ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

in cotutela con University of Luxembourg - Université du Luxembourg

DOTTORATO DI RICERCA IN

LAW, SCIENCE AND TECHNOLOGY

Ciclo 36

Settore Concorsuale: 12/C1 - DIRITTO COSTITUZIONALE

Settore Scientifico Disciplinare: IUS/08 - DIRITTO COSTITUZIONALE

ONLINE HATE SPEECH AND INTERMEDIARY LIABILITY IN THE AGE OF ALGORITHMIC MODERATION

Presentata da: Pietro Dunn

Coordinatore Dottorato

Monica Palmirani

Supervisore

Oreste Pollicino

Supervisore

Mark Cole

Co-supervisore

Giovanni Sartor

Esame finale anno 2024

PhD-FDEF-x
The Faculty of Law, Economics and Finance

Department of Legal Studies

DISSERTATION

Defence held on 04/07/2024 in
Bologna to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG EN DROIT
and
DOTTORE DI RICERCA IN LAW, SCIENCE AND TECHNOLOGY

by

Pietro DUNN
Born on 15 November 1995 in Mondovì (Italy)

ONLINE HATE SPEECH AND INTERMEDIARY LIABILITY IN THE AGE OF ALGORITHMIC MODERATION

Dissertation defence committee

Prof. Dr. Mark D. Cole, dissertation supervisor
Full Professor, Université du Luxembourg

Prof. Dr. Oreste Pollicino, co-supervisor
Full Professor, Università Commerciale "L. Bocconi"

Prof. Dr. Giovanni Ziccardi, chair
Full Professor, Università degli Studi di Milano

Prof. Dr. Marina Castellaneta, vice-chair
Full Professor, Università degli Studi di Bari "A. Moro"

Prof. Dr. Giulio Enea Vigevani
Full Professor, Università degli Studi di Milano Bicocca

Prof. Maria Romana Allegri
Associate Professor, Sapienza Università di Roma

Online Hate Speech and Intermediary Liability in the Age of Algorithmic Moderation

TABLE OF CONTENTS

Abstract

This research aims to investigate the impact of liability-enhancing legal strategies in the context of the governance of online hate speech. Indeed, the increased reliance of the law on the role of private platforms for the purposes of moderating and removing hate speech deeply affects constitutional principles and individual fundamental rights. For instance, the enhancement of intermediary liability and responsibilities can contribute to the phenomenon of the over-removal of user content, with little regard to basic constitutional guarantees. Furthermore, research has shown that the ever-increasing use of automated systems for hate speech moderation gives rise to a whole new set of challenges and issues related to the concrete risk of errors and biased results, leading to a disproportionate removal of content produced by minority, vulnerable, or discriminated groups of people. After dealing with the question concerning the rationale(s) of hate speech regulation and arguing for an increased role for the principle of substantive equality in this regard, this work investigates the developing trends concerning the imposition of forms of intermediary liability with respect to the spread of hate speech content across the Internet, keeping a close eye on the evolving European framework. In doing so, this work also explores the relationship between platforms' content moderation practices and the promotion of fundamental rights and values – including the principle of substantive equality – especially in the light of the ever-increasing use of artificial intelligence systems for the detection and removal of hate speech. In the context of the European Union, it is held that such reflections are of utmost importance particularly following the adoption of the Digital Services Act. In this respect, the work argues for the need for a renewed code of conduct on hate speech, with a view to further protecting constitutional values and the rights of users.

Keywords: Hate Speech; Intermediary Liability; Non-Discrimination; EU; Content Moderation; Artificial Intelligence; Platform Governance; Freedom of Expression; Substantive Equality; Internet.

List of Acronyms

ACHR	American Convention on Human Rights
ACLU	American Civil Liberties Union
AG	Advocate General
AGCOM	Italian Communications Regulatory Authority
AI	Artificial Intelligence
AOL	America Online
ARCOM	*Authorité de Regulation de la Communication Audiovisuelle et Numérique*
AVMSD	Audiovisual Media Services Directive
BVerfG	*Bundesverfassungsgericht*
CAI	Committee on Artificial Intelligence (of the Council of Europe)
CDA	Communications Decency Act
CFREU	Charter of Fundamental Rights of the European Union
CJEU	Court of Justice of the European Union
CoC	Code of Conduct
CoE	Council of Europe
Cons Cons.	*Conseil Constitutionnel*
CSA	*Conseil Supérieur de l'Audiovisuel*
CSAM	Child Sexual Abuse Material
DMA	Digital Markets Act
DMCA	Digital Millennium Copyright Act
DSA	Digital Services Act
DSC	Digital Services Coordinator
DSM	Digital Single Market
EBDS	European Board for Digital Services
ECHR	Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights)
ECRI	European Commission against Racism and Intolerance
ECtHR	European Court of Human Rights
ECD	Directive 2000/31/EC (e-Commerce Directive)
EDSM	European Digital Single Market Strategy
EU	European Union
FRA	European Union Agency for Fundamental Rights
GDPR	General Data Protection Regulation
ICCPR	International Covenant on Civil and Political Rights

ICERD	International Convention on the Elimination of All Forms of Racial Discrimination
INRA	(Polish) Institute of National Remembrance Act
ISIS	Islamic State of Iraq and Syria
ISP	Internet service provider
LGBTQIA+	Lesbian, gay, bisexual, transgender, queer, intersexual, asexual, etc.
LLM	Large Language Model
LSSI	*Ley de Servicios de la Sociedad de Información y Comercio Electrónico*
NetzDG	*Netzwerkdurchsetzungsgesetz*
NGO	Non-Governmental Organization
NLP	Natural Language Processing
OB	Oversight Board
OECD	Organization for Economic Co-operation and Development
Ofcom	Office of Communications (UK)
OHWP	Online Harms White Paper
OSA	Online Safety Act
OTT	Over-the-top
POCs	People of colour
PragerU	Prager University
SCOTUS	Supreme Court of the United States
TERREG	Terrorist Content Online Regulation
TEU	Treaty on the European Union
TFEU	Treaty on the Functioning of the European Union
TGI	*Tribunal de Grande Instance*
TUSMA	*Testo Unico dei Servizi di Media Audiovisivi*
UK	United Kingdom of Great Britain and Northern Ireland
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNGPs	United Nations Guiding Principles on Business and Human Rights
URL	Uniform Resource Locator
US(A)	United States (of America)
VLOP	Very large online platform
VLOSE	Very large online search engine
VSP	Video-sharing platform

1.

Introduction

Summary: 1.1. Objectives of the research. – 1.1.1. Background of the research: old and new challenges in the fight against hate speech. – 1.1.2. Objectives and research questions. – 1.2. Notes on methodology. – 1.2.1. Material scope of the research. – 1.2.2. Territorial scope of analysis. – 1.2.3. Aspects of interdisciplinarity. – 1.3. Structure of the work – 1.3.1. Chapter 2: Setting the framework on hate speech governance. – 1.3.2. Chapter 3: Intermediary liability and hate speech in Europe. – 1.3.3. Chapter 4: Comparative perspectives. – 1.3.4. Chapter 5: Platform standards and automated moderation.

1.1. Objectives of the research

1.1.1. *Background of the research: old and new challenges in the fight against hate speech*

Hate speech regulation has long been a controversial topic for discussion, due to the inevitable repercussions that a legal response aimed at curbing the phenomenon has on freedom of expression. The debate, both in academia and politics, has been particularly prolific during the second half of the twentieth century – notably because of the many legislative reactions (domestic and international) enacted against hate speech in the wake of World Wars I and II – and has reemerged in recent years as a result of the birth of the Internet and of online platforms which, while representing extraordinary tools for the expansion of the right to freedom of expression and information, have also proven to be an avenue for the dissemination of hateful and discriminatory content.[1]

The act of defining what hate speech actually is from a legal perspective and of identifying which utterances fall within the scope of the term raises itself important and significant challenges, not only because different jurisdictions may choose to adopt their own definitions of the conducts subject to being sanctioned, but also because the expression has often been adopted in the context of the general public debate as well as in the context of philosophical, linguistic, sociological, and psychological discussion.

[1] In this sense see, for example, European Commission, 'Communication from the Commission to the European Parliament and the Council, A More Inclusive and Protective Europe: Extending the List of EU Crimes to Hate Speech and Hate Crime' COM(2021) 777 final.

Therefore, a variety of interpretations and connotations of "hate speech" are nowadays available and the challenge that the law faces, *vis-à-vis* the plethora of possible meanings, is that of identifying the appropriate boundaries between permissible and impermissible speech and, consequently, the appropriate boundaries beyond which the imposition of legal sanctions or restrictions ceases to be an acceptable political choice and starts representing an unconstitutional impingement on freedom of expression.

Traditionally, the debate has indeed been mainly focused, precisely, on addressing these questions, that is, whether (and to what extent) regulation on hate speech is compatible with the democratic asset of the state. Different responses have been given by different jurisdictions. Thus, against the backdrop of a constitutional framework where the protection of free speech under the First Amendment is treated as an almost absolute value, the US have generally limited the admissible scope for legal intervention only to those rare cases where "hate speech" takes the forms of a true threat of an imminent lawless action or of low-value "fighting words",[2] provided that such interventions are not motivated by the goal of punishing the expression of a certain – albeit disparaging and discriminatory – viewpoint.[3]

Conversely, within the European context, hate speech has generally been found to represent a phenomenon directly infringing the dignity and right to equality of those individuals or groups it targets and, as such, to be deserving of being constrained with a view to balancing the protection of freedom of expression with the promotion of other equally important constitutional values and principles. In many cases, the European Court of Human Rights (ECtHR) has held that the utterance of certain, particularly egregious, forms of hate speech amounts in fact to an "abuse of right" under the European Convention on Human Rights (ECHR)[4] and, as such, is removed from the guarantees Article 10 sets for freedom of expression and information.[5]

The main challenge that jurisdictions have traditionally had to face, therefore, has thus been that of establishing what the boundaries and limits to free speech are, based on their own constitutional value framework,[6] and of identifying when, conversely, a certain expression leaves the domain of admissible speech, becoming something else – a "fighting word", a true threat, an abuse of right, or, more in general, an utterance constituting illegal speech. Far from being solved, the debate around what should be the contours of legal and illegal hate speech is still ongoing and has, even recently, been at the centre of highly polarized narratives in certain jurisdictions. Think, for instance, of the highly debated Zan

[2] That is, those words "which, by their very utterance, inflict injury or tend to incite an immediate breach of the peace". *Chaplinsky v New Hampshire* 315 US 568 (1942) 582. See *infra*, §2.2.1.

