, E. and Owen, T., 'The Platform Press: How Silicon Valley Reengineered Journalism', Columbia Journalism Review, Tow Center for Digital Journalism at Columbia's Graduate School of Journalism, New York, 2017, available at: https://www.cjr.org/tow_center_reports/platform-press-how-silicon-valley-reengineered-journalism.php/, (UK CMA, 2019, 222).

[450] Bell, E. and Owen, T., 'The Platform Press: How Silicon Valley Reengineered Journalism', Columbia Journalism Review, Tow Center for Digital Journalism at Columbia's Graduate School of Journalism, New York, 2017, available at: https://www.cjr.org/tow_center_reports/platform-press-how-silicon-valley-reengineered-journalism.php/ (UK CMA, 2019, 222).

[451] Bell, E. and Owen, T., 'The Platform Press: How Silicon Valley Reengineered Journalism', *Columbia Journalism Review,* Tow Center for Digital Journalism at Columbia's Graduate School of Journalism, New York, 2017, available at: https://www.cjr.org/tow_center_reports/platform-press-how-silicon-valley-reengineered-journalism.php/.

[452] Newman, N. et. al., *Reuters Institute Digital News Report 2020*, Reuters Institute for the Study of Journalism, University of Oxford, Oxford, 2020, p. 22, available at: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf.

[453] Newman, N. et. al., *Reuters Institute Digital News Report 2020*, Reuters Institute for the Study of Journalism, University of Oxford, Oxford, 2020, p. 22, available at: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf.

[454] Newman, N. et. al., *Reuters Institute Digital News Report 2020*, Reuters Institute for the Study of Journalism, University of Oxford, Oxford, 2020, p. 22, available at: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf.

[455] Bell, E. and Owen, T., 'The Platform Press: How Silicon Valley Reengineered Journalism', *Columbia Journalism Review,* Tow Center for Digital Journalism at Columbia's Graduate School of Journalism, New York, 2017, available at: https://www.cjr.org/tow_center_reports/platform-press-how-silicon-valley-reengineered-journalism.php/.

Reuters Institute report also highlights the emergence of a relatively new model of relying on donations for news.[456] Here too, however, it is also the main national brands that seem to be benefiting most: in the UK in 2019, almost half of all donations (42%) went to the *Guardian*.[457] Major brands also have greater capacities to invest in technology and expertise to make their own advertising products. For example, the *New York Times* launched a direct ad sales business that uses its own data in 2020.[458]

In addition, despite these more positive developments, it remains the case that about 70% of news consumers are still not paying for online news, and that many of those who paid get their subscription through their employers or educational institutions.[459] The most important factor for those paying for news was the distinctiveness and quality of the content.[460]

c. The fragmentation of the public sphere

Many scholars have argued that the internet and social media contribute to the fragmentation of the public sphere by creating echo-chambers - or networks where individuals only or predominantly access content and opinions already similar to theirs. The public sphere is often defined as a space in which citizens are provided with information, ideas and debates around public affairs, so that they can acquire an informed opinion and participate in democratic politics.[461] Academic discussion on this topic is usually anchored in philosopher Jürgen Habermas' influential account of the public sphere as a crucial part of social life, in which citizens express public opinion through rational discourse, and which facilitates debate over common issues. [462] In the 1960s, Habermas famously argued that mass media failed to provide such a space due to its commercialisation and the influence of public relations. More recently, revisiting his earlier work, he has expressed concern that the rise of social media has exacerbated the commercialisation and superficiality of the media, compromising their ability to intermediate and form public opinion. [463]

A recognised framework developed by Dahlgren classifies the public sphere in three dimensions: structural, representational and interactional. [464] The structural dimension refers to the way the communicative space is organised - issues of access, freedom of speech and its dynamics, inclusivity and exclusivity, etc. Applied to social media, it can refer to how these spaces are configured in their

---

[456] Bell, E. and Owen, T., 'The Platform Press: How Silicon Valley Reengineered Journalism', *Columbia Journalism Review,* Tow Center for Digital Journalism at Columbia's Graduate School of Journalism, New York, 2017, available at: https://www.cjr.org/tow_center_reports/platform-press-how-silicon-valley-reengineered-journalism.php/.

[457] Bell, E. and Owen, T., 'The Platform Press: How Silicon Valley Reengineered Journalism', *Columbia Journalism Review,* Tow Center for Digital Journalism at Columbia's Graduate School of Journalism, New York, 2017, available at: https://www.cjr.org/tow_center_reports/platform-press-how-silicon-valley-reengineered-journalism.php/.

[458] Team, T.N.O., 'To Serve Better Ads, We Built Our Own Data Program', *Medium*, December 17, 2020, available at: https://open.nytimes.com/to-serve-better-ads-we-built-our-own-data-program-c5e039bf247b.

[459] Eurobarometer, *Media & News Survey 2022,* n.d. July 2022, p. 30, available at: https://europa.eu/eurobarometer/surveys/detail/2832.

[460] Newman, N. et. al., *Reuters Institute Digital News Report 2020*, Reuters Institute for the Study of Journalism, University of Oxford, Oxford, 2020, p. 11, available at: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf.

[461] Dahlgren, P., Media and political engagement. Citizens, communication and democracy. Cambridge University Press, Cambridge, 2009.

[462] Habermas, J. The structural transformation of the public sphere: An inquiry into a category of bourgeois society. Cambridge, U.K.,: Polity, 1989.

[463] Habermas, J., 'Reflections and Hypotheses on a Further Structural Transformation of the Political Public Sphere', *Theory, Culture and Society*, Vol. 39, No. 4. 2023.

[464] Batorski, D., and Grzywińska, I., 'Three Dimensions of the Public Sphere on Facebook', *Information, Communication & Society*, Vol. 21, No. 3, March 4, 2018, pp. 356–374.

legal, social and economic and technical features. [465] For example, the popularity of social media and the decreasing cost of internet access have brought larger populations into online political debate. .[466]

The representational dimension refers to different media outputs, such as agenda-setting, pluralism of views, and accuracy of coverage. [467] The academic literature has shown that online communities can serve as discussion forums providing alternatives to traditional media discourse, for example when dissenting opinions are excluded from mainstream politics in authoritarian countries. [468] At the same time, scholars have shown that the alternative communities that social media platforms enable can be highly partisan and can be used to promote disinformation or extremism. [469]

The interactional dimension refers to the realisation of the promise of the public sphere: an exchange of views and opinions amongst participants. [470] It can be divided into interactions between citizens and the media and between citizens themselves. [471] Many policy makers and academics have focused on this aspect in attempting to understand whether social media are pushing users into filter bubbles, as is often suggested by mainstream media and some researchers. As Chapter 4 on disinformation explained in more detail, empirical evidence on this issue is mixed: such echo chambers may exist in some contexts, but reflect user behaviour and choices as well as the influence of platforms. [472]

In the social media context, the possibility of increased polarisation has been linked to limited attention, the increasing number of information sources, and the so-called rise of an 'attention economy': the idea that given the abundance of information on the internet, attention is now a scarce and valuable resource, so content must be designed to attract and keep users' attention, which can often be achieved by presenting it in exaggerated, sensationalist or emotive ways which play on people's identities and conflicts. [473] Additionally, social media architecture typically encourages users to identify and connect with people and media publishers with similar opinions and characteristics, and to consume content which is optimised to meet users' individual preferences. [474]

---

[465] Batorski, D., and Grzywińska, I., 'Three Dimensions of the Public Sphere on Facebook', *Information, Communication & Society*, Vol. 21, No. 3, March 4, 2018, pp. 356–374.

[466] Kushin, M.J., and Kitchener, K., 'Getting Political on Social Network Sites: Exploring Online Political Discourse on Facebook', *First Monday*, 2009.

[467] Dahlgren, P., Media and political engagement. Citizens, communication and democracy. Cambridge University Press, Cambridge, 2009.

[468] Etling, B. et. al., Blogs as an alternative public sphere: The role of blogs, mainstream media, and TV in Russia's media ecology (Berkman Center Research Publication No. 8). Berkman Center Research Publication, 2014, Retrieved from http://www.ssrn.com/abstract=2427932.

[469] Iosifidis, P., 'The public sphere, social networks and public service media,' *Information, Communication & Society*, *Vol. 14* Issue 5, 2011, p. 619–637; Kaiser, J., and Rauchfleisch, A., 'Unite the Right? How YouTube Recommendation Algorithm Connects The U.S. Far-Right,' *D&S Media Manipulation: Dispatches from the Field*, 11 April, 2018 https://www.hiig.de/en/how-youtube-helps-to-unite-the-right/.

[470] Dewey, J., The public and its problems. New York, NY: Swallow Press, 1954.; Habermas, J., Between facts and norms: Contributions to a discourse theory of law and democracy. Cambridge: Polity Press, 1996.

[471] Dahlgren, P., Media and political engagement. Citizens, communication and democracy. Cambridge: Cambridge University Press, 2009.

[472] *See* Chapter 3, above; Brown, M. et. al. "Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users." *SSRN Electronic Journal*, 2022, https://ssrn.com/abstract=4088828.

[473] Davenport, T. and Beck, J., *The Attention Economy: Understanding the New Currency of Business*. Cambridge: MA: Harvard Business School Press, 2001.

[474] Bakshy, E. et. al., 'Exposure to ideologically diverse news and opinion on Facebook', *Science Vol. 348, Issue 6239,* 2015, p. 1130–1132. https://doi.org/10.1126/science.aaa1160.

Several studies on Twitter have shown that political discussions are grouped into clusters of like-minded users.[475] Similarly, a 2018 study showed that YouTube's recommendation algorithm contributes to the formation of far-right communities in both the US and Germany.[476] Brown et. al. distinguish ideological echo chambers, extremist rabbit holes and platform-wide ideological bias, and use these different concepts to analyse a large survey of US-based YouTube users. An echo chamber is a distribution of content that is tightly centred around an individual's particular ideological position. A rabbit hole exists when that individual is pushed towards increasingly extreme content after showing interest in a given topic. Finally, 'platform-wide ideological biases occur when, at a system level, users are pushed towards videos that are systematically in one ideological direction.'[477] They find only limited evidence that YouTube's recommendation algorithm pushes users into ideological echo chambers or extremist rabbit holes, but find stronger evidence that there is a platform-wide bias toward more conservative content. The bias, according to their research, is toward a moderately conservative space, not to ideological extremes. Notably, a peer-reviewed study conducted by Twitter's internal researchers found similar results showing a system-level bias towards right-wing political content.[478]

However, understanding how social media platforms influence these dynamics is challenging, due in large part to limitations in access to data, which lead to conflicting results and an incomplete understanding of these issues. It is not possible to generalise from the limited studies that are available to all platforms and contexts.[479] For example, a recent study comparing Google Search, Google News, Facebook, YouTube and Twitter found little evidence for 'filter bubbles' based on users' ideologies on any platform, and suggested that recommendations actually tend to have a homogenising effect, favouring the biggest news brands regardless of the user's characteristics, However, the authors ultimately concluded that each platform has its own dynamics as to which kinds of content it favours.[480] The possible existence and nature of echo chambers and system-level ideological biases on other platforms, possible variation between different political contexts, and their implications for media pluralism are important questions that demand more independent research. In this regard, the possibility for vetted independent researchers to request access to in-depth internal data from platforms under Article 40(4) DSA is an important step. European policymakers should ensure the

---

[475] Yardi, S. and Boyd, D., 'Dynamic debates: An analysis of group polarization over time on twitter.' *Bulletin of Science, Technology & Society*, *30(5)*, 2010, p 316–327. http://doi-org.acces-distant.sciencespo.fr/10.1177/0270467610380011.

[476] Kaiser, J., and Rauchfleisch, A., 'Unite the Right? How YouTube Recommendation Algorithm Connects The U.S. Far-Right,' *D&S Media Manipulation: Dispatches from the Field*, 11 April, 2018 https://www.hiig.de/en/how-youtube-helps-to-unite-the-right/.

[477] Brown M. et. al. *Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users,* May 11, 2022, p. 5. Available at: https://www.brookings.edu/research/echo-chambers-rabbit-holes-and-ideological-bias-how-youtube-recommends-content-to-real-users/.

[478] Huszár F. et. al. 'Algorithmic amplification of politics on Twitter,' *Proceedings of the National Academy of Science Vol. 119, No. 1, 2021*, January 5, 2022, https://www.pnas.org/doi/10.1073/pnas.2025334119?cookieSet=1%20.

[479] It is worth noting that academic research has disproportionately focused on Twitter because the platform makes more data available to researchers than other platforms, and not because other platforms are considered less significant (indeed, given their larger and more diverse user bases, rather the opposite). See Tufeczi, Z., 'Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodologies' ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media, 2014, https://arxiv.org/ftp/arxiv/papers/1403/1403.7400.pdf Matamoros-Fernández, A. and Farkas, J., 'Racism, Hate Speech, and Social Media: A Systematic Review and Critique,' Television & New Media Vol. 22, Issue 2 https://doi.org/10.1177/1527476420982230.

[480] Nechushtai, E., Zamith, R. and Lewis, S.C., 'More of the Same? Homogenization in News Recommendations When Users Search on Google, YouTube, Facebook, and Twitter', *Mass Communication and Society*, 2023.

process for such data access is as streamlined as possible and offer financial and practical support for research into how social media shapes media content and consumption.

5.4. Market concentration and the challenge to local news

Market concentration in the media ecosystem has put increasing pressure on local and regional news media.[481] Market concentration is the opposite of media pluralism: it refers to the dominance of a few large actors in the media industry. In the EU, the EUI Media Pluralism Monitor gave market plurality an average risk score of 66% in 2022. This indicator is designed to assess the risks to media pluralism that arise from the legal and economic context in which media actors operate. It deals with the structure of the market concentration but also other factors such as ownership and transparency and economic sustainability. Market concentration has always been a feature of the European media market, but it has tended to increase as legacy media organisations merge and consolidate to face digital disruption.[482] As has been described throughout this chapter, market concentration in the online news ecosystem is characterised by two main features. The first is the high concentration of the ad market and dominance of large technology companies. Second, as this section outlines, many mainstream media outlets have merged and consolidated to face digital disruption, and the market exhibits a winner-takes-most tendency.[483] Although the digital ecosystem has opened up opportunities for new media outlets to be established,[484] it has not been enough to counteract the trends towards concentration.

This winner-takes-most tendency has put increasing pressure on local and regional news media.[485] It is often a few of the largest news brands that are best positioned to benefit from online payments and donations as a substitute for advertising revenue. The accessibility of digital news content has also meant that local news publishers must compete for audience attention against more global or general online news outlets.[486] At the same time, they face competition from specialised websites and apps providing services like weather forecasts or job boards that were previously more reliant on local newspapers, as well as from social media feeds and groups that enable communities to form around specific subjects and areas. Similarly, local authorities, businesses and politicians now use websites or

---

[481] Centre for Media Pluralism and Media Freedom, Monitoring media pluralism in the digital era: application of the Media Pluralism Monitor in the European Union, Albania, Montenegro, the Republic of North Macedonia, Serbia and Turkey in the year 2021 - Media Pluralism Monitor, European University Institute, San Domenico di Fiesole, 2022, p. 151, available at: http://hdl.handle.net/1814/74712.

[482] Centre for Media Pluralism and Media Freedom, Monitoring media pluralism in the digital era: application of the Media Pluralism Monitor in the European Union, Albania, Montenegro, the Republic of North Macedonia, Serbia and Turkey in the year 2021 - Media Pluralism Monitor, European University Institute, San Domenico di Fiesole, 2022, p. 156, available at: http://hdl.handle.net/1814/74712.

[483] Centre for Media Pluralism and Media Freedom, Monitoring media pluralism in the digital era: application of the Media Pluralism Monitor in the European Union, Albania, Montenegro, the Republic of North Macedonia, Serbia and Turkey in the year 2021 - Media Pluralism Monitor, European University Institute, San Domenico di Fiesole, 2022, p. 152, available at: http://hdl.handle.net/1814/74712.

[484] Centre for Media Pluralism and Media Freedom, Monitoring media pluralism in the digital era: application of the Media Pluralism Monitor in the European Union, Albania, Montenegro, the Republic of North Macedonia, Serbia and Turkey in the year 2021 - Media Pluralism Monitor, European University Institute, San Domenico di Fiesole, 2022, p. 152, available at: http://hdl.handle.net/1814/74712.

[485] Centre for Media Pluralism and Media Freedom, Monitoring media pluralism in the digital era: application of the Media Pluralism Monitor in the European Union, Albania, Montenegro, the Republic of North Macedonia, Serbia and Turkey in the year 2021 - Media Pluralism Monitor, European University Institute, San Domenico di Fiesole, 2022, p. 151, available at: http://hdl.handle.net/1814/74712.

[486] Schulz, A., 'Local news unbundled: where audience value still lies,' Reuters Institute Digital News Report 2021, 10th Edition, Oxford, Reuters Institute for the Study of Journalism, June 23, 2021, available at: Reuters Institute Digital News Report 2021.

social media pages, instead of local newspapers, to provide updates and communicate with constituents.[487] Local newspapers across the world are working to develop new products to differentiate themselves, such as series featuring stories about local people, editorial newsletters, and a greater focus on sports. However, the Reuters Institute highlights that it is unclear how far those efforts can solve the problems that the unbundling of local information creates for the business model of local news outlets, given the competition from platforms and other digital alternatives.[488]

In a 2021 survey, 50-60% of respondents across 38 markets still considered traditional local media - including newspapers, TV and local radio - most valuable for covering hard news topics such as local politics, crime and the economy, as well as softer subjects like local sport. Newspapers were also valued as publishers of formal announcements, such as deaths and births.[489] However, social media are most valued for information about shops and restaurants (49%), local services (47%), and things to do in the area (46%).[490] The weakening of local news poses particularly acute risks to democracy. Local and regional news media play a critical role in informing citizens about democratic processes, holding local politicians to account, and fostering democratic participation and community building.[491]

Figure 5 presents data from a across-markets survey conducted by YouGov and commissioned by the Reuters Institute for the Study of Journalism. It shows how respondents from value different sources differently to access news and information on topics that had been traditionally covered by local news outlets.[492]

[487] Schulz, A., 'Local news unbundled: where audience value still lies,' *Reuters Institute Digital News Report 2021, 10th Edition,* Oxford, Reuters Institute for the Study of Journalism, June 23, 2021, available at: Reuters Institute Digital News Report 2021.

[488] Schulz, A., 'Local news unbundled: where audience value still lies,' *Reuters Institute Digital News Report 2021, 10th Edition,* Oxford, Reuters Institute for the Study of Journalism, June 23, 2021, available at: Reuters Institute Digital News Report 2021.

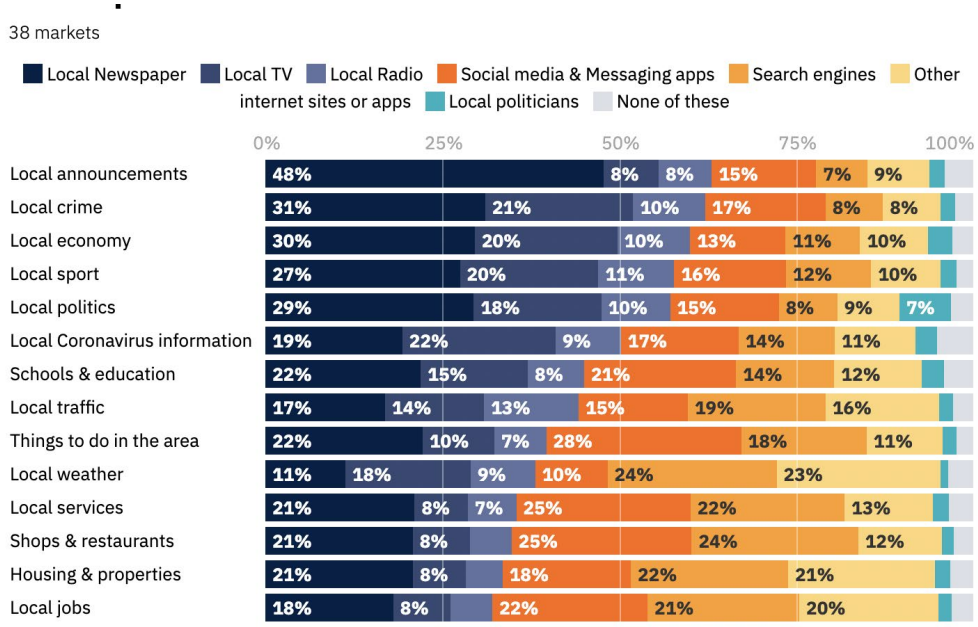
[489] Schulz, A., 'Local news unbundled: where audience value still lies,' *Reuters Institute Digital News Report 2021, 10th Edition,* Oxford, Reuters Institute for the Study of Journalism, June 23, 2021, available at: Reuters Institute Digital News Report 2021.

[490] Schulz, A., 'Local news unbundled: where audience value still lies,' *Reuters Institute Digital News Report 2021, 10th Edition,* Oxford, Reuters Institute for the Study of Journalism, June 23, 2021, available at: Reuters Institute Digital News Report 2021.

[491] Schulz, A., 'Local news unbundled: where audience value still lies,' *Reuters Institute Digital News Report 2021, 10th Edition,* Oxford, Reuters Institute for the Study of Journalism, June 23, 2021, available at: Reuters Institute Digital News Report 2021.

[492] Schulz, A., 'Local news unbundled: where audience value still lies,' *Reuters Institute Digital News Report 2021, 10th Edition,* Oxford, Reuters Institute for the Study of Journalism, June 23, 2021, Reuters Institute Digital News Report 2021.

Figure 5: What source of news are considered best for different local content



Source: Schulz, A., 'Local news unbundled: where audience value still lies,' *Reuters Institute Digital News Report 2021*, Reuters Institute Digital News Report 2021

5.5. Legal framework and regulatory developments

Media pluralism is established as a pillar of democracy in the EU, along with freedom of expression, which includes the right to receive and impart information. These rights are protected in Article 11 of the EU Charter of Fundamental Rights, which mirrors Article 10 of the European Convention for the Protection of Human Rights and Fundamental Freedom. Media freedom and pluralism has additionally become a priority for the European Commission's work and reporting on the rule of law. In 2020, the Commission published the first Rule of Law Report, which included a section on media freedom and pluralism. In 2021 the Commission proposed a European Media Freedom Act (EMFA) and in 2022 it adopted a European Democracy Action Plan aiming at improving the safety of journalists. The revised Audiovisual Media Services Directive, also published in 2022, is expected to strengthen independence of media regulators, transparency of media ownership and media literacy.[493] Finally, the new Copyright Directive contains rules that aim to increase press publishers' remuneration for the use of their content by online platforms, like social media.[494] This section focuses on the regulation of social media in the European Media Freedom Act and the Copyright Directive, as the DSA has already been addressed in the preceding chapters.

5.5.1. The European Media Freedom Act

The European Media Freedom Act was initially proposed in 2021, and seeks to address problems affecting the functioning of the internal market for media services and the operation of media service

[493] European Commission, 'Media freedom and pluralism', *Shaping Europe's digital future,* n.d. available at: https://digital-strategy.ec.europa.eu/en/policies/media-freedom.

[494] European Commission, 'Media freedom and pluralism', *Shaping Europe's digital future,* n.d. available at: https://digital-strategy.ec.europa.eu/en/policies/media-freedom.

providers.[495] It amends certain provision of Directive 2010/13/EU on audiovisual media services. Amongst the different challenges identified by the proposal, two are especially relevant for media pluralism in relation to social media. The first is 'the increasing digitalisation of media service distribution, and the risks to free provision of media services on very large online platforms, to the detriment of a level playing field of the internal market.'[496] Second is 'the opacity and possible biases in audience measurement systems also to the detriment of the level playing field in the internal market'.[497] These risks are thus associated with the competitive advantage platforms gain from controlling, and not disclosing, audience measurement systems, and the risks to legacy and professional media posed by digitisation of the news ecosystem. These topics are mainly addressed in Articles 17 and 23 of the proposed Act.

Article 17 deals with content from media service providers on very large online platforms. The definition of 'media service' follows the definition in the Treaty of the EU, that is, media where the principal purpose of the service 'consists in providing programmes or press publications to the general public, by any means, in order to inform, entertain or educate, under the editorial responsibility of a media service provider.'[498] A media service provider is a natural or legal person whose professional activity is to provide a media service and has editorial choice - such as news anchors, newspapers and journalists.[499] Very large online platform is defined in accordance with the DSA as 'online platforms which provide their services to a number of average monthly active recipients of the service in the Union equal to or higher than 45 million', and includes leading social media services like Facebook.[500] According to Article 17, social media platforms that exercise editorial responsibility must provide explanations to media service providers when they consider that these providers' content is incompatible with their terms and conditions. They should endeavour to do this before the restriction takes effect, although this should not prevent social media companies from taking expedited measures against illegal content.[501]

Additionally, Article 17(1) provides that very large online platforms must provide a functionality allowing recipients of their services to declare that they are a media service provider as defined by the Act; that they are independent; and that they are subject to regulatory requirements for the exercise of editorial responsibility in one or more Member States, or adhere to equivalent self-regulatory or co-regulatory and widely recognised standards.'[502] The Directive further establishes a European Board for Media Services, which will be independent and will be in charge of applying the Directive, promoting cooperation between national regulators, and advising the Commission on issues related to the application of the Directive.[503] Article 17(4) establishes that media service providers can submit a

[495] European Commission, 'Explanatory Memorandum', Proposal for a Regulation of the European Parliament and of the Council establishing a common framework for media services in the internal market (European Media Freedom Act) and amending Directive 2010613/EU, Brussels, 16 September, 2022.

[496] European Commission, 'Explanatory Memorandum', Proposal for a Regulation of the European Parliament and of the Council establishing a common framework for media services in the internal market (European Media Freedom Act) and amending Directive 2010613/EU, Brussels, 16 September, 2022, p. 8.

[497] European Commission, 'Explanatory Memorandum', Proposal for a Regulation of the European Parliament and of the Council establishing a common framework for media services in the internal market (European Media Freedom Act) and amending Directive 2010613/EU, Brussels, 16 September, 2022.

[498] Art. 2(1) of the Media Service Directive.

[499] Art. 2(2) of the Media Service Directive.

[500] DSA, Art. 25.

[501] Proposed Regulation Media Freedom Act, Article 17.1.

[502] Proposed Regulation Media Freedom Act, Article 17.1.

[503] Proposed Regulation Media Freedom Act, Article 17.

declaration to this Board when they consider that 'a provider of a very large online platform frequently restricts or suspends the provision of its services in relation to content provided by the media service provider without sufficient grounds.' In such circumstances, 'the provider of a very large online platform shall engage in a meaningful and effective dialogue with the media service provider, upon its request, in good faith with a view to finding an amicable solution for terminating unjustified restrictions or suspensions and avoiding them in the future. The media service provider may notify the outcome of such exchanges to the Board.'[504]

Article 23 of the Act contains the audience measurement provisions, which seek to address the transparency problem by mandating that platforms that provide audience measurements disclose them to media service providers. The second paragraph establishes that '[w]ithout prejudice to the protection of undertakings' business secrets, providers of proprietary audience measurement systems shall provide, without undue delay and free of cost, to media service providers and advertisers, as well as to third parties authorised by media service providers and advertisers, accurate, detailed, comprehensive, intelligible, and up-to-date information on the methodology used by their audience measurement system. This provision shall not affect the Union's data protection and privacy rules.'[505] These obligations are without prejudice to new and related obligations under the Digital Markets Act, including those related to sharing information on rankings and self-preferencing.[506]

5.5.2. The Copyright Directive

As outlined in Chapter 2, Article 15 of the 2019 Copyright Directive creates a right ancillary to copyright which arises when press publications are reproduced and made available by online publishers. It provides that 'Member States shall provide publishers of press publications established in a Member State with the rights provided in Article 2 and Article 2(3) of Directive 2001/29/EC *for the online use of their press publications by information society service providers*.'[507] (italics added). The rights at stake are 'the exclusive right to authorise or prohibit direct or indirect, temporary or permanent reproduction by any means and in any forms, in whole or in part,'[508] and the 'exclusive right to authorise or prohibit the making available to the public, by wire or wireless means, in such a way that members of the public may access them from a place and at a time individually chosen by them.'[509] Nevertheless, hyperlinking, individual words or 'very short extracts of a press publication' are not covered.[510] According to Recital 56, the protection also excludes 'websites, such as blogs, that provide information as part of an activity that is not carried out under the initiative, editorial responsibility and control of a service provider, such as a news publisher.'[511] This right expires two years after the press publication is published.[512]

Press publications, the object of the protection, are defined by Article 2(4) as 'a collection composed mainly of literary works of a journalistic nature, but can also include other works or other subject matter, and which (a) constitutes an individual item within a periodical or regularly updated publication under a single title, such as a newspaper or a general or special interest magazine; (b) has the purpose of providing the general public with information related to news or other topics; and (c) is published in