[3] *Brandenburg v Ohio* 395 US 444 (1969); *RAV v City of St Paul* 505 US 377 (1992).

[4] Convention for the Protection of Human Rights and Fundamental Freedoms 1950 art 17.

[5] See, *ex multis*, *Garaudy v France* (dec) [2003] ECtHR 65831/01, ECHR 2003-IX; *Witzsch v Germany (2)* (dec) [2005] ECtHR 7485/03; *Norwood v the United Kingdom* (dec) [2004] ECtHR 23131/03, ECHR 2004-XI; *Pavel Ivanov v Russia* (dec) [2007] ECtHR 35222/04; *M'bala M'bala v France* (dec) [2015] ECtHR 25239/13, ECHR 2015-VIII. See more *infra*, §2.2.3.1.

[6] On the role of the value framework of a country in the creation and application of law, with specific regard to the governance of the digital sphere, see Oreste Pollicino, 'The Quadrangular Shape of the Geometry of Digital Power(s) and the Move towards a Procedural Digital Constitutionalism' (2023) 29 European Law Journal 10.

Draft Law,[7] a project – ultimately rejected by the Senate – for a legislative reform of the Italian framework on hate speech that aimed to amplify the scope of action of the relevant provisions of the Criminal Code, so as to include sexual orientation, gender identity, gender, sex, and disability among the protected grounds of discrimination.[8]

Following the creation of the Internet and the increasing spread of online digital platforms for freedom of expression, regulators have had to deal with a whole new set of issues and have had to redefine their strategies. Indeed, the fight against harmful or illegal content has to deal, today, with the specific characteristics of the contemporary algorithmic age, described by Balkin in the following terms:

> The Algorithmic Society features the collection of vast amounts of data about individuals and facilitates new forms of surveillance, control, discrimination and manipulation, both by governments and by private companies. Call this the problem of Big Data. The Algorithmic Society also changes the practical conditions of speech as well as the entities that control, limit, and censor speech. first, digital speech flows through an elaborate privately-owned infrastructure of communication. Today our practical ability to speak is subject to the decisions of private infrastructure owners, who govern the digital spaces in which people communicate with each other. This is the problem of private governance of speech.[9]

Against this backdrop, lawmakers across the world have progressively turned towards forms of speech governance attempting to harness the computational power[10] of private owners of digital infrastructures with a view, in particular, to increasing their spheres of liability and accountability with respect to the online presence of illegal or harmful content.[11] Through the implementation of such strategies, the goal is to push providers of intermediary services, especially those offering hosting or online platform services, to take the necessary actions to reduce as much as possible the presence of content that is illegal or at least considered to be at odds with the interests of the public at large.

This developing trend in the overall governance of online speech has recently become increasingly relevant – and will likely become even more important – also in the context of the fight against hate speech. For instance, the new Regulation (EU) 2022/2065,[12] commonly known as the Digital Services Act, has set the basis for a new era for the European regulation of content moderation practices.[13] Similarly, legislative attempts in the same direction have been made at the level of domestic state law, as showcased by the examples, for instance, of Germany and France.[14]

---

[7] AS 2005 (XVIII), *Misure di prevenzione e contrasto della discriminazione e della violenza per motivi fondati sul sesso, sul genere, sull'orientamento sessuale, sull'identità di genere e sulla disabilità*.

[8] See *infra*, §4.2.2.2.

[9] Jack M Balkin, 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation' (2018) 51 U.C. Davis Law Review 1149, 1153.

[10] Massimo Durante, *Computational Power: The Impact of ICT on Law, Society and Knowledge* (Routledge 2021).

[11] On the rise of such forms of regulation at the European level, see *infra*, §3. With regard to other jurisdictions, see *infra*, §4.

[12] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act), OJ L 277/1.

[13] See *infra*, §3.5.4.

[14] See *infra*, §§4.2.1., 4.2.2.1.

However, in the light of the increased reliance on the private owners of digital infrastructures for the attainment of the goal of protecting public debate from unwarranted pollutions of the informational ecosystem, it is necessary to investigate and reflect upon the potential impacts this might have on the constitutional framework and the fundamental rights of individual users affected. In particular, it has correctly been noted on multiple sides that the enhancement of intermediary liability and responsibilities will likely have the effect of causing the over-removal of users' content with little regard to the necessary guarantees for the protection of their freedom of expression rights.[15]

Furthermore, addressing these matters is particularly important in light of the rise in the use of automated systems for content moderation and content curation. Artificial intelligence (AI) is today an essential and inescapable resource for platforms to detect, remove, and filter out unwarranted items from the Internet.[16] The deployment of those tools, nevertheless, gives rise to a whole new set of challenges and issues, not only due to the limited transparency characterizing the functioning of implemented algorithms, but also due to the concrete risk of errors and biased results leading to a disproportionate removal of content produced by minority, vulnerable, or discriminated groups of people. Such an issue is particularly problematic when it comes to the governance of the phenomenon of hate speech.[17]

1.1.2. *Objectives and research questions*

The purpose of the present work is to investigate the ways in which "new-school"[18] speech regulation strategies have been developing in recent years – both inside and outside Europe – and how those strategies actually relate to the governance of online hate speech, with a view to mapping out what challenges are lying ahead and to suggesting possible courses of action to address and face those challenges on the European level.

To this end, Chapter 2 first focuses on the preliminary and inescapable set of questions concerning the rationale(s) behind the legislative choice to intervene to restrict and limit the scope of freedom of expression with a view to reducing the spread of hate speech. What is, in particular, the constitutional stance of such a choice? What interests does the proscription of hate speech aim to protect? How are those interests balanced with the fundamental right to freedom of expression? Clearly, the answer to such questions is not univocal, both because the responses will vary from jurisdiction to jurisdiction and because jurisdictions proscribing hate speech generally offer a plurality of reasons justifying their choice. Nevertheless, the present work argues for a regulatory approach towards hate

[15] In this sense, with specific regard to the Digital Services Act, see, among others, Joan Barata, 'The Digital Services Act and Its Impact on the Right to Freedom of Expression: Special Focus on Risk Mitigation Obligations' (*DSA Observatory*, 27 July 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/> accessed 3 December 2021. See more *infra*, §3.5.3.4.

[16] See *infra*, §5.3.

[17] See *infra*, §5.4.

[18] Jack M Balkin, 'Old-School/New-School Speech Regulation' (2014) 127 Harvard Law Review 2296. See *infra*, §3.2.2.

speech governance that – cognizant of the fact that hate speech inherently perpetuates dynamics of dominance between speaker and targets – aims to serve as a remedy precisely against such dynamics. In this sense, this work suggests using the lens of substantive equality when dealing with the regulation of hate speech, which should ultimately be oriented towards the active promotion of an equal standing of all demographics in society.

The second set of questions, addressed within Chapter 3, concerns the ways in which the law has evolved in Europe – both at the level of the Council of Europe and at the level of the European Union (EU) – with regard to the area of intermediary liability in general and of liability for third-party hate speech in particular. This analysis is done with due regard to critically investigating how the resulting framework does, in fact, relate with constitutional values and fundamental rights and whether the principle of (substantive) equality enters or not into such a framework. How has the ECtHR case law evolved with respect to intermediary duties to remove illegal content and, specifically, hate speech? Is intermediary liability for third-party hate speech consistent, in the context of the ECHR framework, with the right to freedom of expression? In parallel to the development of that case law, which trends have been followed on these matters by the EU? What novelties, in particular, has the 2022 Digital Services Act introduced in the context of the regulation of online content moderation and to what extent do these novelties apply to the case of hate speech moderation? What are the main limitations of the Digital Services Act and how will such novelties affect freedom of expression and the right to equality?

These questions are, furthermore, strictly related with another set of questions addressed within Chapter 4. Indeed, the evolving EU legal framework on hate speech moderation is not set within a void. Its implementation will necessarily have to take into account the domestic legislation of the various EU Member States but may also, given the transnational character of online content, clash with the legal systems of foreign jurisdictions. To what extent, therefore, is EU law consistent with the law of Member States? How are different legal systems outside of Europe addressing hate speech and the presence of illegal content on the Internet? What challenges may arise with respect, in particular, to the relationship between the Digital Services Act and the constitutional framework of the United States (US)? Are other jurisdictions following a regulatory model similar to the European one?

Finally, precisely because developing legal trends – both inside and outside of Europe – are progressively shifting towards increasing the liability and responsibility of providers of intermediary services to remove unwarranted content, a fourth set of research questions, dealt with in Chapter 5, will focus specifically on the ways in which private platforms enforce their own duties and content moderation practices. Indeed, the manner in which these private actors govern online speech, and hate speech in particular, has highly significant consequences in terms of how users' fundamental rights are affected and in terms of whether such practices actually enable to reach the inherent goals of hate speech governance. In this context, close consideration is taken of the use of AI for the purposes of detecting and removing hateful content. More specifically, how is hate speech defined and treated under private platforms' terms and conditions and what is the system of values

underpinning those terms and conditions? How are these private rules enforced from a technical point of view? What are the main characteristics, capabilities, and limitations of automated systems of content moderation? What is the overall impact of private content moderation practices on freedom of expression and the right to equality? In light of such impact, what are the challenges that lie ahead for the implementation of EU law, notably the Digital Services Act? Can the principle of substantive equality represent a valid focal point to orient future legislative and policy choices?

1.2. Notes on methodology

1.2.1. *Material scope of the research*

As described above, the present research is focused on the analysis of legislative responses against the dissemination of online illegal or harmful content, with a close eye on the phenomenon of hate speech. In this respect, the notion itself of hate speech is not always clear, also due to the significant increase in the use of the term in everyday language and in the context of non-legal debates and discussions. As better clarified within Chapter 2, the present research mainly refers to a concept of hate speech that is, in its essence, comparable to that adopted by the European Commission against Racism and Intolerance inside its General Policy Recommendation No. 15.[19]

Admittedly, the definition contained within that Recommendation is in fact rather ample as regards its scope of application, because it considers as relevant a wide variety of expressive conducts. Nevertheless, that definition is highly relevant inasmuch as it clarifies that the specific feature distinguishing hate speech is that it is "based on a non-exhaustive list of personal characteristics or status that includes 'race', colour, language, religion or belief, nationality or national or ethnic origin, as well as descent, age, disability, sex, gender, gender identity and sexual orientation".[20] In other words, hate speech is inherently rooted in and is a direct expression of discrimination.[21]

It is for this reason that, within the present work, specific attention is given to the analysis of the phenomenon of hate speech under the lens of its strict interrelation with the typical categories of anti-discrimination law. In particular, the work borrows from that field the concept of substantive equality, intended as the – constitutionally relevant – aspiration towards the active elimination of the barriers to the pursuit of true equality between societal demographics. More specifically, the work tends to refer to the concept of substantive equality as theorized by Sandra Fredman, who, rather than considering it as a unitary principle, identifies it as a complex one, composed of a variety of dimensions.[22] Particular attention will be given to the "participative" dimension of substantive equality,

[19] European Commission against Racism and Intolerance, 'General Policy Recommendation No. 15 on Combating Hate Speech' (Council of Europe 2015) CRI(2016)5. See *infra*, §2.2.3.1.