[504] Proposed Regulation Media Freedom Act, Article 17.
[505] Proposed Regulation Media Freedom Act, art. 23.
[506] Proposed Regulation Media Freedom Act, numeral 46.
[507] DSM, Article 15.1.
[508] Directive 2001/29/EC, Art. 2.
[509] Directive 2001/29/EC, Art. 3.2.
[510] DSM Article 15.1.
[511] DSM, art. 15.1, Recital 56.
[512] DSM, art. 15.4.

any media under the initiative, editorial responsibility and control of a service provider. Periodicals that are published for scientific or academic purposes, such as scientific journals, are not press publications for the purposes of this Directive.'[513]

Commentators have noticed that the protection is very broad and ill-defined.[514] Article 2(4) states that the subject matter must be of journalistic nature, but need not only be journalistic; there is no limiting requirement that the content is original, or has been expensive to produce. Richard Danbury notes that '(t)he absence of an expenditure requirement is particularly curious, seeing as the rationale for the right as set out in the recitals is the fact that large Internet companies are free riding on news publishers' investments.'[515] This broad definition creates uncertainty but also broad scope for infringements, which may raise the protection of established players at the cost of dissuading new entrants - many blogs, for example, evolve to become journalism outlets.[516] This encourages the concentration of the market and could thus harm media pluralism.[517] Indeed, a letter directed at the European Commission signed by 160 European academics working in related fields highlighted that the protection is likely to raise transaction costs significantly, as permission will be needed for virtually any use of news content. It could even disadvantage journalists and other non-institutional creators and producers of news, since payments are due to institutional news providers.[518]

The intended effect of the new press publishers' right is to force online platforms which link to news stories with short excerpts of their content - including social media, but also Google and other search engines - to negotiate and pay for licences to use that content. In effect, this right will give press publishers leverage in negotiations and allow them to secure a new revenue stream. In France, the first member state to transpose the CD, national authorities have taken an active role in support of media publishers, with ADLC (the French competition authority) ruling that Google's dominant market position meant it could be required to negotiate with publishers in good faith, rather than simply opting not to publish snippets of news content alongside search results, as it had previously done in Spain and Germany after similar legislation was introduced.[519] Meta and Google have now agreed payment schemes with publishers in France and other Member States.

However, experts have highlighted that these negotiation processes and payment schemes tend above all to benefit the largest publishing companies and associations, due to their political influence and greater capacities to bargain with platforms.[520] By providing funding to crisis-hit sectors of the news industry and funding training programmes and projects which involve closer collaboration with platforms and use of their own services, Google and Meta can ultimately reinforce publishers'

---

513 DSM Directive, Art. 2(4).

514 Danbury, R., 'The DSM Copyright Directive: Article 15: What? Part II', *Kluwer Copyright Blog*, April 29, 2021, available at: http://copyrightblog.kluweriplaw.com/2021/04/29/the-dsm-copyright-directive-article-15-what-part-ii/.

515 Danbury, R., 'The DSM Copyright Directive: Article 15: What? Part II', *Kluwer Copyright Blog*, April 29, 2021, available at: http://copyrightblog.kluweriplaw.com/2021/04/29/the-dsm-copyright-directive-article-15-what-part-ii/.

516 Danbury, R., 'The DSM Copyright Directive: Article 15: What? Part II', *Kluwer Copyright Blog*, April 29, 2021, available at: http://copyrightblog.kluweriplaw.com/2021/04/29/the-dsm-copyright-directive-article-15-what-part-ii/.

517 Danbury, R., 'The DSM Copyright Directive: Article 15: What? Part II', *Kluwer Copyright Blog*, April 29, 2021, available at: http://copyrightblog.kluweriplaw.com/2021/04/29/the-dsm-copyright-directive-article-15-what-part-ii/.

518 Ricolfi, M. et. al., 'Academics against Press Publishers' Right: 169 European Academics warn against it,' n.d, available at: https://www.ivir.nl/publicaties/download/Academics_Against_Press_Publishers_Right.pdf.

519 Papaevangelou, C. and Smyrnaios, N., 'The political stakes of online platforms' deals with French publishers', *HAL Sciences Humaines et Sociales*, 2022, available at: https://shs.hal.science/halshs-03747847/.

520 Papaevangelou, C. and Smyrnaios, N., 'The political stakes of online platforms' deals with French publishers', *HAL Sciences Humaines et Sociales*, 2022. https://shs.hal.science/halshs-03747847/; Papaevangelou, C., 'Funding Intermediaries: Google and Facebook's Strategy to Capture Journalism', *Digital Journalism*, 2023.

dependence on their platforms and strengthen their own influence over the news industry.[521] There is a concerning lack of transparency about how platforms allocate funding and which organisations ultimately benefit.[522] As such, while they could increase the funding available for news journalism and provide more sustainable revenue streams for publishers, their implications for media pluralism might ultimately be mixed or negative.[523] Victor Pickard, a leading expert in the political economy of news journalism, has argued that rather than trying to tweak economic incentives in a market that will ultimately continue to favour the biggest news conglomerates, governments should directly tax platforms and use the revenue to subsidise public media, as well as smaller, local and independent news organisations.[524]

However, such subsidies cannot be regarded as a simple solution, as they do not address all of the impacts of social media on media pluralism - for example, platforms' influence over editorial decisions - and raise further difficult policy questions, such as how to prevent governments from using subsidy programmes to favour preferred media organisations and agendas. Establishing effective schemes for governments to identify which media outlets to subsidise - without favouring certain sectors of the news media market, or giving governments dangerous levels of influence over news media - is a formidable challenge. However, carefully-designed subsidy programmes should be one element of the EU's media pluralism policy and could achieve Article 15 CD's ultimate aim of transferring revenue from platforms to news publishers in a way that more effectively promotes media pluralism. Box 14 examines one promising proposal as to how this could be achieved.

Box 14: Indirectly subsidising media through 'journalism vouchers'

Julia Cagé, economist and professor at Sciences Po Paris, has argued that strengthening the 'critical infrastructure of democracy' - political parties and independent media - and preventing them from being captured by the wealthy requires more equal and democratic participation in allocating funding. In Cagé's view, instead of subsidising media organisations directly (which raises concerns about their independence and about state influence), the best way of ensuring media autonomy is crowdfunding it, through so-called 'journalism vouchers'. These would be vouchers of a set amount, issued and funded by governments and distributed to each citizen for them to donate to their media organisation(s) of choice. Cagé suggests that this could strengthen accountability, participation, and representation by providing sustainable funding for media which would be independent from both the market and the state.

This measure offers many potential advantages: under the principle of equality of opportunity in political participation, journalism vouchers will enable all citizens to finance the media outlets and journalists they wish to hear more from, in an egalitarian manner which does not depend on individual ability to pay. This could not only strengthen the sustainability and independence of the media, but reduce bias in favour of wealthy audiences and donors. As Jan-Werner Müller suggests, if media outlets 'contribute to citizen judgment, it matters that citizens can also judge them'.

---

[521] Papaevangelou, C., 'Journalism, Platforms, and the Challenges of Public Policy', *Tech Policy Press*, February 6, 2023, available at: https://techpolicy.press/journalism-platforms-and-the-challenges-of-public-policy/.

[522] Papaevangelou, C., 'Journalism, Platforms, and the Challenges of Public Policy', *Tech Policy Press*, February 6, 2023, available at: https://techpolicy.press/journalism-platforms-and-the-challenges-of-public-policy/.

[523] Pickard, V. 'Can Journalism Survive in the Age of Platform Monopolies? Confronting Facebook's Negative Externalities', in Flew, T. and Martin, F.R., *Digital Platform Regulation: Global Perspectives on Internet Governance*, Palgrave Macmillan, 2022, pp. 23-41.

[524] Pickard, V. 'Can Journalism Survive in the Age of Platform Monopolies? Confronting Facebook's Negative Externalities', in Flew, T. and Martin, F.R., *Digital Platform Regulation: Global Perspectives on Internet Governance*, Palgrave Macmillan, 2022, pp. 23-41.

Journalism vouchers would give everyone a fair way to influence power and visibility in the media.

Nonetheless, such a scheme could also have disadvantages. It might reinforce winner-takes-all effects, where most people just donate to the biggest or best-known organisations. To the extent that this results from a lack of visibility of less popular outlets, it cannot only be corrected through allocating funding for news production, but also requires consideration of distribution and exposure. Citizens may engage in strategic voting, for example by funding highly partisan and polarised media as a way of supporting their favoured political causes, instead of those that provide the most reliable news reporting. Finally, implementing such a scheme would require state institutions to determine many practical details - for example, which organisations are eligible to receive voucher funding, and under what criteria. These could undermine the scheme's aim of securing independence from state interests, by allowing governments to favour their preferred outlets.

Overall, journalism voucher schemes are promising, but raise many unresolved questions. A good first step would be to introduce pilot schemes at the local or regional level, which could provide learnings and best practices for potential wider application.

Sources: Cagé, J., *Saving the Media: Capitalism, Crowdfunding, and Democracy*, Harvard University Press, 2016; Müller, J. 'Liberal Democracy's Critical Infrastructure. How to think about Intermediary Powers', Scripts Working Paper No. 16. Berlin, Scripts-Berlin, Nov. 16, 2022.

5.6. Recommendations

a. Strengthening independent and professional journalism

- EU policymakers should explore funding and policy programmes to strengthen independent and professional journalism, for example by subsidising independent newspapers and broadcasters. In this context, it is essential that funding projects are structured in a way that maintains the independence of the funded entities, limiting the influence of state institutions and private interests in allocating funding and shaping editorial decisions.
- Along these lines, the Council of Europe Declaration by the Committee of Ministers on the financial sustainability of quality journalism in the digital age promotes the implementation of mechanisms to ensure financial sustainability in national media ecosystems.[525] The Declaration includes major online platforms, which have strongly impacted advertising and broadcasting. Among the measures included, the Declaration encourages member states to financially support schemes for regional, local and not-for-profit media, a beneficial tax regime for the production and distribution of journalistic content, and other funding schemes including private-public partnerships to support quality journalism.[526]
- One way of achieving this would be to increase EU funding for journalism funding programmes at the national and regional level, which are already established and regarded as ensuring effective safeguards for journalistic independence. EU institutions should also serve as a forum

[525] Committee of Ministers, Council of Europe, 'Declaration by the Committee of Ministers on the financial sustainability of quality journalism in the digital age.' Adopted on 13 February 2019, available at: https://eos.cartercenter.org/uploads/document_file/path/733/13._Declaration_on_Sustainability_of_Journalism_in_Digital_Age_EN.pdf.

[526] Committee of Ministers, Council of Europe, 'Declaration by the Committee of Ministers on the financial sustainability of quality journalism in the digital age.' Adopted on 13 February 2019, p. 3, available at: https://eos.cartercenter.org/uploads/document_file/path/733/13._Declaration_on_Sustainability_of_Journalism_in_Digital_Age_EN.pdf.

for the dissemination of knowledge and best practices across Europe with regard to such programmes.

- In cooperation with Member State governments, the EU should also offer funding and support for novel media subsidy programmes, such as 'journalism vouchers' (see Box 14), which offer a promising way to allocate subsidies in a democratic, decentralised manner that minimises direct state influence. Such programmes should be piloted at national or local level, involving independent stakeholders as well as EU and Member State institutions in evaluating their success and developing future best practices.
- Given the evidence that local journalism has been hit particularly hard by the ongoing economic disruptions to news publishing, and that it can be particularly valuable in promoting political accountability, reducing polarisation and contributing to a trustworthy media environment, all such funding programmes should focus particularly on supporting local and regional publishers.

b. DSA enforcement

- With the DSA in force, policymakers should pursue close collaboration with academic and independent researchers to understand and analyse the vast amounts of information platforms will now have to report in relation to their moderation and audience measurement practices. Effective oversight and independent research will help realise the promises of enhanced transparency, so that news publishers will be able to benefit from a better understanding of content moderation, audience measurement, recommender systems and other dynamics of the online media environment.

6. RECOMMENDATIONS

6.1. Introduction

This chapter presents all of the study's key recommendations, based on the in-depth analysis in the foregoing chapters. It first provides a brief overview of the most important points for EU and national institutions in three key areas: implementation and enforcement of the new DSA framework; further legislative reform; and funding and policy programmes. This is followed by a comprehensive recap of all recommendations in each of the three policy areas examined: hate speech, disinformation and media pluralism.

6.2. Core priorities

6.2.1. DSA enforcement

The DSA came into force on November 2022, and will be directly applicable across the EU in early 2024. This will be the start, rather than the end of a reform process. Building the institutional architecture to enforce the DSA, establishing cooperation and best practices among regulators, and developing more concrete norms will be a major project for EU and national institutions in the coming years.[527] This is an important opportunity to ensure the goals of the DSA - strengthening democracy, fundamental rights, and the rule of law in the context of social media - can be successfully realised. In this section, based on our analysis of issues around online hate speech, disinformation and media pluralism, we highlight three key areas that should be priorities in the coming years: developing a new Code of Conduct on Hate Speech to further develop and concretise platforms' obligations in this area, issuing official guidance to clarify the scope and interpretation of the regulation, and ensuring regulators have sufficient technical and human resources for effective, in-depth oversight.

a. A new Code of Conduct on Hate Speech

Given the abstract and open-ended nature of important DSA provisions – in particular very large online platforms' obligations to assess and mitigate systemic risks – developing codes of conduct under Article 45 will be an important tool to further concretise these obligations, and to establish more specific, stringent and consistent standards for regulatory compliance. A key priority for the Commission should be to drive forward the development of an expanded and updated Code of Conduct on Hate Speech, following the example of the successful effort to update and strengthen the CoP on Disinformation.

As detailed in Chapter 3, platforms' existing legal obligations to moderate reported instances of illegal hate speech are insufficient to protect marginalised users from hate and harassment, while strengthening these obligations would pose severe risks to freedom of expression and non-discrimination. Developing a new Code of Conduct with a focus on improving design and operational practices to provide a safer environment for users would effectively concretise platforms' obligations to address systemic risks to their users' safety and equality, and create clear incentives for them to invest in improving their safety and equality policies. The Commission should convene maximally diverse and inclusive multistakeholder discussions to begin drafting such a Code.