[20] ibid.

[21] See more *infra*, §2.2.4.

[22] Sandra Fredman, 'Substantive Equality Revisited' (2016) 14 International Journal of Constitutional Law 712. See *infra*, §2.5.2.1.

which, according to the present work, provides a lens of utmost importance for the definition of responses to the upcoming challenges of hate speech governance.

Furthermore, the analysis contained within the present work shall focus specifically on the interrelations between liability-enhancing regulation, private governance of online speech, and related impacts on fundamental rights of users.

In this respect, it is first of all important to stress that, from a terminological point of view, terms relating to "Internet intermediaries", "Internet service providers", "online platforms", etc., shall generally be used interchangeably as umbrella expressions to refer to the composite and extremely wide category of private actors that are active in the market of digital services. Nevertheless, when discussing the content or application of specific legislative acts, the work shall rely on the technical terms and definitions contained within those sources. Thus, for instance, when referring to the framework established by the Digital Services Act, the term "online platform" shall be intended as referring specifically to "a hosting service that, at the request of a recipient of the service, stores and disseminates information to the public".[23]

Second, because the purpose of the research is to investigate how the law can influence the hate speech moderation practices of private actors and how those practices affect, in turn, the liberties of users – and, consequently, the governance of hate speech from an anti-discriminatory perspective –, this work, while acknowledging the rise of new non-human purveyors of hate speech, shall not deal specifically with that aspect. In particular, the spread of more and more advanced generative AI systems, including large language models, has raised the challenge of the emergence of new forms of hate speech originating from those technologies. While representing a critical challenge for the future, such an issue falls outside the scope of this research. AI will, instead, be considered inasmuch as it is increasingly used by platforms for the purposes of detecting and removing hate speech and may thus impact, in particular, the right for users to enjoy online their right to freedom of expressions in conditions of equality.[24]

1.2.2. *Territorial scope of analysis*

The research mainly considers the European legal framework on hate speech moderation, taking into consideration the developments occurring both within the case law of the ECtHR and within the body of legislation of the EU. Within the Old Continent, indeed, digital policies, including policies concerning the governance of online speech, are increasingly confronted with on a supranational – rather than merely national – level, with significant regulatory interventions especially from EU institutions.

Thus, Chapter 2 mainly considers the debate on hate speech regulation by focusing on the way European Courts and European legal and policy documents have addressed the matter. Chapter 3, similarly, contains an extensive review of the European framework on intermediary liability regulation. Chapter 5, dealing with platform governance practices

[23] DSA art 3, lett (i). See *infra*, §3.5.3.
[24] See *infra*, §5.

and on the use of automated systems for hate speech moderation, investigates how EU law can face the human rights challenges raised by these practices and tools, with a view to fostering the injection within them of principles and values that are the expression of the European constitutional framework.

At the same time, the research, acknowledging in particular the transnational nature of the phenomenon of hate speech, also considers other legal frameworks from a comparative perspective. In Chapter 2, for example, specific regard is given to the approach of the US towards hate speech governance against the background of the evolution of First Amendment jurisprudence across the last two centuries. Additionally, Chapter 4 considers the topic of intermediary liability legislation and hate speech governance precisely by taking a comparative overview of jurisdictions both within and outside the EU.

1.2.3. *Aspects of interdisciplinarity*

The research mainly addresses the topic of hate speech governance from a legal perspective. Thus, in this respect, the work is based on an extensive review of relevant literature on this topic and related issues, as well as upon landmark case law and legislation. As already mentioned, the analysis mainly addresses the European landscape, but comparative elements are also present throughout the work. Through this analysis, the work aims to identify the rationale justifying the adoption of measures against hate speech, the novel challenges brought about in this respect by the Internet, and the issues in terms of fundamental rights related to the development of new legislative responses. The goal is, ultimately, to suggest a key for the interpretation of the phenomenon as a whole and, thus, to suggest preliminary tools to address the challenges still lying ahead.

Nevertheless, the full understanding of the phenomenon of online hate speech, as well as of the role and impact of contemporary practices of (private) content moderation, also requires considering relevant technological aspects. In this respect, the legal and policy analysis is complemented by a review of relevant technical literature. Specifically, the analysis contained in Chapter 5 aims to give an overview of the technical aspects of the AI systems deployed to remove hate speech content from the Internet, with a view to highlighting those systems' limitations and the consequent effects on fundamental rights and public speech governance policies.

1.3. Structure of the work

While the previous sections have already highlighted the key aspects of the present work, the following subsections will give a more detailed overview of the content of the dissertation's Chapters.

1.3.1. *Chapter 2: Setting the framework on hate speech governance*

Chapter 2 introduces the many issues and challenges relating to the development of an adequate hate speech governance system, both within the online and offline environment.

First, it aims to give the reader the necessary background information concerning the origins of the notion of "hate speech" in the US system and to give an overview of how the international framework on hate speech has evolved throughout the twentieth century.[25] It also introduces the European regional framework on hate speech, considering both the case law of the ECtHR and the legislation of the EU. In this way, the concept itself of hate speech is better investigated from a legal point of view, serving as a baseline for the remainder of the work.

The Chapter then moves on to address the debate concerning the main rationales behind the possible legal options *vis-à-vis* the phenomenon of hate speech. The US framework is, in particular, taken as a model of a "liberal" and "tolerant" approach towards the "thought that we hate".[26] Conversely, the European perspective, especially that enshrined within the judgments of the ECtHR, is taken as a model of a more "militant" approach, oriented towards the protection of the rights, dignity, and equality of groups traditionally targeted by hate speech.

In this respect, the peculiar aspects characterizing specifically the online dimension of hate speech are also investigated, with a view to highlighting the emerging challenges set by the Internet and to showcasing how digital technologies have themselves been described in different terms across the two sides of the Atlantic. In the US, the narrative has in fact generally been optimistic, with the recognition of the Internet as an extraordinary avenue for free speech, whereas on the Eastern side of the Atlantic more attention has been given to the new risks and threats posed by it.

The Chapter, finally, argues for an interpretation of the hate speech phenomenon as inherently grounded in its relationship to the perpetuation of the dynamics of power and dominance within the social fabric, starting from some basic notions and concepts taken from speech act theory. As a result, the Chapter suggests that the purpose of hate speech governance should be precisely to combat the dominance dynamics entailed by it and argues that, in order to do so, legal strategies in this area should be buttressed by following, as a target, the principle of substantive equality.

1.3.2. *Chapter 3: Intermediary liability and hate speech in Europe*

Chapter 2 having explored the main features characterizing the phenomenon of hate speech both offline and online, Chapter 3 delves into the developments undergone by ECtHR case law and EU legislation in terms of intermediary liability for third-party content.

[25] See, in particular, International Covenant on Civil and Political Rights 1966 arts 19–20; International Convention on the Elimination of All Forms of Racial Discrimination 1965 art 4.

[26] *Matal v Tam* 582 US __ (2017) 25.

With respect to the ECtHR, specific attention is given to the landmark judgments of *Delfi*[27] and *MTE*,[28] as well to the subsequent legacy of those decisions.[29] In this sense, the Chapter discusses, in particular, how the ECtHR case law has established a rather exceptional approach towards intermediary liability for third-party hate speech content, as opposed to other types of unlawful material. Indeed, whereas from *MTE* onwards the Strasbourg Court has adopted a narrow approach towards the governmental enforcement of forms of intermediary liability for the dissemination of illegal content, due to concerns related to Article 10 ECHR, hate speech, representing itself an abuse of freedom of expression, has generally been considered to be deserving of more invasive state intervention.

As regards the EU, Chapter 3 stresses the shift from an inherently liberal original phase towards an increasingly interventionist approach. In this respect, the Chapter first investigates the active role of the Court of Justice of the EU (CJEU) in adapting the interpretation of the e-Commerce Directive[30] in the light of the evolving technological paradigm. Then, the work gives an overview of the most recent (from the end of the 2010s onwards) legislative trends characterizing the Union's policy strategies on content moderation, critically assessing the characteristics of the developing framework and the challenges arising from a constitutional and human rights law perspective.

Finally, the Chapter moves on to analyse the significant development in EU law represented by the enactment of the already mentioned Regulation (EU) 2022/2065, that is, the Digital Services Act. The new Regulation, indeed, operates a general and horizontally applicable reform of the system established in 2000 by the e-Commerce Directive. In particular, the Chapter aims to give an overview of the new legislation, focusing on the new set of rules on providers' due diligence obligations "for a transparent and safe online environment", while also investigating the relationship between the Act and the challenge of hate speech moderation.

1.3.3. *Chapter 4: Comparative perspectives*

Chapter 4 gives a broad overview, from a comparative perspective, of how the challenges raised by online hate speech have – or have not – been addressed by different jurisdictions.

First, the Chapter explores the relationship between the EU framework and the domestic legislation of some notable Member States. Among these, specific consideration is

[27] *Delfi AS v Estonia* [2015] ECtHR [GC] 64569/09, ECHR 2015.

[28] *Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v Hungary* [2016] ECtHR 22947/13.

[29] See, in particular, *Pihl v Sweden* (dec) [2017] ECtHR 74742/14; *Høiness v Norway* [2019] ECtHR 43624/14; *Standard Verlagsgesellschaft Mbh v Austria (no 3)* [2021] ECtHR 39378/15; *Sanchez v France* [2023] ECtHR [GC] 45581/15, ECHR 2023.

[30] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce'), OJ L 178/1.

given to the German Network Enforcement Act[31] which, enacted in 2017, has since then served as an internationally relevant blueprint for the regulation of intermediary liability with respect to user-generated hate speech. Subsequently, the experiences of three major EU countries – France, Italy, and Spain – are described. All three jurisdictions, indeed, have addressed the phenomenon of online hate speech differently – with more or less successful outcomes – and thus showcase the variety of domestic legal tools that the application of the Digital Services Act will have to take into consideration. Additionally, the Chapter critically discusses the online speech governance approaches of two Eastern European countries, Poland and Hungary, that have suffered in recent years from forms of democratic backsliding. Once again, the relationship of those national approaches to the EU's Digital Services Act is the main focus, especially in the light of the adoption in those countries of much debated "memory laws".

Second, Chapter 4 describes the recently adopted UK Online Safety Act, with a view to outlining its material, subjective, and territorial scope of application, the new set of duties imposed upon providers of Internet services, and the role the Act shall play in the fight against online hate speech across the UK. The Online Safety Act, indeed, offers many interesting terms of comparison with the Digital Services Act, the two pieces of legislation having aims and goals that largely coincide.

Third, the Chapter takes once again a look at the legal framework of the US concerning intermediary liability, a framework that is, in fact, radically different from the one characterizing the EU. In this respect, Chapter 4 addresses in particular the rise, at the end of the 1990s, of the famous Section 230 of the Communications Decency Act,[32] outlining the fundamental role played by the provision in the development of the US case law on intermediary liability. The interplay between Section 230, the state action doctrine, and the First Amendment is also dealt with, as well as the increasing critiques moved both by conservatives and liberals towards the current system and the attempts that have thus been made on both sides to amend the provision. Indeed, the success or failure of such attempts may well support or hamper a positive relationship between the Digital Services Act and US constitutional law.

Finally, the Chapter briefly outlines some other legislative approaches worldwide. It is indeed important to highlight the plurality of techniques that can and have been adopted with respect to online speech governance and to bear in mind, specifically, that the regulatory strategies of Western democracies may have to deal with other regulatory frameworks.

1.3.4. *Chapter 5: Platform standards and automated moderation*

The goal of Chapter 5 is mainly that of investigating how providers of intermediary services themselves have addressed the phenomenon of hate speech, both in terms of the

[31] *Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken* (*Netzwerkdurchsetzungsgesetz* - NetzDG) 2017 (BGBI I S 3352).

[32] Communications Decency Act 1996.

policies adopted and in terms of the practical means of enforcement of those policies. This analysis is indeed considered to be essential for the purposes of identifying what further challenges still lie ahead in the governance of online hate speech and of defining the future strategies to be implemented by the EU in this respect.

The Chapter is, in its essence, structured into two parts. The first part focuses on the private anti-hate speech strategies applied by major providers of intermediary services. In particular, the work deals with the policies, standards, and terms and conditions formulated by those actors, with a close eye on the case of Meta platforms – whose terms and conditions are analysed in the light of the decisions rendered in recent years by the Meta Oversight Board – and also considering the cases of X, YouTube, and TikTok. The goal is, in this respect, to search for common patterns and features, as well as to compare those platforms' policy instruments with the European legal framework and, importantly, their consistency with the principle of substantive equality. From a more technical perspective, the Chapter also considers the technical means through which hate speech is actually moderated by private platforms, focusing specifically on the rise of AI detection systems and giving an overview of their main features, their functioning and limitations.

The second part of the Chapter contemplates the challenges that the ways in which platforms moderate hate speech pose to the law and, specifically, to European hate speech governance and the protection of constitutional values and fundamental rights. In this respect, the work highlights how the resort to AI systems for content moderation and content curation necessarily entails the presence of certain margins of error – thus requiring policymakers and lawmakers to define the limits of "acceptability" of error – and suggests substantive equality as a proxy to determine the borders of acceptable errors in the context of hate speech moderation in Europe. The Chapter also indicates some areas of action to be addressed – namely, the areas of transparency, rule of law, and due process – and underlines how the Digital Services Act may indeed serve as the baseline for such mitigating interventions within the European context.

Most notably, the Chapter argues that the adoption of more specific guidelines with regard to the moderation of hate speech could represent a noteworthy asset. In this respect, the Chapter calls for a renovation of the current EU Code of Conduct on Illegal Hate Speech.[33]

[33] Code of Conduct on Countering Illegal Hate Speech Online 2016.

2.

Hate Speech and Substantive Equality: A Theoretical Framework

Summary: 2.1. Introduction. – 2.2. The concept of hate speech in the global and European context. – 2.2.1. Origins of the term and constitutional approach to hate speech in the United States. – 2.2.2. Lessons from international human rights law. – 2.2.2.1. Article 20 ICCPR. – 2.2.2.2. Article 4 ICERD. – 2.2.3. Hate speech in Europe. – 2.2.3.1. The Council of Europe. – 2.2.3.2. The European Union. – 2.2.4. *Interim* conclusions. – 2.3. The transatlantic debate on hate speech regulation. – 2.3.1. The liberal approach: the US model of the free marketplace of ideas. – 2.3.2. The militant approach: the case of Europe. – 2.4. Hate speech and the Internet. – 2.4.1. Free speech and information in the digital age. – 2.4.2. Main characters of online hate speech. – 2.4.2.1. Permanence. – 2.4.2.2. Itinerancy. – 2.4.2.3. Anonymity. – 2.4.2.4. Cross-jurisdictional nature of online content. – 2.4.3. The role of algorithmic content moderation and curation. – 2.5. Anti-discrimination perspectives on hate speech: a substantive equality approach. – 2.5.1. Hate speech as domination: some takeaways from speech act theory. – 2.5.2. Substantive equality as a lodestar for hate speech governance. – 2.5.2.1. The concept of substantive equality. – 2.5.2.2. Substantive equality and hate speech in the European multi-level human rights protection system. – 2.5.3. Hate speech governance and substantive equality in the world of bits. – 2.6. Conclusions.

2.1. Introduction

The purpose of the present Chapter is to introduce the many issues and challenges relating to the development of an adequate hate speech governance system both within the online and offline environment. The concept of "hate speech", indeed, is not univocal, nor are the legal approaches to such a phenomenon within the global context. The rationale, itself, behind a legislative reaction against hate speech has long been the topic of a doctrinal and political debate which is far from being solved. The perspective adopted within the present work is, nonetheless, that hate speech governance, at least within the European context, should be driven, primarily, by the goal of fostering and promoting the substantive equality of the individuals and groups of individuals that are more commonly vulnerable to hate speech victimization.

The Chapter is structured as follows. Section 2.2 aims to give the reader the necessary background information concerning the origins of the notion of "hate speech" (§2.2.1),

as well as about the international (§2.2.2) and regional – namely, European (§2.2.3) – human rights framework on hate speech, so as to identify common patterns and/or inconsistencies (§2.2.4).

Section 2.3 underscores the different, and often opposing, approaches that the law can take with respect to the discussed phenomenon: for this purpose, the United States (§2.3.1) and European (§2.3.2) perspectives are considered, as they represent key models for "liberal" *versus* "militant" approaches to hate speech regulation.

Section 2.4 focuses on the context of the Internet, highlighting, in particular, how the specific characters of online communication and information (§2.4.1) can influence the way hate speech is disseminated and distributed and the way this may affect its targets (§2.4.2). Due regard is also given to the increasingly important role played, in the context of expression and information rights across the Internet, by the resort to algorithmic practices of content moderation and content curation (§2.4.3).

Section 2.5, moving from some preliminary notions grounded in speech act theory, investigates the links and connections between the European approach to hate speech governance, as described in the previous sections, and the principle of substantive equality. In particular, after having stressed the capability of hate speech to produce illocutionary effects consisting of the perpetuation of dynamics of power and dominance within the social tissue (§2.5.1), the Section moves on to argue that the principle of substantive equality could (and should) be invoked as a lens to interpret the goals of hate speech governance, the purpose of which could be intended precisely as providing a remedy against those dynamics of power and dominance (§2.5.2). It is also noted that, that being the case, governing the phenomenon of hate speech in the digital sphere raises specific challenges related, in particular, to the (more and more automated) private moderation systems deployed by platforms (§2.5.3).

Finally, Section 2.6 briefly provides some *interim* conclusions which shall represent the steppingstone for Chapter 3.

2.2. The concept of hate speech in the global and European context

When addressing the phenomenon of "hate speech", one of the most significant challenges is that of identifying what the expression actually means. Admittedly, there is in fact no universally accepted definition of the term. On the one hand, if one considers the phenomenon of hate speech from a legal perspective, one is confronted with an extraordinary variety of legal frameworks across the globe, which may vary not only with respect to the solutions adopted but also with respect to the actual scope of the notion of "hate speech" itself. On the other hand, "hate speech" is not only a legal concept, as it is also relevant for other fields of knowledge such as philosophy, linguistics, psychology, and

sociology. Additionally, the expression has increasingly entered the ordinary and everyday language of people who are not professionals of the law.[1]

The purpose of the present section is not that of offering a solution to the interpretive challenges set by the term but, rather, that of presenting an overview of its content under landmark international human rights law, as well as under the European human rights framework (hereby including both the Council of Europe and the European Union systems). In other words, the purpose is to highlight what forms of speech and what types of hate may be included within the umbrella expression "hate speech", at least within the Old Continent, and thus to identify the fundamental features characterizing the phenomenon of "hate speech" as intended for the purposes of the present research.

2.2.1. *Origins of the term and constitutional approach to hate speech in the United States*

The Oxford English Dictionary – in defining "hate speech" as speech, address or written material capable of inciting hatred or intolerance, especially against a particular social group on the basis of its members' ethnicity, religious beliefs, sexuality, etc. – clarifies that the origins of the term can be traced back to the United States.[2] Indeed, the debate concerning hate speech and free speech in the US constitutional system dates back to the 1920s, when historical victims of prejudice and discrimination launched a concerted effort to react against the forms of oppression they had traditionally been subjected to. In so doing, these groups entered into disagreement with the recently born American Civil Liberties Union (ACLU), dedicated primarily to the promotion and defence of the values of free speech.[3]

Subsequently, throughout the twentieth century, US constitutional jurisprudence on hate speech, under the guidance of the Supreme Court (SCOTUS), underwent a significant evolution. Most notably, after a brief period where the phenomenon was categorized as a form of "group libel" and was considered to be legitimately subjectable to punishment in the aftermath of *Beauharnais v Illinois*,[4] the SCOTUS inaugurated with the 1969 decision of *Brandenburg v Ohio*[5] a consistent strand of case law, still applicable today, cutting down significantly the possibility for the government to impose limitations and restrictions upon the utterance of hate speech. Indeed, the inherent rejection of any form of content- or viewpoint-based regulation, characterizing US First Amendment

[1] Alexander Brown, *Hate Speech Law: A Philosophical Examination* (Routledge 2015); Alexander Brown, 'What Is Hate Speech? Part 1: The Myth of Hate' (2017) 36 Law and Philosophy 419; Alexander Brown and Adriana Sinclair, *The Politics of Hate Speech Laws* (Routledge 2019); Irene Spigno, *Discorsi d'odio. Modelli Costituzionali a Confronto* (Giuffrè 2018).

[2] 'Hate, n.' <https://www.oed.com/view/Entry/84550> accessed 28 December 2022.