As starting points for such a drafting process, this new code should consider establishing a broader definition of hate speech which recognises intersectional oppression, and should additionally require

[527] https://mastodon.social/@jjaursch/109631847758379330.

signatories to tackle other forms of harassment and abuse which target marginalised groups. It should require platforms to invest significantly more in adequately trained, paid and supported moderation staff, and in reliable and thoroughly tested technical tools, in order to effectively identify and respond to hate speech in all languages and markets where they operate. It should also require them to investigate, develop and test proactive measures, including design changes, to discourage hate speech and create safer online environments.

b. Guidance and interpretation

Regulators can also further concretise platforms' DSA obligations by issuing official guidance on how they will interpret relevant provisions for the purposes of evaluating compliance (as provided for example by Article 35(3) on systemic risks). In this context, the Commission and national DSCs should issue guidance clarifying that obligations to have regard to fundamental rights under Article 14 and to address systemic risks to rights under Articles 34-35 preclude the use of indiscriminate or clearly discriminatory automated moderation systems, and that they require adequate moderation capacities in all languages widely spoken by a platform's users. The guidance should further require platforms to clearly document the design, operation and performance of their automated and manual moderation processes.

As regards disinformation, the guidance should build on the existing commitments to implementing 'safe design practices' set out in the CoP on Disinformation by emphasising that such practices should be the primary response to disinformation, and that legal disinformation content should only be deleted where it poses immediate dangers to public safety.

c. Capacity building

The Commission and national DSCs should further ensure that they invest sufficiently in staff with relevant technical and UX/UI design expertise to be able to effectively oversee and enforce the DSA and the relevant codes and guidance. In this regard, to use such resources effectively, collaboration and knowledge-sharing between regulators are also essential.

6.2.2. Legislative reform

a. Regulating moderation work

Online hate speech and disinformation are complex issues, and dealing with them raises many intractable problems and deeply contested questions; however, much could be achieved simply by requiring sufficient investments in staff and resources to consistently implement established best practices across all markets. Available evidence clearly indicates that, even at the biggest and most highly-resourced platforms, investment in moderators and other trust and safety staff is far from adequate – particularly in smaller European (and global) markets and those whose languages are less widely spoken. In addition, research indicates that the labour of content moderation – an essential service which protects fundamental rights and access to online media for society as a whole – is precarious, unsafe and undervalued.

To address the poor working conditions of content moderators, while also improving the quality and reliability of content moderation, the Commission should consult on the possibilities to propose EU-level legislation regulating the staffing and operation of content moderation teams. Given the prevalence of outsourcing, this should apply broadly to platforms operating in the EU, whether or not the moderation staff are based inside the EU. Such legislation could, for example, establish minimum staffing levels for the various languages and markets in which a platform operates, and regulate

moderation workers' training, working conditions and hours. In the absence of such EU legislation, Member State governments should consider the possibilities of proposing similar legislation at national level as regards moderators based in and/or moderating content from that Member State.

b. Safeguards against censorship

Certain aspects of the existing intermediary liability framework create significant risks of state censorship which threaten EU citizens' rights to freedom of expression, freedom of information and non-discrimination, as well as endangering the freedom of democratic debate more broadly. In particular, the deferral to national law to define 'illegal content' which platforms can be required to remove, without further fundamental rights safeguards, means that a wide range of national speech laws which are highly problematic from a fundamental rights perspective can be used by state authorities or private individuals to demand removal of social media content. Importantly, using the notice-and-takedown framework enables state authorities to circumvent the legal safeguards attached to formal removal orders, and given the business incentives created by liability risks, such laws can be used to put effective pressure on platforms to remove content even where it is doubtful that the law could be applied or enforced against the user posting the content.

This creates broad possibilities for unaccountable censorship and should be an urgent priority for legislative reform. The Commission should begin consultations including civil society and fundamental rights experts on the possibility of introducing further fundamental rights safeguards within the harmonised EU intermediary liability framework: for example, by specifying that platforms only lose their immunity for hosting known illegal content if that content poses a direct risk to public safety or the fundamental rights of others, and ensuring effective judicial oversight of this condition.

c. Targeted advertising

The capacity to target political advertising to narrowly-defined segments of the population (microtargeting) undermines free and open democratic debate and equal political participation by all citizens. While it can be used in disinformation operations, these risks are much broader: microtargeting with accurate information still creates many possibilities to evade accountability for political claims and rhetoric, exploit and exacerbate social divisions, and exclude certain audiences from political debate.

Importantly, banning targeting of political adverts based on 'sensitive data' (race, religion, sexuality etc) does very little to mitigate these risks, which rather arise from the detailed profiling and segmenting of audiences based on more specific combinations of characteristics. Segmenting audiences based on non-sensitive characteristics will still tend to, and can intentionally be used to, create audiences which correspond closely with existing patterns of discrimination and marginalisation. In light of these considerations, EU legislators should positively consider banning microtargeting of political ads entirely in the draft Political Advertising Regulation, permitting targeting only based on certain broad characteristics like location.

6.2.3. Funding and policy programmes

a. Support for trust & safety professional associations

The development of industry-wide professional associations for trust and safety professionals represents a promising way to develop safe design practices and best practices for risk mitigation; to strengthen the position of platform companies' employees when they attempt to mitigate risks and prevent ethical abuses; and to leverage industry expertise for more effective regulation. The EU should

support existing trust and safety associations both financially (for example, by making grants available for research projects) and practically (for example, by inviting them to participate in stakeholder discussions on DSA codes of conduct and other relevant multistakeholder processes and consultations). It should also encourage the development of European and regional professional associations.

b. Media literacy programmes

Media literacy should not be over-relied upon in tackling disinformation, but should be one element of a holistic approach. The most effective approaches to enhancing media literacy, and those that are most relevant to addressing disinformation and strengthening trust in the media ecosystem, are likely to vary strongly across Member States. The EU's role in this area should thus be to facilitate the development of existing successful programmes and the dissemination of evidence-based approaches to media literacy education at the national and subnational levels, for example through grant funding.

c. Subsidising independent media

An ecosystem of trustworthy, independent and pluralistic media institutions with sufficient resources to provide essential media services and hold political actors accountable requires funding sources that are not solely reliant on advertising and other marketised business models. In this context, European policymakers should built on and extend existing traditions of public service media and subsidising journalism to provide additional public funding for independent news media, in particular at the local level. A citizen's voucher system which decentralises choices about how to direct funding to the population as a whole could be one promising way of doing this, though implementing it at the European level would be technically complex and piloting it in one or more Member States could be a helpful interim step. EU policymakers should begin multistakeholder consultations on the best approaches to extending state subsidies for independent media while safeguarding journalistic independence.

As a shorter-term measure, EU institutions could also extend financial support for independent media outlets – including and especially at the local level – to provide fact-checking services and disseminate reliable scientific information in public health emergencies and other crisis situations. This would simultaneously provide counter-narratives to disinformation regarding crises, and provide an additional revenue stream for such media organisations. Establishing a clear framework for decision-making, accountability and oversight would however be essential to avoid actual or perceived threats to journalistic independence.

6.3. Detailed recommendations

6.3.1. Hate speech

a. A new Code of Conduct on Online Hate Speech

- In order to strengthen and concretise very large online platforms' obligations to mitigate systemic risks under the DSA, the Commission should take the lead on establishing multistakeholder discussions to update and expand the 2016 Code of Conduct on Hate Speech and Harassment.
- These discussions should include a diverse range of independent researchers and civil society organisations from all over Europe. Representing marginalised communities such as Roma people, LGBTQ+ people and migrants should be a top priority in convening these discussions. Funding should be available to support participation by organisations who may otherwise lack the resources.

As a starting point, the new code of conduct should:

- Establish a broader definition of online hate speech as incitement to hatred or violence based on any characteristic protected by Article 21 of the Charter of Fundamental Rights. To recognise intersectional forms of marginalisation, this should also extend to combinations of characteristics where any one of those characteristics is protected by the Charter.
- Broaden the scope of platforms' obligations beyond hate speech. Platforms should additionally commit to tackle all forms of threats, harassment and privacy violations which target a person or group based on a protected characteristic.
- Require platforms to establish adequate moderation staff and technical resources for all languages which are widely spoken in markets where they operate, and to publish detailed reports on their moderation capabilities in all such languages.
- Establish clear and specific commitments from platforms to investigate, develop and test proactive measures (including design changes) to discourage hate speech and support affected users. Platforms should also commit to ongoing consultation and participation from stakeholder groups representing affected communities as part of these processes.
- Establish clear and specific standards on the working conditions (e.g. pay, training, performance quotas, working hours, psychological support) of all platform staff working on content moderation. These should also apply to staff working on behalf of a platform via outsourcing companies, and staff based outside the EU.

b. DSA enforcement

- The Commission and national DSCs should issue guidance stating that, in accordance with the ECJ decision in Poland v Parliament and Council [2022], platforms' obligations to have due regard to fundamental rights (under Article 14(4) DSA and Article 5 TCR) and very large online platforms' obligations to address systemic risks to fundamental rights (under Articles 34-35 DSA) preclude the use of automated moderation tools which are indiscriminate or clearly discriminatory. The guidance should further state that platforms must clearly document the design, use, performance and outcomes of such tools, including industry-standard accuracy and bias metrics, to establish regulatory compliance.

- The Commission and national DSCs should also issue guidance stating that the obligation for platforms to enforce their content policies in a diligent, objective and proportionate manner under Article 14(4) DSA requires adequate moderation capacities in all languages widely spoken by their users, including adequate investment in competent moderation staff. All relevant moderation processes should be clearly and publicly documented to establish compliance.
- In overseeing and enforcing very large platforms' systemic risk mitigation obligations under Articles 34-35 DSA, the Commission should place significant weight on design changes and other interventions which aim to proactively discourage and prevent the occurrence of online hate speech, harassment and other systemic risks, as opposed to moderating or removing content retroactively. Risk assessments and audit reports which indicate that platforms are not investing in such proactive risk mitigation measures should not be regarded as compliant.
- The Commission and national regulators should ensure that they have sufficient staff with relevant technical and UX/UI design expertise to effectively assess compliance with these obligations. This would also be aided by effective procedures for collaboration, co-investigations and knowledge sharing between different regulatory agencies.

c. Legislative reform

The Commission should consider and consult on proposing EU-level legislation to regulate the staffing and operation of platforms' content moderation teams. This could include:

- Minimum thresholds for numbers of staff with relevant language and market expertise for each EU country in which a platform operates;
- Regulation of the working conditions (e.g. training, performance quotas, working hours, psychological support) of content moderation staff.
- Member States should consider similar legislation to regulate the staffing and working conditions of content moderation staff based in the relevant Member State and/or moderating content from that Member State.

6.3.2. Disinformation

a. DSA enforcement

- Building on the obligations and commitments already established in the DSA and Code of Practice, the Commission and national DSCs should issue guidance stating that safe design practices should be a primary line of defence against disinformation, and should be prioritised over content moderation except where disinformation directly endangers the public or threatens the rights of others.
- In overseeing and enforcing very large platforms' systemic risk mitigation obligations under Articles 34-35 DSA, the Commission should place significant weight on design changes and other interventions which aim to proactively discourage and prevent the occurrence of online hate speech, harassment and other systemic risks, as opposed to moderating or removing content retroactively. Risk assessments and audit reports which indicate that platforms are not investing in such proactive risk mitigation measures should not be regarded as compliant. The Commission should ensure that it has sufficient staff with expertise in UX/UI design to effectively assess compliance with these obligations.

- The Commission and national DSCs should also issue guidance stating that the obligation for platforms to enforce their content policies in a diligent, objective and proportionate manner under Article 14(4) DSA requires adequate moderation capacities in all languages widely spoken by their users, including adequate investment in competent moderation staff. All relevant moderation processes should be clearly and publicly documented to establish compliance.

b. Legislative reform

- The Commission should consider and consult on amending Articles 3 and 6 DSA to create a narrower and more fundamental-rights-compliant definition of 'illegal content' which can attract liability for platforms. For example, the amended DSA could specify that platforms retain their intermediary liability immunity even where they have knowledge of illegal content, except where that content creates a direct and specific threat to public safety or the fundamental rights of others.
- The Commission should positively consider proposals to entirely ban or very significantly restrict the personalised targeting of political advertising, recognising that microtargeting of political messaging has negative impacts for civic and political debate even where it does not infringe the rights of individual users.

c. Strengthening trust and safety

- Recognising that countering organised disinformation operations and other emerging threats requires flexible response capacities within the social media industry and civil society, EU policy should make it a priority to strengthen the online trust and safety profession. This should include support for professional associations of platform engineers and moderation staff and consultation with such organisations in the development of industry best practices and safety standards under the DSA.

d. Enhance media literacy, but with caution

- Through media campaigns, in schools, and in other civic spaces, the EU should promote and fund new and existing programmes which teach individuals about best practices to evaluate the reliability of online content, as well as identifying bots and strategically-promoted disinformation read and identify bots, potentially harmful and/or mis-informative content online
- However, policymakers should not over-rely on media literacy as a solution. Not only does it emphasise individual agency and control over more consequential structural issues, research has also shown that it can backfire, as individuals may also learn to doubt trustworthy content.[528] Media literacy education should be one component of a broader policy programme aimed at promoting a trustworthy information environment.

e. Promoting reliable independent media

- As detailed in Chapter 5 on media pluralism, the EU's disinformation policy should be part of a broader policy programme to strengthen independent journalism and trust in media, for example through funding programmes.

[528] Boyd, D. 'You think you want media literacy… do you?', *Data & Society: Points*, March 9, 2018, available at: https://points.datasociety.net/you-think-you-want-media-literacy-do-you-7cad6af18ec2.

- Public media and independent journalism institutions across the EU should be supported to provide fact-checking services and to create easily shareable, accurate information on sensitive political topics (e.g. public health risks, conflict situations).