[3] Samuel Walker, *Hate Speech: The History of an American Controversy* (University of Nebraska Press 1994) 9–10.

[4] *Beauharnais v Illinois* 343 US 250 (1951).

[5] *Brandenburg v Ohio* 395 US 444 (1969).

jurisprudence,[6] generally entails the exclusion of the constitutional legitimacy of hate speech bans, which can only be adopted in specific cases, such as when those expressions amount to "true threats"[7] or, even more importantly, when they constitute "fighting words" – that is, when "by their very utterance" they "inflict injury or tend to incite an immediate breach of the peace".[8]

With regard to the latter category, it is worth mentioning the case of *Chaplinsky v New Hampshire*, where, while defining for the first time the concept of fighting words in the context of US free speech jurisprudence, the SCOTUS held that this category, representing a form of low-value speech, should not be considered worthy of full First Amendment protection, so that the adoption of legal reactions against it should generally be considered as allowed by the US Constitution. Indeed, "such utterances are no essential part of any exposition of ideas, and are of such slight social value as a step to truth that any benefit that may be derived from them is clearly outweighed by the social interest in order and morality".[9] Quite evidently, the Court's approach to fighting words in Chaplinsky could have opened the doors to the constitutional legitimacy of many forms of hate speech bans in the US. Nonetheless, subsequent case law from the SCOTUS went on to reduce the scope of applicability of the category. First, the 1971 decision of *Cohen v California* redefined that class of speech, concluding that it included only those words that, "when addressed to the ordinary citizen, are, as a matter of common knowledge, inherently likely to provoke violent reaction".[10] Through this judgment, the SCOTUS thus significantly heightened the standards required for the adoption of measures against fighting words, as showcased, namely, by the famous *Skokie* judicial saga.[11]

---

[6] Laurence H Tribe, *American Constitutional Law* (2nd edn, Foundation Press 1988) 789–792; Martin H Redish, 'The Content Distinction in First Amendment Analysis' (1981) 34 Stanford Law Review 113; Geoffrey R Stone, 'Content Regulation and the First Amendment' (1983) 25 William & Mary Law Review 189; Susan H Williams, 'Content Discrimination and the First Amendment' (1991) 139 University of Pennsylvania Law Review 615; Leslie Kendrick, 'Content Discrimination Revisited' (2012) 98 Virginia Law Review 231. See also, *ex multis*, *Police Department of the City of Chicago v Mosley* 508 US 92 (1972).

[7] "'True threats' encompass those statements where the speaker means to communicate a serious expression of an intent to commit an act of unlawful violence to a particular individual or group of individuals … a prohibition on true threats 'protect[s] individuals from the fear of violence' and 'from the disruption that fear engenders,' in addition to protecting people 'from the possibility that the threatened violence will occur.' … Intimidation in the constitutionally proscribable sense of the word is a type of true threat, where a speaker directs a threat to a person or group of persons with the intent of placing the victim in fear of bodily harm or death". *Virginia v Black* 538 US 343 (2003) 359–360.

[8] *Chaplinsky v New Hampshire* 315 US 568 (1942) 582.

[9] ibid 572. See, on *Chaplinsky* and on the concept of "low-value speech", Genevieve Lakier, 'The Invention of Low-Value Speech' (2015) 128 Harvard Law Review 2166.

[10] *Cohen v California* 403 US 15 (1971) 20.

[11] Indeed, in light of *Cohen*'s redefinition of "fighting words", the Illinois Supreme Court held that the Village of Skokie's refusal to allow a neo-Nazi parade was unconstitutional because the wearing of the Swastika and of Nazi regalia could not be considered to constitute a case of "fighting words": "The display of the swastika, as offensive to the principles of a free nation as the memories it recalls may be, is symbolic political speech intended to convey to the public the beliefs of those who display it … It does not, in our opinion, fall within the definition of "fighting words," and that doctrine cannot be used here to overcome the heavy presumption against the constitutional validity of a prior restraint". *Village of Skokie v Nat'l Socialist Party of America* 373 NE2d 21 (Ill 1978) 24.

Additionally, the 1992 decision of *RAV v City of St. Paul*[12] added further important limitations to the possibility for local, state, and federal authorities to impose governmental restrictions on the phenomenon of hate speech. In this case, the applicant – a juvenile at the time of the facts – had been sentenced, together with other people, for having burnt a cross in front of the house of an African American who had recently moved into their neighbourhood. Such conduct had been punished under a local statute passed by the City of St. Paul, Minnesota, which made the placement on private or public property of symbols, objects, appellations, characterizations or graffiti, with the knowledge or reasonable expectancy that such action would stir anger, alarm or resentment in others on the basis of race, colour, creed, religion or gender, a misdemeanour. The SCOTUS unanimously held that the ordinance represented an inadmissible restriction on freedom of speech. Most notably, the majority, although accepting the view that the statute only specifically dealt with the class of fighting words, concluded nonetheless that its purpose was precisely that of prohibiting otherwise permitted speech solely on the basis of the subjects the speech addressed.[13] In other words, the Court's majority argued that the choice of the statute to only address those fighting words that were based on the categories of race, colour, creed, religion, and gender inherently indicated the actual goal not of proscribing fighting words as such but, rather, of opposing the utterance of a specific point of view. As a result, the ordinance, precisely because of its "underbreadth",[14] was considered to be vitiated by viewpoint discrimination and, therefore, unconstitutional under the First Amendment.[15]

As a result, the US constitutional framework has repeatedly proven to be, in general terms, opposed to the adoption of forms of hate speech bans as such, precisely because the expression of discriminatory and dehumanizing opinions cannot, *per se*, be subjected to governmental constraints without these translating into unwarranted limitations on specific viewpoints and, thus, upon the free marketplace of ideas protected by the First Amendment. Quite curiously, the global approach towards hate speech regulation has

[12] *RAV v City of St Paul* 505 US 377 (1992).

[13] ibid 381.

[14] The expression "underbreadth", in fact, was adopted in critical terms within Justice White's concurring opinion. See ibid 402.

[15] "Although the phrase in the ordinance, 'arouses anger, alarm or resentment in others,' has been limited by the Minnesota Supreme Court's construction to reach only those symbols or displays that amount to 'fighting words,' the remaining, unmodified terms make clear that the ordinance applies only to 'fighting words' that insult, or provoke violence, 'on the basis of race, color, creed, religion or gender.' Displays containing abusive invective, no matter how vicious or severe, are permissible unless they are addressed to one of the specified disfavored topics. Those who wish to use 'fighting words' in connection with other ideas – to express hostility, for example, on the basis of political affiliation, union membership, or homosexuality – are not covered. The First Amendment does permit St. Paul to impose special prohibitions on those speakers who express views on disfavored subjects … moreover, the ordinance goes even beyond mere content discrimination, to actual viewpoint discrimination … 'fighting words' that do not themselves invoke race, color, creed, religion, or gender – aspersions upon a person's mother, for example – would seemingly be usable *ad libitum* in the placards of those arguing *in favor* of racial, color, etc., tolerance and equality, but could not be used by those speakers' opponents". ibid 391. In this respect see, among others, Akhil Reed Amar, 'The Case of the Missing Amendments: R.A.V. v. City of St. Paul' (1992) 106 Harvard Law Review 124; Michel Rosenfeld, 'Hate Speech in Constitutional Jurisprudence: A Comparative Analysis' (2002) 24 Cardozo Law Review 1523.

evolved in a manner which is rather different from that of the jurisdiction where the term originated. Not only has the body of international human rights protection laws provided significantly for the introduction of legal measures and responses against the discussed phenomenon, but many regional, as well as national, frameworks have increasingly moved towards the imposition of limitations on the utterance and spread of forms of hate speech. In this respect, the following subsections will focus, specifically, on the UN and European landscapes.

2.2.2. *Lessons from international human rights law*

International human rights documents represent the necessary starting point of any discussion concerning the imposition of legal limitations and restrictions to hate speech. Most notably, the International Covenant on Civil and Political Rights (ICCPR)[16] and the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD)[17] have been paramount in shaping the subsequent development of hate speech regulations on a global scale.[18]

2.2.2.1. Article 20 ICCPR

Consistently with the Universal Declaration of Human Rights,[19] Article 19 of the ICCPR explicitly recognizes individuals' right to freedom of expression, which includes the freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers and through any means chosen. Nonetheless, paragraph 3 of the Article also recognizes that freedom of expression, since it "carries with it special duties and responsibilities", may be subjected to certain restrictions when these are provided by the law and are necessary in order to guarantee the rights and reputation of others or to protect publicly relevant goods such as national security, public order, or public health or morals.

Additionally, Article 20, paragraph 2, notably affirms that "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law",[20] a provision which is quite unique within the Covenant itself as it is the only one requiring (and not prohibiting) an active intervention by states parties.[21] Although it does not employ the term "hate speech", the ICCPR is thus considered to be one of the first and most significant documents introducing its notion and concept at an international level, as it recognized as legally relevant a set of conducts which pertain specifically to the sphere of what hate speech is: that is, incitement to discrimination,

[16] International Covenant on Civil and Political Rights 1966.

[17] International Convention on the Elimination of All Forms of Racial Discrimination 1965.

[18] Stephanie Farrior, 'Molding The Matrix: The Historical and Theoretical Foundations of International Law Concerning Hate Speech' (1996) 14 Berkeley Journal of International Law 1.

[19] Universal Declaration of Human Rights 1948 art 19.

[20] International Covenant on Civil and Political Rights art 20.

[21] Ivan Hare, 'Extreme Speech Under International and Regional Human Rights Standards' in Ivan Hare and James Weinstein (eds), *Extreme Speech and Democracy* (Oxford University Press 2009) 70.

incitement to hostility, and incitement to violence through the advocacy of hatred that is based either on ethnicity or religious beliefs.[22]

The relationship between Article 19 and Article 20 has raised suspicions as to their coherence. Such doubts, however, were first rejected by the Human Rights Committee in its General Comment No. 11 (1983), according to which "these required prohibitions are fully compatible with the right of freedom of expression".[23] Subsequently, the Committee confirmed its position once again in its General Comment No. 34 (2011), where it clarified that Article 20, paragraph 2, is to be considered as a *lex specialis* of Article 19, paragraph 3: this meant, according to the Committee, that states parties, when implementing the hate speech prohibition, must comply with the threefold requirement set therein (i.e., prior provision by the law; legitimate aim; and proportionality).[24]

Furthermore, Article 20 does not require states to prohibit any type of advocacy of hatred, but only those forms of advocacy that constitute "incitement", that is, those that aim at provoking specific reactions and are in fact capable of producing contingent harm.[25] As a result, the threshold set by the provision is rather high and "does not ban hate speech outright but only requires the prohibition of certain qualified types of hate speech".[26] As underlined by Temperman, the act of incitement under Article 20 implies a triangular scheme where an advocator produces an "imminent risk" or "likelihood" that the audience will be stirred to discrimination, hostility and violence against the target group.[27]

---

[22] In this respect, the Committee on the Elimination of Racial Discrimination clarified that, although "the term hate speech is not explicitly used in the Convention, this lack of explicit reference has not impeded the Committee from identifying and naming hate speech phenomena and exploring the relationship between speech practices and the standards of the Convention". Committee on the Elimination of Racial Discrimination, 'General Recommendation No. 35. Combating Racist Hate Speech' (United Nations 2013) CERD/C/GC/35 para 5.