6.3.3. Media pluralism

a. Strengthening independent and professional journalism

- EU policymakers should explore funding and policy programmes to strengthen independent and professional journalism, for example by subsidising independent newspapers and broadcasters. In this context, it is essential that funding projects are structured in a way that maintains the independence of the funded entities, limiting the influence of state institutions and private interests in allocating funding and shaping editorial decisions.
- Along these lines, the Council of Europe Declaration by the Committee of Ministers on the financial sustainability of quality journalism in the digital age promotes the implementation of mechanisms to ensure financial sustainability in national media ecosystems.[529] The Declaration includes major online platforms, which have strongly impacted advertising and broadcasting. Among the measures included, the Declaration encourages Member States to financially support schemes for regional, local and not-for-profit media, a beneficial tax regime for the production and distribution of journalistic content, and other funding schemes including private-public partnerships to support quality journalism.[530]
- One way of achieving this would be to increase EU funding for journalism funding programmes at the national and regional level, which are already established and regarded as ensuring effective safeguards for journalistic independence. EU institutions should also serve as a forum for the dissemination of knowledge and best practices across Europe with regard to such programmes.
- In cooperation with Member State governments, the EU should also offer funding and support for novel media subsidy programmes, such as 'journalism vouchers' (see Box 14), which offer a promising way to allocate subsidies in a democratic, decentralised manner that minimises direct state influence. Such programmes should be piloted at national or local level, involving independent stakeholders as well as EU and member state institutions in evaluating their success and developing future best practices.
- Given the evidence that local journalism has been hit particularly hard by the ongoing economic disruptions to news publishing, and that it can be particularly valuable in promoting political accountability, reducing polarisation and contributing to a trustworthy media environment, all such funding programmes should focus particularly on supporting local and regional publishers.

b. DSA enforcement

- With the DSA in force, policymakers should pursue close collaboration with academic and independent researchers to understand and analyse the vast amounts of information platforms

---

[529] Committee of Ministers, Council of Europe, 'Declaration by the Committee of Ministers on the financial sustainability of quality journalism in the digital age.' Adopted on 13 February 2019 https://eos.cartercenter.org/uploads/document_file/path/733/13._Declaration_on_Sustainability_of_Journalism_in_Digital_Age_EN.pdf.

[530] Committee of Ministers, Council of Europe, 'Declaration by the Committee of Ministers on the financial sustainability of quality journalism in the digital age.' Adopted on 13 February 2019, p. 3 https://eos.cartercenter.org/uploads/document_file/path/733/13._Declaration_on_Sustainability_of_Journalism_in_Digital_Age_EN.pdf.

will now have to report in relation to their moderation and audience measurement practices. Effective oversight and independent research will help realise the promises of enhanced transparency, so that news publishers will be able to benefit from a better understanding of content moderation, audience measurement, recommender systems and other dynamics of the online media environment.

7. REFERENCES

Case Law

- Cour d'appel de Bruxelles , Belgische Vereniging van Auteurs, Componisten En Uitgevers CVBA (SABAM) v Netlog NV, 2012.
- Cour d'appel de Bruxelles Scarlet Extended SA v Société Belge des Auteurs, Compositeurs et Éditeurs SCRL (SABAM), 2011.
- Judgement of 6 March 2001, Connolly v. Commission of the European Communities, C-274/99 P, ECLI:EU:C:2001:127.
- Judgement of 27 April 2004 of the European Court of Human Rights, Salov v. Ukraine, application no. 65518/01.
- Judgement of 8 October 2008 of the European Court of Human Rights, *Kita v. Poland*, application no. 57659/00.
- Judgement of 2 February 2016, of the European Court of Human Rights, Case of Magyar Tartalomszolgáltatók Egyesülete and Index.Hu Zrt v. Hungary, application no. 22941/13
- Judgement of 7 February 2017 of the European Court of Human Rights, PIHL v. Sweden, application no. 74742/14.
- Judgement of 25 July 2019 of the European Court of Human Rights, Brzeziński v. Poland, application no. 47542/07.
- Judgement of the Court of Justice of 13 May 2014 Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González, C-131/12, ECLI:EU:C:2014:317.
- Judgement of the Court of Justice of May 24, 2019 Republic of Poland v European Parliament and Council of the European Union, C-419/19, ECLI: EU: C:2019:270:TOC.
- Meta Oversight Board, *UK music drill 2022-007-IG-MR*.

Policy Documents

- European Commission (EC) (L 124/36) Recommendation 2003/361/EC, concerning the definition of micro, small and medium-sized enterprises (notified under document number C(2003) 1422), 6 May 2003.
- European Commission, '2022 Strengthened Code of Practice on Disinformation ', Policy and Legislation, June 16, 2022.
- European Commission, 'Commission Welcomes European Parliament´s Adoption of Digital Services Package | Shaping Europe's Digital Future', n.d. https://digital-strategy.ec.europa.eu/en/news/commission-welcomes-european-parliaments-adoption-digital-services-package.
- European Commission, 'Explanatory Memorandum', Proposal for a Regulation of the European Parliament and of the Council establishing a common framework for media services in the internal market (European Media Freedom Act) and amending Directive 2010613/EU, Brussels, 16 September, 2022.

- European Commission, 'Guidance on Strengthening the Code of Practice on Disinformation', Brussels, 2021.
- https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52021DC0262&qid.
- European Commission, 'Media freedom and pluralism', Shaping Europe's digital future, n.d. https://digital-strategy.ec.europa.eu/en/policies/media-freedom.
- European Commission, 'The EU Code of Conduct on Countering Illegal Hate Speech Online', n.d., p.1. https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en.
- European Commission, 2022 Rule of law report, n.d. July 2022, https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/upholding-rule-law/rule-law/rule-law-mechanism/2022-rule-law-report_en.
- European Commission, SWD (2020)180 Final - Assessment of the Code of Practice on Disinformation, 2020, https://digital-strategy.ec.europa.eu/en/library/assessment-code-practice-disinformation-achievements-and-areas-further-improvement.
- European Data Protection Board, 'Facebook and Instagram Decisions: "Important Impact on Use of Personal Data for Behavioural Advertising"', European Data Protection Board, January 12, 2023. https://edpb.europa.eu/news/news/2023/facebook-and-instagram-decisions-important-impact-use-personal-data-behavioural_hu.
- European Digital Rights, 'On New Crisis Response Mechanism and Other Last Minute Additions to the DSA,' EDRi, April 12, 2022, https://edri.org/wp-content/uploads/2022/04/EDRi-statement-on-CRM.pdf.
- European Parliament. Directorate General for External Policies of the Union., The Impact of Disinformation Campaigns about Migrants and Minority Groups in the EU: In Depth Analysis., Publications Office, LU, 2021.

Research and reports

- Aggarwal, M., et al., 'A 2 Million-Person, Campaign-Wide Field Experiment Shows How Digital Advertising Affects Voter Turnout', *Nature Human Behaviour*, January 12, 2023.
- Ahmad, S., and M. Greb, 'Automating Social Media Content Moderation: Implications for Governance and Labour Discretion', *Work in the Global Economy, Vol. 2, No. 2,* November 2022, pp. 176–198.
- Alaphilippe, A., et. al. 'Automated Tackling of Disinformation: Major Challenges Ahead, European Parliament,' European Parliament Think Tank, Brussels, 2019.
- Alexander, J., 'YouTube Moderation Bots Punish Videos Tagged as "Gay" or "Lesbian," Study Finds', *The Verge*, September 30, 2019.
- Andi, S., 'How and why do consumers access news on social media?,' Reuters Institute Digital News Report 2021.

- Andrejevic, M., *Automated Media,* Routledge, London ; New York, NY, 2020.
- Antheaume, A. et. al.,'The Changing Business of Journalism and its Implications for Democracy Reuters Institute for the Study of Journalism, Department of Politics and International Relations, University of Oxford, 2010.
- Appelman, N., and Leerrssen, P., 'On "Trusted" Flaggers', *Yale-Wikimedia Initiative on Intermediaries & Information*, July 12, 2022.
- Aristotle, Nicomachean Ethics, Book VIII, Chapter 10 (1160a.31-1161a.9). Internet Classics Archive. Retrieved 21 June 2018.
- Armitage, C., Botton, N., Dejeu-Castang, L., and Lemoine, L., 'Towards a more transparent, balanced and sustainable digital advertising ecosystem: Study on the impact of recent developments in digital advertising on privacy, publishers and advertisers,' Publications Office of the European Union, January 31, 2023.
- Bakshy, E., et. al., 'Exposure to ideologically diverse news and opinion on Facebook', *Science Vol. 348, Issue 6239, 2015*, p. 1130–1132.
- Barata, J., 'The Digital Services Act and Its Impact on the Right to Freedom of Expression: Special Focus on Risk Mitigation Obligations', DSA Observatory, 2021.
- Barberá, P., Vaccari, C., Valeriani, A., 'Social Media, Personalisation of News Reporting, and Media Systems' Polarisation in Europe', *Social Media and European Politics*, Palgrave Macmillan, United Kingdom, 2017, pp. 25-52.
- Barrett P., and Hendrix J., 'A Platform 'Weaponized': How YouTube Spreads Harmful Content—And What Can Be Done About It', NYU Stern, 2022.
- Batorski, D., and Grzywińska, I., 'Three Dimensions of the Public Sphere on Facebook', *Information, Communication & Society, Vol. 21, No. 3*, March 4, 2018, pp. 356–374.
- Bayer J., et. al., European Parliament. Directorate General for External Policies of the Union., Disinformation and Propaganda: Impact on the Functioning of the Rule of Law in the EU and Its Member States : 2021 Update., Publications Office, LU, 2021.
- Bayer, J., et. al., 'The fight against disinformation and the right to freedom of expression,' European Union, 2021.
- Bayer, J., and Bárd, P., 'Hate Crime and Hate Speech in Europe: Comprehensive Analysis of International Law Principles, EU-Wide Study and National Assessments', Policy Department for Citizens' Rights and Constitutional Affairs, July 2020.
- Bell, E., and Owen, T., 'The Platform Press: How Silicon Valley Reengineered Journalism', Columbia Journalism Review, Tow Center for Digital Journalism at Columbia's Graduate School of Journalism, New York, 2017.
- Bellanova, R., and M. de Goede, 'Co-Producing Security: Platform Content Moderation and European Security Integration', *JCMS: Journal of Common Market Studies, Vol. 60, No. 5*, September 2022, pp. 1316–1334.

- Benkler, Y., *The Wealth of Networks*, Yale University Press, New Haven, US.
- Bennett, L., et. al., 'Treating Root Causes, Not Symptoms: Regulating Problems of Surveillance and Personal Targeting in the Information Technology Industries', Hertie School, 2021.
- Bennett, O., 'The Promise of Financial Services Regulatory Theory to Address Disinformation in Content Recommender Systems', *Internet Policy Review, Vol. 10, No. 2*, May 11, 2021.
- Bhatia A., 'Election Disinformation in Different Languages is a Big Problem in the U.S.', Center for Democracy and Technology, 2022.
- Big Brother Watch, Ministry of Truth: The secretive government units spying on your speech, n.d. 2023.
- Bingham, T., The Rule of Law, Penguin Books Limited, 2011.
- Bloch-Wehba, H., 'Content Moderation as Surveillance,' *Berkeley Technology Law Journal, Vol. 36, Iss. 3, 2022*, pp. 1297-1340.
- Borelli, M., 'Social Media Corporations as Actors of Counter-Terrorism', New Media & Society, Vol. 0, No. 0, August 8, 2021
- Borges do Nascimento, I.J., et. al., 'Infodemics and Health Misinformation: A Systematic Review of Reviews', *Bulletin of the World Health Organization, Vol. 100, No. 9, September 1, 2022,* pp. 544–561.
- Borgesius, F.J., et. al., 'Online Political Microtargeting: Promises and Threats for Democracy', *Utrecht Law Review, Vol. 14, No. 1*, February 9, 2018, p. 82.
- Bradshaw S., et. al., 'Industrialized Disinformation: 2020 Global Inventory of Organised Social Media Manipulation. Working Paper 2021.1', Project on Computational Propaganda, Oxford, 2021.
- Bridy, A., 'The Price of Closing the Value Gap: How the Music Industry Hacked EU Copyright Reform', *Vanderbilt Journal of Entertainment & Technology Law, Vol. 22, No. 2*, 2020, pp. 323–358.
- Broniatowski, D.A., et. al,, 'Facebook's Architecture Undermines Vaccine Misinformation Removal Efforts', arXiv:2202.02172 [cs.SI], 2022
- Brown M. et. al. 'Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users,' May 11, 2022, p. 5. Available at https://www.brookings.edu/research/echo-chambers-rabbit-holes-and-ideological-bias-how-youtube-recommends-content-to-real-users/.
- Bruns, A., S. Harrington, and E. Hurcombe, 'Corona? 5G? Or Both?': The Dynamics of COVID-19/5G Conspiracy Theories on Facebook', *Media International Australia, Vol. 177, No. 1*, November 2020, pp. 12–29.
- Buerger, Cathy, 'Speech as a Driver of Intergroup Violence: A Literature Review', Dangerous Speech Project, June 16, 2021.

- Butler, A., and A. Parrella, 'Tweeting with Consideration', Twitter, 2021.
- Caplan, R., 'Content or Context Moderation?', Data & Society, November 14, 2018.
- Cappello M. (ed.), 'New actors and risks in online advertising,' IRIS Special 2022-1, European Audiovisual Observatory, Strasbourg, 2022.
- Cavaliere, P., 'The Truth in Fake News: How Disinformation Laws Are Reframing the Concepts of Truth and Accuracy on Digital Platforms', *European Convention on Human Rights Law Review, Vol. 3, No. 4, October 11*, 2022, pp. 481–523.
- Centre for Media Pluralism and Media Freedom, 'Monitoring media pluralism in the digital era : application of the Media Pluralism Monitor in the European Union, Albania, Montenegro, the Republic of North Macedonia, Serbia and Turkey in the year 2021,' Media Pluralism Monitor, European University Institute, San Domenico di Fiesole, 2022.
- Centre for Media Pluralism, 'Media Pluralism Monitor 2022 at 'MPM2022 Results', Centre for Media Pluralism and Freedom, n.d., 2022, https://cmpf.eui.eu/mpm2022-results/.
- Chang, B., 'From Internet Referral Units to International Agreements: Censorship of the Internet by the UK and EU', Columbia Human Rights Law Review, NY, October 31, 2017.
- Chemaly, S., 'Demographics, Design, and Free Speech: How Demographics Have Produced Social Media Optimized for Abuse and the Silencing of Marginalized Voices', by S. Chemaly, Free Speech in the Digital Age, Oxford University Press, 2019, pp. 150–169.
- Chowdhury N., 'Automated Content Moderation: A Primer', Program on Platform Regulation, Stanford, 2022.
- Clapham, A. '*Human Rights: A Very Short Introduction (2nd Edn.)',* Oxford, Oxford University Press, 2015.
- Cohen, J.E., 'How (Not) to Write a Privacy Law', Knight First Amendment Institute at Columbia University, March 23, 2021.
- Competition & Markets Authority, 'Online platforms and digital advertising: Market study final report', Competition & Markets Authority, London, 2019, p. 220.
- Copland, S., 'Reddit Quarantined: Can Changing Platform Affordances Reduce Hateful Material Online?', *Internet Policy Review, Vol. 9, No. 4*, 2020.
- Costanza-Chock, S., Design Justice: Community-Led Practices to Build the Worlds We Need, Information Policy, The MIT Press, Cambridge, Massachusetts, 2020.
- Crawford, K., and T. Gillespie, 'What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint', *New Media & Society, Vol. 18, No. 3,* March 2016, pp. 410–428.
- Crenshaw, Kimberle, 'Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics', *The University of Chicago Legal Forum*, 1989, p. 139-167.
- Cunningham, S., and D. Craig, 'Creator Governance in Social Media Entertainment', Social Media + Society, Vol. 5, No. 4, October 2019.