[23] Human Rights Committee, 'General Comment No. 11. Prohibition of Propaganda for War and Inciting National, Racial or Religious Hatred (Art. 20)' (United Nations 1983) para 2.

[24] Human Rights Committee, 'General Comment No. 34. Article 19: Freedom of Opinion and Expression' (United Nations 2011) CCPR/C/GC/34 paras 51–52. Similarly, in *Ross v. Canada*, the Human Rights Committee had declared that "restrictions on expression which may fall within the scope of article 20 must also be permissible under article 19, paragraph 3". *Malcolm Ross v Canada* [2000] Human Rights Committee CCPR/C/70/D/736/1997 [10.6]. Prior to such clarifications, in fact, international law experts disagreed on whether art 20, para 2, was to be recognized as a mere elaboration of art 19, para 3, or whether it were to be interpreted as a different and additional basis for the imposition of restrictions on freedom of expression: see Ineke Boerefijn and Joanna Oyediran, 'Article 20 of the International Covenant on Civil and Political Rights' in Sandra Coliver (ed), *Striking a Balance. Hate Speech, Freedom of Expression and Non-Discrimination* (Article 19 1992) 30.

[25] Susan Benesch, 'Contribution to OHCHR Initiative on Incitement to National, Racial, or Religious Hatred' (UN OHCHR 2011 Expert workshop on the prohibition of incitement to national, racial or religious hatred, Vienna, February 2011) <https://www2.ohchr.org/english/issues/opinion/articles1920_iccpr/docs/ContributionsOthers/S.Benesch.doc> accessed 26 December 2022.

[26] Jeroen Temperman, 'Blasphemy versus Incitement: An International Law Perspective' in Christopher S Grenda, Chris Beneke and David Nash (eds), *Profane: Sacrilegious Expression in a Multicultural Age* (University of California Press 2014) 285.

[27] ibid 297–303. According to the Human Rights Committee, "the action advocated through incitement speech does not have to be committed for said speech to amount to a crime. Nevertheless, some degree of risk of harm must be identified". Human Rights Committee, 'Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred That Constitutes Incitement, to Discrimination, Hostility

A problematic aspect of the provision at hand is, nonetheless, represented by the definition of the objects themselves of the inflammatory conduct: that is, the definition of what "discrimination", "hostility", and "violence" are, as well as their relationship with hatred itself. In this respect, the Human Rights Committee has not formally provided any further clarifications. According to an influential study prepared by the NGO Article 19 for the UN, nevertheless, "discrimination" should be understood as "any distinction, exclusion, restriction or preference" based on the membership of a certain category or group of persons, "which has the purpose or effect of nullifying or impairing the recognition, enjoyment or exercise, on an equal footing, of human rights and fundamental freedoms";[28] "violence" is defined as "the intentional use of physical force or power … that either results in or has a high likelihood of resulting in injury, death, psychological harm, maldevelopment, or deprivation";[29] finally, "hostility" is to be distinguished from "hatred" in that, where the latter is a "state of mind" characterized by intense and irrational emotions of opprobrium, enmity and detestation, the former rather implies a "manifested action" which is, therefore, an outward and material projection of hatred itself.[30]

2.2.2.2. Article 4 ICERD

The second essential provision concerning hate speech regulation within the international human rights framework is represented by Article 4 ICERD,[31] which presents at least two fundamental differences from Article 20, paragraph 2. The first difference concerns the protected grounds of discrimination: whereas the ICCPR addressed advocacy of hatred based on "national, racial or religious" grounds, the ICERD ignores the phenomenon of religious hate speech focusing, rather, on "race", "colour", and "ethnic origin". The second difference concerns the reaction against hate speech required by the Convention: indeed, whereas the ICCPR simply obliges states to "prohibit" the conducts described above, leaving to them the choice to resort to civil, administrative, or criminal sanctions,[32] the ICERD compels them to adopt the latter.

---

or Violence' (United Nations 2013) A/HRC/22/17/Add.4 para 29. Similarly, the "Camden Principles" state that "incitement" refers to "statements about national, racial or religious groups which create *an imminent risk* of discrimination, hostility or violence against persons belonging to those groups" (emphasis added). Article 19, 'The Camden Principles on Freedom of Expression and Equality' (April 2009) <https://www.article19.org/data/files/pdfs/standards/the-camden-principles-on-freedom-of-expression-and-equality.pdf> accessed 27 December 2022 principle 12.

[28] Article 19, 'Towards an Interpretation of Article 20 of the ICCPR: Thresholds for the Prohibition of Incitement to Hatred' (Regional expert meeting on article 20, Vienna, 9/02 2010) 7 <https://www2.ohchr.org/english/issues/opinion/articles1920_iccpr/docs/CRP7Callamard.pdf> accessed 27 December 2022.

[29] ibid.

[30] ibid.

[31] With regard to Article 4 ICERD, see Hare (n 21); Patrick Thornberry, *The International Convention on the Elimination of All Forms of Racial Discrimination: A Commentary* (Oxford University Press 2016); Karl Josef Partsch, 'Racial Speech and Human Rights: Article 4 of the Convention on the Elimination of All Forms or Racial Discrimination' in Sandra Coliver (ed), *Hate Speech, Freedom of Expression and Non-Discrimination* (Article 19 1992).

[32] As a matter of fact, the Rabat Plan of Action explicitly states: "Criminal sanctions related to unlawful forms of expression should be seen as last resort measures to be applied only in strictly justifiable situations". Human Rights Committee, 'Rabat Plan of Action' (n 27) 34.

Article 4 ICERD opens with a general condemnation of all propaganda and organizations "which are based on ideas or theories of superiority of one race or group of persons of one colour or ethnic origin, or which attempt to justify or promote racial hatred and discrimination in any form" and requires states parties to adopt immediate and positive measures to eradicate not only acts of discrimination but also all incitement to such discrimination. To reach this end, the provision orders that, with "due regard" to the principles embodied in the Universal Declaration of Human Rights and within Article 5 ICERD,[33] states parties declare as offences punishable by law "all dissemination of ideas based on racial superiority or hatred" and "incitement to racial discrimination, as well as acts of violence or incitement to such acts".[34]

In its General Recommendation No. 35, nonetheless, the Committee on the Elimination of Racial Discrimination specifically addressed the interpretation and scope of application of Article 4.[35] Most notably, the Committee underlined that criminalization should only be resorted to in the most severe cases of racist expressions and should be enacted granting due respect to the principles of legality, proportionality and necessity.[36] Additionally, whereas the Committee had previously attached to Article 4 a strict or absolute liability regime,[37] General Recommendation No. 35 adopted a much more careful approach. Indeed, at least with respect to the conduct of incitement, it explicitly required that states parties take into account the intention of the speaker, as well as "the imminent risk or likelihood that the conduct desired or intended by the speaker will result from the speech in question".[38]

---

[33] "The phrase due regard implies that, in the creation and application of offences, as well as fulfilling the other requirements of article 4, the principles of the Universal Declaration of Human Rights and the rights in article 5 must be given appropriate weight in decision-making processes. The due regard clause has been interpreted by the Committee to apply to human rights and freedoms as a whole, and not simply to freedom of opinion and expression, which should however be borne in mind as the most pertinent reference principle when calibrating the legitimacy of speech restrictions". Committee on the Elimination of Racial Discrimination (n 22) para 19.

[34] Unsurprisingly, many states parties have adopted reservations to the ICERD or have chosen approaches diverging from that of the Committee on the Elimination of Racial Discrimination because of concerns and/or constitutional incompatibilities with art 4 as interpreted by the Committee itself: see Farrior (n 18) 53–60. With respect to the relationship between art 4 ICERD and the United States constitutional system, see Mari J Matsuda, 'Public Response to Racist Speech: Considering the Victim's Story' (1989) 87 Michigan Law Review 2320.

[35] Committee on the Elimination of Racial Discrimination (n 22). With respect to General Recommendation No. 35, see Tarlach McGonagle, 'General Recommendation 35 on Combating Racist Hate Speech' in David Keane and Annapurna Waughray (eds), *Fifty Years of the International Convention on the Elimination of all Forms of Racial Discrimination: A Living Instrument* (Manchester University Press 2017). One of the paramount goals of the Recommendation was to reconcile the ICERD with Articles 19 and 20 ICCPR. In fact, the Convention was previously looked at as an outlier within the field, due to its reliance on the tools of criminal law as a means to fight racism.

[36] Committee on the Elimination of Racial Discrimination (n 22) para 12.

[37] According to a 1983 study of the Committee, "what is penalized … is the mere act of incitement, without any reference to any intention on the part of the offender or the result of such incitement, if any". Committee on the Elimination of Racial Discrimination, 'Positive Measures Designed to Eradicate All Incitement to, or Acts of, Racial Discrimination: Implementation of the International Convention on the Elimination of All Forms of Racial Discrimination, Article 4' (United Nations 1986) CERD/2 para 96.

[38] Committee on the Elimination of Racial Discrimination (n 22) para 16.

However, General Recommendation No. 35 has not fully resolved the debate concerning the criminalization of the conduct of dissemination of ideas based on racial superiority or hatred, with respect to which the threshold is arguably lower pursuant to the text of Article 4. Indeed, the General Recommendation, although recognizing the need that a range of contextual factors be taken into account in order to avoid an excessive restriction of freedom of expression, including the objectives of the speech and thus the intention of the speaker,[39] does not seemingly extend to this conduct the requirement of likelihood or existence of a high risk of impact. This seems to be implicitly confirmed by the General Recommendation where it distinguishes the two conducts, declaring that whereas the provisions of Article 4 on dissemination of ideas "attempt to discourage the flow of racist ideas upstream", those on incitement "address their downstream effects".[40]

The framework resulting from the ICCPR and ICERD has helped give a fundamental impulse on a global scale with respect both to the definition of the phenomenon of hate speech and with respect to the adoption of legal responses to it. First, they offer an insight into the variety of conducts pertaining to the umbrella term "hate speech", most notably by addressing both the case of "incitement" (to discrimination, violence, or hostility) and that of "dissemination of ideas". Second, the two provisions suggest what the response of the law can (and should) be, by requiring states to scale the measures adopted based on the seriousness of the conduct and on a range of contextual features and conditions. Third, both Article 20 ICCPR and Article 4 ICERD attach to the notion of "hate" a nuclear content by identifying, in particular, who the targets of hatred should be in order for hate speech to be relevant under international law: that is, those individuals and groups that are subjected to victimization and discrimination due to a particular identifying feature.

In fact, the international human rights regime on hate speech is rather sectoral and nuclear if compared to the legal regimes actually developed, in the following years, across regional and domestic frameworks. Many jurisdictions adopting hate speech regulations have most notably extended the scope of grounds of discrimination addressed, providing, for example, for measures also encompassing sexist, homophobic, transphobic, or ableist speech.

Nonetheless, the historical role of the ICCPR and ICERD in setting the standards and in propelling state action in this field has been remarkable. Moreover, one of the most relevant merits of the treaties has possibly been the establishment of a direct link between hate speech and the violation of human rights and of the paramount principle of non-

[39] ibid 15. In this respect, however, the UN High Commissioner for Human Rights had expressed a few years before a very different view, arguing that under art 4 ICERD "the dissemination of the idea itself is what attracts sanction without any further requirement about its intent or impact". United Nations High Commissioner for Human Rights, 'Incitement to Racial and Religious Hatred and the Promotion of Tolerance' (United Nations 2006) A/HRC/2/6.

[40] Committee on the Elimination of Racial Discrimination (n 22) para 30. Be that as it may, the question regarding the requirement of the element of likelihood still represents quite an open debate. See, on this point, Article 19, 'Prohibiting Incitement to Discrimination, Hostility or Violence' (2012) <https://www.article19.org/data/files/medialibrary/3548/ARTICLE-19-policy-on-prohibition-to-incitement.pdf> accessed 28 December 2022; Giovanni Ziccardi, *Online Political Hate Speech in Europe: The Rise of New Extremisms* (Edward Elgar Publishing 2020) 36.

discrimination,[41] a connection which has become more and more explicit throughout the subsequent years. Thus, for instance, the UN Special Rapporteur on minority issues highlighted in 2015 that hate speech and incitement to hatred and violence are capable of damaging the "entire social fabric, unity and stability of societies" and that tolerance of and inaction against them "reinforce the subordination of targeted minorities, making them more vulnerable to attacks".[42]

2.2.3. *Hate speech in Europe*

Although the international human rights framework has played a paramount role in the worldwide development of hate speech regulation, regional international and supranational frameworks have also been fundamental – perhaps even more – in orienting state policies at a more de-centralized level. In the context of European countries, both the Council of Europe and the European Union have indeed been extremely influential with respect to this field, as will be underlined throughout the following subsections.

2.2.3.1. The Council of Europe

A variety of sources pertaining to the system of the Council of Europe (CoE) address the issue of hate speech from different perspectives and angles. The most relevant source is inevitably represented by the ECHR,[43] whose provisions have stimulated the ECtHR to take an active role in shaping the way hate speech is dealt with in the Old Continent.

At least two provisions represent the backbone of the development of the Court's case law in this field, that is, Article 10 on freedom of expression and Article 17 on the abuse of rights. To a certain extent, especially in recent years, Article 14 on the right to non-discrimination has also garnered increasing importance.[44] Although recognizing that freedom of expression is applicable also to those ideas and that information "that offend, shock or disturb the State or any sector of the population",[45] the ECtHR has in fact progressively recognized the possibility for states to impose restrictions and limitations upon such freedom when it comes to confronting the phenomenon of hate speech. In this respect, the Court of Strasbourg has developed a two-tiered approach[46] by which, while it

[41] Besides, both the ICCPR and the ICERD address the principle of the right to non-discrimination, respectively at art 26 and art 2. On the relationship between hate speech regulation and the right to (substantive) equality, see *infra*, §2.5.

[42] Rita Izsák, 'Report of the Special Rapporteur on Minority Issues' (United Nations 2015) A/HRC/28/64 para 25.

[43] Convention for the Protection of Human Rights and Fundamental Freedoms 1950.

[44] The role of non-discrimination under the ECHR human rights framework was significantly extended following the adoption of Additional Protocol No. 12 in 2000, prohibiting contracting states from any form of discrimination with respect to the enjoyment of any right recognized by the state (and not only with respect to the fundamental rights and freedoms set directly within the ECHR): see Protocol No. 12 to the Convention for the Protection of Human Rights and Fundamental Freedoms 2000 (ETS No 177) art 1.

[45] *Handyside v the United Kingdom* [1976] ECtHR 5493/72, Series A 24 [49].

[46] David Keane, 'Attacking Hate Speech under Article 17 of the European Convention on Human Rights' (2007) 25 Netherlands Quarterly of Human Rights 641; Mario Oetheimer, 'Protecting Freedom of Expression: The Challenge of Hate Speech in the European Court of Human Rights Case Law Symposium:

generally addresses the matter of the consistency of such measures by applying the three-based test set by Article 10, paragraph 2, of the Convention, it classifies nevertheless the most egregious forms of hate speech as altogether amounting to forms of abuse of freedom of expression under Article 17.

When applying Article 10, the ECtHR, in order to evaluate the consistency of the imposition of formalities, conditions, restrictions or penalties (civil, administrative, and/or criminal) on freedom of expression to combat hate speech, must assess the existence of a prior legislation setting that measure, the pursuit of one of the legitimate aims indicated by the ECHR itself,[47] and the necessity of such a measure in a democratic society (i.e., the respect of the principle of proportionality). In this respect, the ECtHR takes into account a variety of factors, including the purpose of the speaker, the content of the utterance, the context where the utterance is expressed, the identity of the speaker, the composition of the audience, the medium employed, as well as the nature and seriousness of the measure adopted and, therefore, of the state interference upon freedom of expression.[48]

Article 17, conversely, establishes that nothing in the Convention "may be interpreted as implying … any right to engage in any activity or perform any act aimed at the destruction of any of the rights and freedoms set forth [t]herein or at their limitation to a greater extent than is provided for in the Convention". In other words, the provision prohibits "the harmful exercise of a right by its holder in a manner that is manifestly inconsistent with or contrary to the purpose for which such right is granted/designed".[49] In the context of hate speech, this means that there are some cases where utterances are of such a nature so as to constitute, per themselves, a violation of other interests protected by the Convention.

The origins of such an approach can be traced back to 1979, when the then European Commission of Human Rights (ECommHR) delivered a decision of inadmissibility for the case of *Glimmerveen and Hagenbeek v the Netherlands*.[50] In that case, the ECommHR had to assess whether the conviction of the applicant, president of a far-right political

---

Comparative Law of Hate Speech' (2009) 17 Cardozo Journal of International and Comparative Law 427; Hannes Cannie and Dirk Voorhoof, 'The Abuse Clause and Freedom of Expression in the European Human Rights Convention: An Added Value for Democracy and Human Rights Protection?' (2011) 29 Netherlands Quarterly of Human Rights 54; Antoine Buyse, 'Dangerous Expressions: The ECHR, Violence and Free Speech' (2014) 63 International & Comparative Law Quarterly 491; Corrado Caruso, 'L'Hate Speech a Strasburgo: Il Pluralismo Militante Del Sistema Convenzionale' (2017) 4 Quaderni costituzionali 963; Marina Castellaneta, 'La Corte Europea Dei Diritti Umani e l'applicazione Del Principio Dell'abuso Del Diritto Nei Casi Di *Hate Speech*' (2017) 11 Diritti umani e diritto internazionale 745.

[47] I.e., national security; territorial integrity; public safety; prevention of disorder or crime; protection of health or morals; protection of the reputation or rights of others; prevention of the disclosure of information received in confidence; maintenance of the authority and impartiality of the judiciary. Convention for the Protection of Human Rights and Fundamental Freedoms art 10, para 2.

[48] See, among others, Anne Weber, *Manual on Hate Speech* (Council of Europe Publishing 2009).

[49] European Court of Human Rights, 'Guide on Article 17 of the European Convention on Human Rights – Prohibition of Abuse of Rights' (Council of Europe 2022) <https://www.echr.coe.int/Documents/Guide_Art_17_ENG.pdf> accessed 6 April 2023.

[50] *Glimmerveen and Hagenbeek v the Netherlands* [1979] ECommHR 8348/78, 8406/78, 18 Decisions and Reports 187.

party, for the possession – with a view to distribution – of leaflets inciting to racial discrimination was consistent with the ECHR. The Commission concluded that the ideas expressed within those leaflets were not at all compatible with a number of conventional values, namely those enshrined in Article 14 on the prohibition of discrimination, so that the expression of such views amounted to activity prohibited within the meaning of Article 17.[51] Subsequent case law by the ECtHR often referred to the relation between hate speech and abuse of rights, sometimes applying Article 17 as a "guillotine" provision and sometimes using it as a parameter to interpret Article 10 itself.[52]

The resort to such a reference, besides, often relies directly on practical aspects of the single cases at issue. Nonetheless, some common patterns have emerged. Indeed, as noted by the ECtHR itself in the case of *Pavel Ivanov v Russia*, Article 17 has been found to be applicable notably to "statements denying the Holocaust, justifying a pro-Nazi policy, alleging the prosecution of Poles by the Jewish minority and the existence of inequality between them, or linking Muslims with a grave act of terrorism".[53] Thus, for instance, the cases of *Ivanov* and *M'bala M'bala*[54] concerned precisely the application of Article 17 to the case of antisemitic propaganda (and satire), by confirming the conviction, respectively, of the author of a series of articles calling for the exclusion of Jewish people from social life and of a French comedian who had enacted a sketch which, in the opinion of the Court, had taken on the nature of an antisemitic rally rather than of an entertainment show. Similarly, in *Norwood*,[55] the Strasbourg judges held that the applicant's display of a poster associating the image of the Twin Towers in flames with the symbol of a crescent and star in a prohibition sign represented, especially in the immediate wake of 9/11, an abuse of rights.