- Dahlgren, P.,. Media and political engagement. Citizens, communication and democracy. Cambridge: Cambridge University Press, 2009.
- Darius, P., and M. Urquhart, 'Disinformed Social Movements: A Large-Scale Mapping of Conspiracy Narratives as Online Harms during the COVID-19 Pandemic', *Online Social Networks and Media, Vol. 26, November 2021*, p. 100174.
- Davenport, T. and Beck, J., The Attention Economy: Understanding the New Currency of Business. Cambridge: MA: Harvard Business School Press, 2001.
- Habermas J., *Between facts and norms: Contributions to a discourse theory of law and democracy*, Cambridge: Polity Press, 1998.
- De Cock Bunig, M., 'A Multi-Dimensional Approach to Disinformation: Report of the Independent High Level Group on Fake News and Online Disinformation,' Publications Office of the European Union, 2018.
- De Gregorio, G., *Digital Constitutionalism in Europe. Reframing Rights and Powers in the Algorithmic Society.*, Cambridge University Press, Cambridge, U.K., 2022.
- De Streel , A. et al. '*Online Platforms' Moderation of Illegal Content Online: Law, Practices and Options for Reform*,' Policy Department for Economic, Scientific and Quality of Life Policies, June 2020.
- Dewey, J.. *The public and its problems*, New York, NY, Swallow Press, 1954.
- Dobber T., et. al., 'The regulation of online political micro-targeting in Europe', *Internet Policy Review,* Vol. 8., Issue 4., 2019.
- Douek, E., 'Content Moderation as Systems Thinking', *Harvard Law Review, Vol. 136, No. 2*, December 2022, pp. 526–607.
- Douek, E., 'Governing Online Speech: From 'Posts-as-Trumps' to Proportionality and Probability', *Columbia Law Review, Vol. 121, No. 3,* April 2021, pp. 759–833.
- Douek, E., *'The Rise of Content Cartels', Essays and Scholarships, Knight First Amendment Institute at Columbia University, February 11, 2020.* http://knightcolumbia.org/content/the-rise-of-content-cartels.
- Dusollier, S., 'The 2019 Directive on Copyright in the Digital Single Market: Some Progress, a Few Bad Choices, and an Overall Failed Ambition', *Common Market Law Review, Vol. 57, No. Issue 4*, August 1, 2020, pp. 979–1030.
- Dvoskin, B., *'Expert Governance of Online Speech', Harvard International Law Journal, Forthcoming,* July 28, 2022.
- Eisenstein, E., *The Printing Revolution in Early Modern Europe*, Cambridge University Press, Cambridge, U.K.
- Etling, B., et. al., 'Blogs as an alternative public sphere: The role of blogs, mainstream media, and TV in Russia's media ecology', *Berkman Center Research Publication No. 8*, Cambridge U.S., 2014.
- Eurobarometer, Media & News Survey 2022, n.d. July 2022, p. 30 https://europa.eu/eurobarometer/surveys/detail/2832.
- Fabbrini, F., *Fundamental Rights in Europe*, Oxford, Oxford University Press, Oxford, 2014.
- Farid, H., 'Creating, Using, Misusing, and Detecting Deep Fakes', *Journal of Online Trust and Safety, Vol. 1, No. 4,* September 20, 2022.

- Franks, M. A., 'Beyond the Public Square: Imagining Digital Democracy', *The Yale Law Journal, Vol. 131, 202*1-2022.
- Freedman, P.,et al., 'Campaign Advertising and Democratic Citizenship', *American Journal of Political Science*, Vol. 48, No. 4, October 2004, pp. 723–74.
- Freelon, D., and C. Wells, 'Disinformation as Political Communication', *Political Communication, Vol. 37, No. 2*, March 3, 2020, pp. 145–156.
- Fuller, L., *The Morality of Law: Revised Edition*, Yale University Press, New Haven, U.S., 1969.
- Gillespie, T., 'Content Moderation, AI, and the Question of Scale', *Big Data & Society, Vol. 7, No. 2*, July 2020.
- Gillespie, T., 'Do Not Recommend? Reduction as a Form of Content Moderation', *Social Media + Society, Vol. 8, No. 3,* July 2022.
- Gillespie, T., *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*, Yale University Press, 2018.
- Griffin, R., 'New School Speech Regulation as a Regulatory Strategy against Hate Speech on Social Media: The Case of Germany's NetzDG', *Telecommunications Policy, Vol. 46, No. 9,* October 2022.
- Griffin, R., 'Public and Private Power in Social Media Governance: Multistakeholderism, the Rule of Law and Democratic Accountability', SSRN Electronic Journal, 2022.
- Griffin, R., 'Rethinking Rights in Social Media Governance: Human Rights, Ideology and Inequality', *Forthcoming in European Law Open, Vol. 2, No. 1,* March 23, 2022.
- Griffin, R., 'The Sanitised Platform', *JIPITEC*, Vol. 3, No.1, 2022, https://www.jipitec.eu/issues/jipitec-13-1-2022/5514/citation.
- Grison, T., and V. Julliard, 'Les Enjeux de La Modération Automatisée Sur Les Réseaux Sociaux Numériques : Les Mobilisations LGBT Contre La Loi Avia', *Communication, Technologies et Développement, No. 10*, May 20, 2021.
- Habermas, J. The structural transformation of the public sphere: An inquiry into a category of bourgeois society. Cambridge, U.K.,: Polity, 1989.
- Habermas, J., The Structural Transformation of the Public Sphere, Polity, 1962 (trans 1989).
- Haimson, O.L., D. Delmonaco, P. Nie, and A. Wegner, 'Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas', P*roceedings of the ACM on Human-Computer Interaction, Vol. 5, No. CSCW2*, October 13, 2021, pp. 1–35.
- Hameleers, M., T.E. Powell, T.G.L.A. Van Der Meer, and L. Bos, 'A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated via Social Media', *Political Communication*, Vol. 37, No. 2, March 3, 2020, pp. 281–301.
- Hare, I., and Weinstein, J. eds., *Extreme Speech and Democracy, 1st ed.*, Oxford University Press, Oxford, 2009.
- Harmer, E. and Lumsden Karen, eds., *Online Othering: Exploring Digital Violence and Discrimination on the Web, 1st ed.*, Palgrave Studies in Cybercrime and Cybersecurity, Springer International Publishing: Imprint: Palgrave Macmillan, Cham, 2019.

- Helberger, N., 'The Political Power of Platforms: How Current Attempts to Regulate Misinformation Amplify Opinion Power', Digital Journalism, Vol. 8, No. 6, 2020, pp. 842–854.
- Heldt, A., 'Reading between the Lines and the Numbers: An Analysis of the First NetzDG Reports', Internet Policy Review, Vol. 8, No. 2, June 12, 2019.
- Hillygus, D.S., 'Campaign Effects and the Dynamics of Turnout Intention in Election 2000', The Journal of Politics, Vol. 67, No. 1, February 2005, pp. 50–68.
- Hoboken, J., and R.Ó. Fathaigh, 'Regulating Disinformation in Europe: Implications for Speech and Privacy', UC Irvine Journal of International, Transnational, and Comparative Law, Vol. 6, No. 1, 2021.
- Humphreys, P., 'Germany's 'Dual' Broadcasting System: Recipe for Pluralism in the Age of Multi-Channel Broadcasting?', *New German Critique, No. 78*, 1999.
- Husovec, M. and Roche Laguna, I., '*Digital Services Act: A Short Primer', Principles of the Digital Services Act*, Oxford, Oxford University, Forthcoming 2023.
- Husovec, M., '(Ir)Responsible Legislature? Speech Risks under the EU's Rules on Delegated Digital Enforcement', Available at SSRN. September 17, 2021.
- Huszár F. et. al. 'Algorithmic amplification of politics on Twitter,' *Proceedings of the National Academy of Science Vol. 119, No. 1, 2021,* January 5, 2022.
- Iacob, O., 'Hate Crime and Hate Speech in Europe: Comprehensive Analysis of International Law Principles, EU-Wide Study and National Assessments', *European Website on Integration*, 2015.
- IHS Markit, 'The Economic Value of Behavioural Targeting in Digital Advertising', Data Driven advertising n.d. Integrity Institute, 'Ranking and Design Transparency', 2021.
- Iosifidis, P., 'The public sphere, social networks and public service media,' *Information, Communication & Society, Vol. 14 Issue 5,* 2011, p. 619–637.
- Ipsos European Public Affairs. 'News and Media Survey 2022', European Parliament, Strasbourg, 2022., p.12, 17-18, https://europa.eu/eurobarometer/surveys/detail/2832.
- Irion, K. et.al., 'Introductory chapter. Outlining the value of safeguarding media pluralism and diversity to Member States, the EU and the relevant competences', Study on Media Plurality and Diversity Online (Centre for Media Pluralism and Media Freedom et. al.), 2022, https://cadmus.eui.eu/bitstream/handle/1814/75099/Study_on_media_plurality_and_diversity_online-KK0722202ENN.pdf?sequence=1.
- ISBA, 'Programmatic Supply Chain Transparency Study', ISBA, May 6, 2020. https://www.isba.org.uk/knowledge/executive-summary-programmatic-supply-chain-transparency-study.
- Jaursch, 'Strengthening EU proposals on deceptive platform design', Stiftung Neue Verantwortung - Policy Briefs, 2022.
- Jaursch, J., 'Overview of DSA Delegated Acts, Reports and Codes of Conduct', Stiftung Neue Verantwortung, September 12, 2022.
- Joshi, P., S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, 'The State and Fate of Linguistic Diversity and Inclusion in the NLP World', *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 6282–6293*, https://arxiv.org/pdf/2004.09095.pdf.

- Jourová V., 'Speech "From pandemic to infodemic"', European Commission, Brussels, 2020.
- Kaiser B. et. al., 'Adapting Security Warnings to Counter Online Disinformation,' USENIX Security '21, 2021.
- Kampourakis, I. et. al., 'Reappropriating the Rule of Law: Between Constituting and Limiting Private Power', Jurisprudence 2022, n.d.
- Katsaros, M., et. al., 'Reconsidering Tweets: Intervening During Tweet Creation Decreases Offensive Content', International AAAI Conference on Web and Social Media, 2022.
- Keller, D., 'Facebook Filters, Fundamental Rights, and the CJEU's Glawischnig-Piesczek Ruling', *GRUR International, Vol. 69, No. 6, June 1*, 2020, pp. 616–623.
- Keller, D., 'Policing Online Comments in Europe: New Human Rights Case Law in the Real World', The Center for Internet and Society, Stanford Law School, 2016.
- Keller, I. C., 'Don't Shoot the Message: Regulating Disinformation Beyond Content', Direito Público, Vol. 18, No. 99, 2021, pp. 486–515.
- Keller, T., T. Graham, D. Angus, A. Bruns, R. Nijmeijer, K.L. Nielbo, A. Bechmann, et al., ''Coordinated Inauthentic Behavior' and other Online Influence Operations in Social Media' AoIR Selected Papers of Internet Research, October 5, 2020.
- Kleis Nielsen, R. et. al, 'Dealing with Digital Intermediaries: A Case Study of the Relations between Publishers and Platforms', New Media & Society, Vol. 20, No. 4, April 2018, pp. 1600–1617.
- Kleis Nielsen, R., and. Levy D., 'The Changing Business of Journalism and its Implications for Democracy,' The Reuters Institute for the Study of Journalism, Department of Politics and International Relation, Oxford, 2010.
- Klompmaker, N., 'Censor Them at Any Cost? A Social and Legal Assessment of Enhanced Action Against Terrorist Content Online', *Amsterdam Law Forum, Vol. 11, No. 3*, June 1, 2019.
- Klonick K., 'The New Governors: The People, Rules, and Processes Governing Online Speech', *Harvard Law Review, vol. 131, no. 6*, 2018.
- Klonick, K., 'The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression', *The Yale Law Journal, Vol. 129, No. 8*, 2020, pp. 2418–2499.
- Krasno, J.S., and D.P. Green, 'Do Televised Presidential Ads Increase Voter Turnout? Evidence from a Natural Experiment', *The Journal of Politics, Vol. 70, No.* 1, January 2008, pp. 245–26.
- Krishnan, N., et. al., 'Research Note: Examining How Various Social Media Platforms Have Responded to COVID-19 Misinformation', *Harvard Kennedy School Misinformation Review*, December 15, 2021.
- Kuczerawy, A., 'From 'Notice and Take Down' to 'Notice and Stay Down': Risks and Safeguards for Freedom of Expression', *The Oxford Handbook of Intermediary Liability Online*, 2019.
- Kumar, S., 'The Algorithmic Dance: YouTube's Adpocalypse and the Gatekeeping of Cultural Content on Digital Platforms', *Internet Policy Review, Vol. 8, No. 2*, June 30, 2019.
- Kushin, M.J., and K. Kitchener, 'Getting Political on Social Network Sites: Exploring Online Political Discourse on Facebook', *First Monday 14(11)*, 2009.
- Kyriakidou, M., S. Cushion, C. Hughes, and M. Morani, 'Questioning Fact-Checking in the Fight Against Disinformation: An Audience Perspective', *Journalism Practice*, July 7, 2022, pp. 1–17.

- Land, M.K., 'Against Privatized Censorship: Proposals for Responsible Delegation', *Virginia Journal of International Law*, 2019.
- Laux, J., et. al., 'Taming the few: Platforms regulation, independent audits, and the risks of capture created by the DMA and DSA,' *Computer Law & Security Review,* Vol. 43, 2021, https://doi.org/10.1016/j.clsr.2021.105613.
- Leerssen, P., 'The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems', *European Journal of Law and Technology, Vol. 11, No. 2*, February 24, 2020.
- Leerssen, P., et al., 'News from the Ad Archive: How Journalists Use the Facebook Ad Library to Hold Online Advertising Accountable', Information, Communication & Society, December 26, 2021, pp. 1–20.
- Lenhart, A., M. Ybarra, K. Zickuhr, and M. Price-Feeney, 'Online Harassment, Digital Abuse, and Cyberstalking in America', Data & Society, November 21, 2016.
- Leonard Cheshire, 'Online Disability Hate Crimes Soar 33%', Leonard Cheshire, 2019.
- Leslie D., et. al. Artificial Intelligence, 'Human Rights, Democracy, and The Rule of Law. A Primer',The Council of Europe, 2021.
- Liu, Z., et al., 'Monolith: Real Time Recommender system With Collisionless Embedding Table', arXiv, September 27, 2022.
- Llansó, E., et. al. 'Artificial Intelligence, Content Moderation, and Freedom of Expression', Transatlantic Working Group, February 26, 2020.
- Lorenz-Spreen, P., L. Oswald, S. Lewandowsky, and R. Hertwig, 'A Systematic Review of Worldwide Causal and Correlational Evidence on Digital Media and Democracy', Nature Human Behaviour, November 7, 2022.
- M. Thompson, 'The Challenging New Economics of Journalism.', Reuters Institute Digital News Report 2016, Reuters Institute for the Study of Journalism, Department of Politics and International Relations, University of Oxford, 2016.
- Mahieu, R., J. van Hoboken, and H. Asghari, 'On the Question of the Controller, "Effective and Complete Protection" and Its Application to Data Access Rights in Europe', *Responsibility for Data Protection in a Networked World, Vol. 10, No. 85*, 2019.
- Marsden, C., T. Meyer, and I. Brown, 'Platform Values and Democratic Elections: How Can the Law Regulate Digital Disinformation?', *Computer Law & Security Review, Vol. 36*, April 2020, p. 105373.
- Marwick, A.E., 'Morally Motivated Networked Harassment as Normative Reinforcement', *Social Media + Society, Vol. 7, No. 2*, April 2021, p. 205630512110213.
- Matamoros-Fernández, A. and Farkas, J., 'Racism, Hate Speech, and Social Media: A Systematic Review and Critique,' Television & New Media Vol. 22, Issue 2 https://doi.org/10.1177/1527476420982230.
- Matsuda, Mari J., ed., Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment, New Perspectives on Law, Culture, and Society, Westview Press, Boulder, Colo, 1993.
- Minow, M., *Saving the News: Why the Constitution Calls for Government Action to Preserve Freedom of Speech,* Oxford, Oxford University Press, 2021.
- Monea, A., Digital Closet: How The Internet Became Straight, Mit Press, S.L., 2023.

- Mühlhoff, R., and T. Willem, 'Social Media Advertising for Clinical Studies: Ethical and Data Protection Implications of Online Targeting', Pre-Print, accepted/in press, *Big Data & Society*., April 2022.
- Müller, J. 'Liberal Democracy's Critical Infrastructure. How to think about Intermediary Powers', *Scripts Working Paper No. 16*. Berlin, Scripts-Berlin, Nov. 16, 2022.
- Müller, K., and C. Schwarz, 'Fanning the Flames of Hate: Social Media and Hate Crime', SSRN Scholarly Paper, Rochester, NY, June 5, 2020.
- Munger, K., and J. Phillips, 'Right-Wing YouTube: A Supply and Demand Perspective', *The International Journal of Press/Politics, Vol. 27, No. 1*, 2022, pp. 186–219.
- Napoli, P.M., and R. Caplan, 'When Media Companies Insist They're Not Media Companies and Why It Matters for Communications Policy', SSRN Scholarly Paper, Rochester, NY, March 18, 2016.
- Narayanan, A., 'Is There a Filter Bubble on Social Media? A Call for Epistemic Humility', Department of Computer Science, Princeton University', Princeton University Media Central, Princeton , 2021.
- Nenadić, I., 'Unpacking the "European approach" to tackling challenges of disinformation and political manipulation,' *Internet Policy Review, Vol. 8. Issue 4*., 2019,
- Newman, N., How Publishers are Learning to Create and Distribute News on TikTok, Reuters Institute for the Study of Journalism, University of Oxford, 2022.
- Nilsen J., Fagan K., Dreyfuss E., and Donovan J., 'Tiktok, the war on Ukraine, and 10 features that make the app vulnerable to misinformation', The media manipulation casebook, 2022.
- Nyhan, B., and J. Reifler, 'The Effect of Fact-Checking on Elites: A Field Experiment on U.S. State Legislators: The Effect of Fact-Checking Elites', *American Journal of Political Science, Vol. 59, No. 3*, July 2015, pp. 628–640.
- Ó Fathaigh, R., et. al. 'The Perils of Legally Defining Disinformation', *Internet Policy Review, Vol. 10, No. 4,* November 4, 2021.
- Open Markets Institute, 'America's Free Press and Monopoly: The Historical Role of Competition Policy in Protecting Independent Journalism in America', Open Markets Institute, n.d., Washington D.C., 2018.
- Oruç, T.H., 'The Prohibition of General Monitoring Obligation for Video-Sharing Platforms under Article 15 of the E-Commerce Directive in Light of Recent Developments: Is It Still Necessary to Maintain It?', *Journal of Intellectual Property, Information Technology and E-Commerce Law, Vol. 13, No. 3*, 2022, pp. 176–199.
- Papaevangelou, C., 'Funding Intermediaries: Google and Facebook's Strategy to Capture Journalism', Digital Journalism, January 13, 2023, pp. 1–22.
- Papaevangelou, C., and N. Smyrnaios, 'Regulating Dependency: The Political Stakes of Online Platforms' Deals with French Publishers', HAL, August 2022.
- Pariser, E., The Filter Bubble: What the Internet Is Hiding from You, Penguin books, London, 2012.
- Park, J.H., J. Shin, and P. Fung, 'Reducing Gender Bias in Abusive Language Detection', Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2799–2804.
- Petre, C., All the News That's Fit to Click, Princeton University Press, New Jersey, 2021, https://press.princeton.edu/books/hardcover/9780691177649/all-the-news-thats-fit-to-click.

- Pickard, V., and A.T. Williams, 'Salvation Or Folly?', *Digital Journalism, Vol. 2, No. 2*, April 3, 2014, pp. 195–213.
- Pickard, V., *Democracy without Journalism?: Confronting the Misinformation Society, 1st ed.*, Oxford University Press, 2020.
- Polletta, F., and J. Callahan, 'Deep Stories, Nostalgia Narratives, and Fake News: Storytelling in the Trump Era', *American Journal of Cultural Sociology, Vol. 5, No. 3*, October 2017, pp. 392–408.
- Raeijmaekers, D. and Maeseele, P., 'Media pluralism and democracy: what's in a name?', *Media, Culture & Society, Vol. 37, Issue 7*, 2015.
- Ranaivoson, H. et. al., 'Chapter B1. Mapping of the measures and data gathering methods concerning the concentration of economic resources to ensure media plurality', European Commission Study on Media Plurality and Diversity Online, Publications Office of the European Union, 2022.
- Raz J., *The authority of law: Essays on law and morality*, Oxford University Press, Oxford, U.K., 1979.
- Richards, N., *Why Privacy Matters*, Oxford, Oxford University Press, 2021.
- Ricolfi, M. et. al., 'Academics against Press Publishers' Right: 169 European Academics warn against it,' n.d.
- Riemer, K., and S. Peter, 'Algorithmic Audiencing: Why We Need to Rethink Free Speech on Social Media', *Journal of Information Technology, Vol. 36, No. 4*, December 2021, pp. 409–426.
- Ritholtz, S., 'Fanning the Flames of Hate: The Transnational Diffusion of Online Anti-LGBT+ Rhetoric and Offline Mobilisation', GNET, n.d.
- Roberts, S.T., *Behind the Screen: Content Moderation in the Shadows of Social Media*, Yale University Press, New Haven, 2021.
- Roozenbeek, J., et. al.'Susceptibility to Misinformation about COVID-19 around the World', *Royal Society Open Science, Vol. 7, No. 10*, October 2020, p. 201199.
- Ross, A. et. al. 'Echo Chambers, Filter Bubbles, and Polarisation: A Literature Review', Reuters Institute for the Study of Journalism, Oxford, Jan. 19, 2022.
- S. Austin, N. Newman, 'Attitudes to Advertising - Digital News Report 2015', Reuters Institute Digital News Report, Oxford, 2015., https://www.digitalnewsreport.org/essays/2015/attitudes-to-advertising/.
- "Sander, B., 'Democratic Disruption in the Age of Social Media: Between Marketized and Structural Conceptions of Human Rights Law,' *European Journal of International Law, Vol. 32, Issue 1, 2021*, p. 159–193.
- Sap, M., D. Card, S. Gabriel, Y. Choi, and N.A. Smith, 'The Risk of Racial Bias in Hate Speech Detection', Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1668–1678.
- Savin, A., *EU Internet Law, Third edition.* Elgar European Law, Edward Elgar Publishing, Cheltenham, UK ; Northampton, MA, USA, 2020.
- Scharkow, M., F. Mangold, S. Stier, and J. Breuer, 'How Social Network Sites and Other Online Intermediaries Increase Exposure to News', Proceedings of the National Academy of Sciences, Vol. 117, No. 6, February 11, 2020, pp. 2761–2763.

- Schiller, T. 'Direct Democracy', Encyclopedia Britannica, 07 Oct. 2022, https://www.britannica.com/topic/direct-democracy.
- Schreiber, M., "Verified' Anti-Vax Accounts Proliferate as Twitter Struggles to Police Content', The Guardian, November 21, 2022, sec. Technology.
- Schulz, A., 'Local news unbundled: where audience value still lies' Reuters Institute Digital News Report 2021, 10th Edition, Oxford, Reuters Institute for the Study of Journalism, June 23, 2021.
- Sellars, A., 'Defining Hate Speech', SSRN Scholarly Paper, Rochester, NY, December 1, 2016.
- Shadmy, T., 'Content Traffic Regulation: A Democratic Framework for Addressing Misinformation', *Jurimetrics, Volume 63, No. 1*, 2022.
- Siapera, E., 'Online Misogyny as Witch Hunt: Primitive Accumulation in the Age of Techno-Capitalism', in D. Ging and E. Siapera (eds.), Gender Hate Online, Springer International Publishing, Cham, 2019, pp. 21–43.
- Singhal, M., et. al., 'SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice', arXiv, October 27, 2022.
- Strand, C.,, Disinformation Campaigns about LGBTI+ People in the EU and Foreign Influence: Briefing, European Parliament, Brussels, 2021.
- Sunstein, C. R., *Republic.com.* Princeton, Princeton University Press, New Jersey, 2001.
- Suzor, N.,et. al., 'Human Rights by Design: The Responsibilities of Social Media Platforms to Address Gender-Based Violence Online: Gender-Based Violence Online', *Policy & Internet, Vol. 11, No. 1*, March 2019, pp. 84–103.
- Szakàcs J., Bognàr E., European Parliament. Directorate General for External Policies of the Union., The Impact of Disinformation Campaigns about Migrants and Minority Groups in the EU: In Depth Analysis., Publications Office, Brussels, 2021.
- Talat, Z., '"It ain't all good:" Machinic abuse detection and marginalisation in machine learning', PhD thesis, University of Sheffield, 2021.
- Tappin, B.M., et al., 'Quantifying the Persuasive Returns to Political Microtargeting', PsyArXiv Preprints, 2022. https://psyarxiv.com/dhg6k/.
- The Economic Value of Behavioural Targeting in Digital Advertising, IHS Markit, London, 2017, p. 2,
- Toparlak R. T., 'Criminalising Pornographic Deep Fakes: A Gender-Specific Inspection of Image-Based Sexual Abuse', Sciences Po Chair Numerique, Paris, 2022.
- Törnberg, P., 'How Digital Media Drive Affective Polarization through Partisan Sorting', *Proceedings of the National Academy of Sciences, Vol. 119, No. 42*, October 18, 2022, p. E2207159119.
- Tucker, J., A. Guess, P. Barbera, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan, 'Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature', SSRN Electronic Journal, 2018.
- Tufeczi, Z., 'Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodologies' ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media, 2014.
- Urban, J.M., et. al., 'Notice and Takedown in Everyday Practice', *UC Berkeley Public Law, Research Paper No. 2755628*, March 22, 2017, pp. 1–182.

- Vaccari, C., and A. Chadwick, 'Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News', *Social Media + Society, Vol. 6, No. 1*, January 2020, p. 205630512090340.
- Vaidhyanathan, S., *Antisocial Media: How Facebook Disconnects US and Undermines Democracy*, Oxford University Press, New York, NY, United States of America, 2018.
- Van Drunen M. et. al., New Actors and Risks in Online Advertising, European Audiovisual Authority, Strasbourg, 2022.
- Vander Maelen, C. "Hardly Law or Hard Law? Investigating the Dimensions of Functionality and Legalisation of Codes of Conduct in Recent EU Legislation and the Normative Repercussions Thereof." *European Law Review, vol. 47, no. 6*, Sweet & Maxwell, 2022, pp. 752–72.
- Vermeulen, M., 'Researcher Access to Platform Data: European Developments', *Journal of Online Trust and Safety, Vol. 1, No. 4*, September 20, 2022.
- Vickery, Jacqueline Ryan, and Tracy Everbach, eds., *Mediating Misogyny: Gender, Technology, and Harassment,* Softcover reprint of the hardcover 1st edition 2018., Palgrave Macmillan, Cham, Switzerland, 2019.
- Waldron, J. 'Is the Rule of Law a Essentially Contested Concept (In Florida)?' *Vol. 21, No. 2, In the Wake of Bush v. Gore: Law, Legitimacy and Judicial Ethics*, 2002.
- Wilman , F., 'The EU's System of Knowledge-Based Liability for Hosting Service Providers in Respect of Illegal User Content – between the e-Commerce Directive and the Digital Services Act', *JIPITEC Vol. 12, 2021*, pp. 317–341.
- Winner L., 'Do Artifacts have politics?', *Daedalus, Vol. 109, No. 1*, 1980.
- Woolley S., 'In Many Democracies, Disinformation Targets the Most Vulnerable', Centre for International Governance Innovation, 2022.
- Wunsch-Vicent, S., The Evolution of News and the Internet, Organisation for Economic Co-operation and Development, France, 2010.
- Yardi, S., & Boyd, D., 'Dynamic debates: An analysis of group polarization over time on twitter.' *Bulletin of Science, Technology & Society, 30(5)*, 2010, p 316–327.
- Zeng, J., and D.B.V. Kaye, 'From Content Moderation to Visibility Moderation: A Case Study of Platform Governance on TikTok', *Policy & Internet, Vol. 14, No. 1*, March 2022, pp. 79–95.
- Zimmermann, F., and M. Kohring, 'Mistrust, Disinforming News, and Vote Choice: A Panel Survey on the Origins and Consequences of Believing Disinformation in the 2017 German Parliamentary Election', *Political Communication, Vol. 37, No. 2*, March 3, 2020, pp. 215–237.
- Zuboff, S., *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, PublicAffairs, 2019.
- Zuiderveen Borgesius, F.J., D. Trilling, J. Möller, B. Bodó, C.H. de Vreese, and N. Helberger, 'Should We Worry about Filter Bubbles?', *Internet Policy Review, Vol. 5, No. 1, March 31*, 2016.

Other documents

- 'The Digital Youth Index - Understand the Impact of Technology on Young People', Digital Youth Index, 2022. https://digitalyouthindex.uk/.
- 'Who we are', Center for Humane Technology, n.d. https://www.humanetech.com/who-we-are#team.
- Alexander, J., 'YouTube Moderation Bots Punish Videos Tagged as "Gay" or "Lesbian," Study Finds', The Verge, September 30, 2019.
- Amnesty International, 'Amnesty Reveals Alarming Impact of Online Abuse against Women', *Amnesty International,* November 20, 2017.
- Article 19, 'CJEU Judgment in Facebook Ireland Case Is Threat to Online Free Speech', Article 19, October 3, 2019. https://www.article19.org/resources/cjeu-judgment-in-facebook-ireland-case-is-threat-to-online-free-speech/.
- Bernstein, J., 'Bad News: Selling the Story of Disinformation', *Harper's Magazine, Vol. September 2021, August 9*, 2021.
- Bobrowsky, M., 'Elon Musk Champions Twitter Fact-Checking Feature That Corrects Him', WSJ, 2022. https://www.wsj.com/articles/elon-musk-champions-twitter-fact-checking-feature-that-corrects-him-11669436937.
- Boullier, D., 'Comment lutter contre le réchauffement médiatique', *Usbek&Rica*, 2019.
- Boyd D,, 'You Think You Want Media Literacy… Do You?', *Data and Society: Points*, 2018.
- Danbury, R., 'The DSM Copyright Directive: Article 15: What? Part II', Kluwer Copyright Blog, April 29, 2021.
- Davy j., 'Amicus Brief for Gonzalez v Google', Integrity Institute, 2022 https://integrityinstitute.org/amicus-brief-for-gonzalez-v-google.
- Debre, I., and F. Akram, 'Facebook While Black: Users Call It Getting 'Zucked,' Say Talking about Racism Is Censored as Hate Speech', USA TODAY, 2021. https://www.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/.
- DiResta, R., 'The Supply of Disinformation Will Soon Be Infinite', The Atlantic, September 20, 2020. https://www.theatlantic.com/ideas/archive/2020/09/future-propaganda-will-be-computer-generated/616400/.
- Dixit, C.S., Ryan Mac, Pranav, ''I Have Blood On My Hands': A Whistleblower Says Facebook Ignored Global Political Manipulation', BuzzFeed News, 2020 https://www.buzzfeednews.com/article/craigsilverman/facebook-ignore-political-manipulation-whistleblower-memo.
- Douek, E., 'What Does "Coordinated Inauthentic Behavior" Actually Mean?', Slate, July 2, 2020.
- Drügemöller, L., 'Pimmel-Gate in Hamburg: Unterhalb der Schwelle', Die Tageszeitung, August 8, 2022, sec. TAZ, Nord.
- Dwoskin E., Menn J., and Zakrzewski C.,'Twitter can't afford to be one of the world's most influential websites', Washington Post, 4 September 2022.

- Eckert, S., C. Felke, and O. Vitlif, 'TikTok schränkt mit Wortfiltern Meinungsfreiheit ein', tagesschau.de, 2022.
- Ghaffary, S., 'Instagram's Surprising Strategy for Bullies: Tell Them to Be Nice', Vox, October 20, 2022.
- Griffin, R., 'Tackling Discrimination in Targeted Advertising: US regulators take very small steps in the right direction – but where is the EU?', Verfassungsblog, June 23, 2022.
- Guynn, J., 'Facebook Still Has Holocaust Denial Content Three Months after Mark Zuckerberg Pledged to Remove It', USA TODAY, Jan. 27, 2021.
- Hagood M., 'Emotional Rescue', *Real Life Magazine*, Dec. 30, 2021.
- Hao K., 'How Facebook got addicted to spreading misinformation', MIT Technology Review, 11 March 2021.
- Horwitz, J.S., Newley Purnell and Jeff, 'Facebook Employees Flag Drug Cartels and Human Traffickers. The Company's Response Is Weak, Documents Show.', Wall Street Journal, September 16, 2021, sec. Tech.
- Hsu, T., 'Worries Grow That TikTok Is New Home for Manipulated Video and Photos', The New York Times, November 4, 2022, sec. Technology.
- Joseph, A.M., et. al., 'COVID-19 Misinformation on Social Media: A Scoping Review', Cureus, April 29, 2022.
- Jütte , B.J., and G. Priora, 'On the Necessity of Filtering Online Content and Its Limitations', Kluwer Copyright Blog, July 20, 2021.
- Kaiser B., et. al., 'Warnings That Work: Combating Misinformation Without Deplatforming', Lawfare, Friday, July 23, 2021.
- Kaiser J., and Rauchfleisch A., 'Unite the Right? How YouTube's Recommendation Algorithm Connects The U.S. Far-Right', D&S Media Manipulation: Dispatches from the Field, April 11, 2018.
- Kayser-Bril, N., 'Automated Moderation Tool from Google Rates People of Color and Gays as "Toxic"', AlgorithmWatch, 2020.
- Kayser-Bril, N., 'Facebook's Moderation Is Wreaking Havoc in Lithuanian Public Discourse', Algorithm Watch, 2022.
- Keller D., 'The DSA's Industrial Model for Content Moderation', Verfassungsblog, 2022.
- Keller, D., 'Daphne Keller and ACLU File Comment to Meta Oversight Board in 'UK Drill'
- Music' Case', ACLU, August 23, 2022.
- Kellner, D., 'The EU's new Digital Services Act and the Rest of the World', Verfassungsblog, November 7, 2022.
- Killen, M., 'Germany supports ban on personal data for political ads', Euroactive, Sept. 7, 2022.
- Knight, W., 'Elon Musk Has Fired Twitter's 'Ethical AI' Team', Wired, 2022.
- Knight, W., 'Here's Proof Hate Speech Is More Viral on Elon Musk's Twitter', Wired, 2022.
- Lapowsky, I., 'Jeff Allen, Sahar Massachi Launch Integrity Institute', *Protocol,* 2021 https://www.protocol.com/policy/integrity-institute.

- Leerssen, P., 'Platform research access in Article 31 of the Digital Services Act: Sword without a shield?', Verfassungsblog, September 7, 2021.
- Levy, S., 'Inside Meta's Oversight Board: 2 Years of Pushing Limits', Wired, November 8, 2022.
- Lomas, N., 'Meta Reports Takedowns of Influence Ops Targeting US Midterms, Ukraine War', TechCrunch, September 27, 2022.
- Lux, D., and Lil Miss Hot Mess, 'Facebook's Hate Speech Policies Censor Marginalized Users', Wired, 2017.
- Ma O., and Feldman B., 'How Google and YouTube are investing in fact-checking', Google News Initiative, 2022.
- Mantas, H., 'Twitter Finally Turns to the Experts on Fact-Checking', Poynter, August 5, 2021.
- Massachi, S., 'How to Save Our Social Media by Treating It like a City', MIT Technology Review, December 20, 2021.
- McGee, P., 'Meta and Alphabet Lose Dominance over US Digital Ads Market', Financial Times, Financial Times, San Francisco, 23 Dec. 2022.
- Merril, J., and Oremus, W., 'Five Points for Anger, One for a 'Like': How Facebook's Formula Fostered Rage and Misinformation', The Washington Post, Oct. 26, 2021, https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/.
- "Merrill J. B., and Oremus W., 'Five points for anger, one for a 'like': How Facebook's formula fostered rage and misinformation', Washington Post, 2021.
- Meta, ' Suicide and Self Injury', Transparency Center, n.d. https://transparency.fb.com/es-es/policies/community-standards/suicide-self-injury/.
- Meta, 'Coordinated Inauthentic Behavior', Meta, n.d.
- Milman, O., '#ClimateScam: Denialism Claims Flooding Twitter Have Scientists Worried', The Guardian, December 2, 2022, sec. Technology.
- Misinformation Amplification Analysis and Tracking Dashboard', Integrity Institute, 2022.
- Morris L., and Oremus W. 'Russian Disinformation Is Demonizing Ukrainian Refugees', Washington Post, December 8, 2022.
- Newton C., and Schiffer Z., 'Twitter, cut in half', Platformer News, 2022.
- Newton, C., 'Facebook Will Pay $52 Million in Settlement with Moderators Who Developed PTSD on the Job', The Verge, May 12, 2020.
- Newton, C., 'The Secret Lives of Facebook Moderators in America', The Verge, February 25, 2019.
- Newton, C., 'Three Facebook Moderators Break Their NDAs to Expose a Company in Crisis', The Verge, June 19, 2019.
- Nimmo B., and Torrey M., 'Taking down coordinated inauthentic behavior from Russia and China', Meta, 2022.
- Oltermann, P., 'Tough New German Law Puts Tech Firms and Free Speech in Spotlight', The Guardian, January 5, 2018, sec. World news.

- Ortolani, P., 'If You Build It, They Will Come: The DSA's "Procedure Before Substance" Approach', Verfassungsblog, November 7, 2022.
- Oversight Board, 'Miembros del Consejo', Consejo asesor de contenido, n.d. https://www.oversightboard.com/meet-the-board/.
- Owen, L.H., 'Facebook's pivot to video didn't just burn publishers. It didn't even work for Facebook', Nieman Lab, September 15, 2021. https://www.niemanlab.org/2021/09/well-this-puts-a-nail-in-the-news-video-on-facebook-coffin/.
- Paul, K., 'A Brutal Year: How the 'techlash' Caught up with Facebook, Google and Amazon', The Guardian, December 28, 2019, sec. Technology.
- Pershan, C., and Sinders, C., 'Why Europe's Digital Services Act Regulators Need Design Expertise', Tech Policy Press, Dec. 22, 2022.
- Pew Rresearch, 'Twitter is much smaller than Meta and Google in terms of total user numbers and revenue, but is the dominant platform used by journalists and other public figures. See Jurkowitz, M., and J. Gottfried, 'Twitter Is the Go-to Social Media Site for U.S. Journalists, but Not for the Public', Pew Research Center, 2022.
- Radsch, C.C., 'Frenemies: Global Approaches to Rebalance the Big Tech v Journalism Relationship', Brookings, August 29, 2022.
- Rankin, J., 'Hungary Passes Law Banning LGBT Content in Schools or Kids' TV', The Guardian, June 15, 2021, sec. World news.
- Reda , F., and P. Keller, 'CJEU Upholds Article 17, but Not in the Form (Most) Member States Imagined', Kluwer Copyright Blog, CJEU, Digital Single Market, European Union, Liability, April 28, 2022.
- Reuters, 'Ex-Facebook Moderator in Kenya Sues over Working Conditions', The Guardian, May 10, 2022.
- Scheck, J.,et. al., 'Facebook Employees Flag Drug Cartels and Human Traffickers. The Company's Response Is Weak, Documents Show.', Wall Street Journal, September 16, 2021.
- Starr, J. Terrell, The Unbelievable Harassment Black Women Face Daily on Twitter, Alternet.Org, 2014.
- Segreti, G., 'Facebook CEO Says Group Will Not Become a Media Company', Reuters, August 29, 2016, sec. Internet News.
- Seufert, E., 'The App Tracking Transparency Recession', Mobiledevmemo.com, January 11, 2023.
- Silverman, R.M., Craig, 'After The US Election, Key People Are Leaving Facebook And Torching The Company In Departure Notes', BuzzFeed News, 2020. https://www.buzzfeednews.com/article/ryanmac/facebook-rules-hate-speech-employees-leaving.
- Solove, D.J., 'Privacy Self-Management and the Consent Dilemma', Harvard Law Review, November 4, 2012.
- Sottiaux, S., 'Conflicting Conceptions of Hate Speech in the ECtHR's Case Law', German Law Journal, Vol. 23, No. 9, December 2022, pp. 1193–1211.

- Spies, S., On Digital Disinformation and Democratic Myths, MediaWell, Social Science Research Council, December 10, 2019.
- Starr, P., *The Creation of the Media: Political Origins of Modern Communication*s, Basic Books, New York City, 2004.
- Stiftung Neue Verantwortung e. V. et. al, 'Open Letter: EU must protect fundamental freedoms for online political speech,' Algorithm watch, November 29, 2022.
- Team, T.N.O., 'To Serve Better Ads, We Built Our Own Data Program', Medium, December 17, 2020.
- Thomsen, I., 'Do Facebook Ads Win Elections? It's Complicated.', *Northeastern Global News*, March 8, 2022.
- Thorburn, L., et. al., 'How to Measure the Causal Effects of Recommenders', Understanding Recommenders, November 23, 2022.
- Tufekci, Z., 'We Pay an Ugly Cost for Ads on Twitter', The New York Times, November 4, 2022, sec. Opinion.
- Vincent, J., 'TikTok Sued by Former Content Moderator for Allegedly Failing to Protect Her Mental Health', The Verge, December 24, 2021.
- Vincent, J., 'Twitter is bringing its 'read before you retweet' prompt to all users / Don't tl;dr that article', The Verge, 2020.
- Zadrozny B, ''Carol's Journey': What Facebook knew about how it radicalized users', NBC News, 2021.

This study, commissioned by the European Parliament's Policy Department for Citizens' Rights and Constitutional Affairs at the request of the LIBE Committee, examines risks that contemporary social media - focusing in particular on the most widely-used platforms - present for democracy, the rule of law and fundamental rights. The study focuses on the governance of online content, provides an assessment of existing EU law and industry practices which address these risks, and evaluates potential opportunities and risks to fundamental rights and other democratic values.

PE 743.400
IP/C/LIBE/2022-093

Print ISBN 978-92-848-0385-9 | doi:10.2861/558930 | QA- 07-23-159 -EN-C
PDF ISBN 978-92-848-0386-6 | doi:10.2861/672578 | QA- 07-23-159 -EN-N