As for the subject of Holocaust denial, the ECtHR has repeatedly held that the utterance of such ideas is not covered under Article 10 ECHR not only because it inherently represents an attack on the Jewish community but also because it goes against historically

[51] ibid 195–196. As a matter of fact, subsequent case law on antisemitic speech and on Holocaust denial initially took a detour from the reasoning expressed in *Glimmerveen*. For instance, in *X v the Federal Republic of Germany* [1982] ECommHR 9235/81, 29 Decisions and Reports 194, concerning a civil lawsuit against a person who had exposed a noticeboard defining the Holocaust a "Zionistic swindle", the ECommHR chose art 10 as the relevant parameter. Besides, in the two subsequent decisions of *Kühnen v the Federal Republic of Germany* [1988] ECommHR 12194/86, 56 Decisions and Reports 205 and *Remer v Germany* [1995] ECommHR 25096/94, the Commission adopted a hybrid approach, as it found the petitions manifestly ill-founded under art 10, para 2, while interpreting nonetheless that provision in the light of art 17. Thus, although art 17 was taken into account not as a principle capable on its own of determining inadmissibility of the request, the remark that the condemned acts had in fact breached the duties enshrined within that provision was employed as an argument to uphold the satisfaction of the proportionality test. *Kühnen* and *Remer* thus seemingly forecast a subsequent return of the ECommHR and, subsequently, of the ECtHR, towards the original model set in *Glimmerveen*.

[52] In this respect, see the already mentioned decisions of *Kühnen v the Federal Republic of Germany* (n 51); *Remer v Germany* (n 51). See also, *ex multis*, *Molnar v Romania* (dec) [2012] ECtHR 16637/06 [23]; *Behar and Gutman v Bulgaria* [2021] ECtHR 29335/13 [105]; *Bonnet v France* (dec) [2022] ECtHR 35364/19.

[53] *Pavel Ivanov v Russia* (dec) [2007] ECtHR 35222/04 [4].

[54] *M'bala M'bala v France* (dec) [2015] ECtHR 25239/13, ECHR 2015-VIII.

[55] *Norwood v the United Kingdom* (dec) [2004] ECtHR 23131/03, ECHR 2004-XI. However, see, *contra*, *Zemmour v France* [2022] ECtHR 63539/19, where the Court chose to address the case based on art 10 rather than based on art 17.

ascertained facts.[56] Coherently with such a strand of case law, in its 2020 judgment for the case of *Ayoub and others v France*,[57] the ECtHR held that the dissolution of some political movements and associations expressing intensely (and aggressively) xenophobic, antisemitic, and revisionist ideas was consistent with the Convention pursuant to Article 17. In such a case, indeed, the Strasbourg judges concluded that those groups, because their conducts amounted to abuse of rights, were not covered by Article 11 on the right of association as interpreted in the light of Article 10.[58]

Besides, the reference to the concept of "abuse of rights" is, in fact, a characteristic feature distinguishing the European human rights multi-level framework, being also acknowledged and recognized by Article 54 of the Charter of Fundamentals Rights of the European Union (CFREU), and thus represents a further defining aspect distinguishing the approach to hate speech – and, in general, to fundamental rights and liberties – taken on the Eastern side of the Atlantic as opposed to the US. Indeed, the liberal perspective on constitutional freedoms, characterizing the US, is not compatible with the notion itself of abuse of rights.[59]

In addition to the ECHR, the CoE framework has addressed the matter of hate speech also through the drafting of other policy documents and treaty-based instruments,[60] often suggesting that contracting states take positive actions against it. Thus, for instance, the Additional Protocol of 2003 to the 2001 Budapest Convention on Cybercrime[61] obliges contracting states to punish the conduct of distributing or making available through a computer system racist and xenophobic material, defined as "any written material, any image or any other representation of ideas or theories, which advocates, promotes or incites hatred, discrimination or violence" based on the grounds of "race", colour, descent, national or ethnic origin, and religion.[62]

Furthermore, it is important to note that, whereas the Convention does not define nor mention the term and thus leaves to the Strasbourg Court the complex task of identifying

---

[56] *Garaudy v France* (dec) [2003] ECtHR 65831/01, ECHR 2003-IX; *Witzsch v Germany (2)* (dec) [2005] ECtHR 7485/03. See, in this respect, Paolo Lobba, 'Holocaust Denial before the European Court of Human Rights: Evolution of an Exceptional Regime' (2015) 26 European Journal of International Law 237. See, *contra*, *Perinçek v. Switzerland*, where the denial of the Armenian genocide was not considered to be able to trigger *per se* art 17 ECHR: *Perinçek v Switzerland* [2015] ECtHR [GC] 27510/08, ECHR 2015. See, in this regard, Luigi Daniele, 'Disputing the Indisputable: Genocide Denial and Freedom of Expression in *Perincek v. Switzerland*' (2016) 25 Nottingham Law Journal 141.

[57] *Ayoub and others v France* [2020] ECtHR 77400/14, 34532/15, 34550/15.

[58] "La Cour conclut que l'État a pu considérer que les associations requérantes et leurs dirigeants poursuivaient des buts prohibés par l'article 17 et qu'ils avaient abusé de leur liberté d'association, en tant qu'organisation radicale menaçant le processus politique démocratique, en contradiction avec les valeurs de tolérance, de paix sociale et de non-discrimination qui sous-tendent la Convention. Dans leur dissolution, la Cour voit l'expression de décisions prise au regard d'une connaissance approfondie de la situation politique interne et en faveur d'une 'démocratie apte à se défendre' … dans un contexte de persistance et de renforcement du racisme et de l'intolérance en France et en Europe". ibid 138.

[59] See, in this respect, Giovanni Pitruzzella and Oreste Pollicino, *Disinformation and Hate Speech* (Bocconi University Press 2020).

[60] Tarlach McGonagle, 'The Council of Europe against Online Hate Speech: Conundrums and Challenges' (Council of Europe 2013) MCM(2013)005.

[61] Convention on Cybercrime 2001 (ETS No 185).

[62] Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems 2003 (ETS No 189) art 2, para 1.

what constitutes hate speech or not,[63] other CoE policy documents are rather relevant in that they offer a clearer insight into this aspect. First, on 30 October 1997, the Committee of Ministers delivered its Recommendation No. R (97) 20 on "Hate Speech",[64] the Appendix of which contained a series of principles meant to guide the action of CoE states. According to the document, hate speech encompasses

> all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin.[65]

In this respect, the Recommendation presents some significant features if compared with the ICCPR and with the ICERD. First, as to the types of conduct considered, it includes within the notion of hate speech not only those expressions that incite to hatred, but also those that simply spread, promote or justify such hatred. Second, as regards the grounds of discrimination to be addressed, the Recommendation, though focusing specifically on racism, seemingly leaves the door open to an expansion of the scope of the term "hate speech" by featuring an open clause. Indeed, the following years saw an increasingly expansive momentum of the set of protected categories.

Thus, Recommendation No. R (2010) 5 on Measures to Combat Discrimination on Grounds of Sexual Orientation or Gender Identity, adopted on 31 March 2010,[66] included amongst the suggestions to contracting states the adoption of "appropriate measures" against all forms of expression "which may be reasonably understood as likely to produce the effect of inciting, spreading or promoting hatred or other forms of discrimination against lesbian, gay, bisexual and transgender persons" and, most notably, the prohibition of such forms of hate speech.[67]

Furthermore, General Policy Recommendation No. 15 on Combating Hate Speech,[68] adopted in December 2015 by the European Commission against Racism and Intolerance (ECRI) of the Council of Europe, contains an even broader notion of "hate speech", stating that the term

> entails the use of one or more particular forms of expression – namely, the advocacy, promotion or incitement of the denigration, hatred or vilification of a person or group of persons, as well any harassment, insult, negative stereotyping, stigmatization or threat of such person or persons and any justification of all these forms of expression – that is based on a non-exhaustive list of personal characteristics or status that includes "race",

---

[63] Françoise Tulkens, 'When To Say Is To Do: Freedom of Expression and Hate Speech in the Case-Law of the European Court of Human Rights' (Seminar on Human Rights for European Judicial Trainers, Strasbourg, 7 July 2015). In fact, the notion of hate speech within the case law of the ECtHR is not always well-defined.

[64] Committee of Ministers of the Council of Europe, 'Recommendation No. R (97) 20 of the Committee of Ministers to Member States on "Hate Speech"' (Council of Europe 1997) CM/Rec(97)20.

[65] ibid, Appendix, Scope.

[66] Committee of Ministers of the Council of Europe, 'Recommendation No. R (2010) 5 of the Committee of Ministers to Member States on Measures to Combat Discrimination on Grounds of Sexual Orientation or Gender Identity' (Council of Europe 2010) CM/Rec(2010)5.

[67] ibid Appendix, I.B.6.

[68] European Commission against Racism and Intolerance, 'General Policy Recommendation No. 15 on Combating Hate Speech' (Council of Europe 2015) CRI(2016)5.

> colour, language, religion or belief, nationality or national or ethnic origin, as well as descent, age, disability, sex, gender, gender identity and sexual orientation.[69]

ECRI's General Policy Recommendation No. 15 thus extends the notion of hate speech to a wide range of forms of expression, including harassment, insulting, negative stereotyping, stigmatization and threats, and also expands significantly the list of grounds of discrimination to be considered. Additionally, it also clarifies that such a list is "non-exhaustive". ECRI's definition has acquired a paramount importance within the framework of the Council of Europe (and, generally, within the European landscape), thus becoming a fundamental standard for the legal and academic debate on hate speech in the Old Continent.[70]

Thus, coherently, the recent Recommendation No. R (2022) 16 on Combating Hate Speech of the Committee of Ministers[71] declaredly built upon ECRI's General Policy Recommendation No. 15 and adopted a similar definition of hate speech encompassing "all types of expression that incite, promote, spread or justify violence, hatred or discrimination … or that denigrates" persons "by reason of their real or attributed personal characteristics such as 'race', colour, language, religion, nationality, national or ethnic origin, age, disability, sex, gender identity and sexual orientation". Although admittedly, in this case, the list of protected grounds of discrimination is not declared to be "non-exhaustive", the enlargement of the scope of the term "hate speech", especially when compared to Recommendation No. R (97) 20, is quite remarkable.

2.2.3.2. The European Union

Within the European Union, Council Framework Decision 2008/913/JHA[72] represents the most significant piece of legislation concerning hate speech, as it requires Member States of the EU to ensure the criminalization of a range of conducts pertaining to the phenomenon. In this respect, the text of the Framework Decision is in great part inspired by the international standards set by the ICCPR and the ICERD, as it obliges Member States to punish the public incitement to violence or hatred against persons or groups "defined by reference to race, colour, religion, descent or national or ethnic origin".[73]

---

[69] ibid 9.

[70] See, for example, Ziccardi (n 40) 39; Lumi Zuleta and Rasmus Burkal, 'Hate Speech in the Public Online Debate' (The Danish Institute for Human Rights 2017) 17.

[71] Committee of Ministers of the Council of Europe, 'Recommendation No. R (2022) 16 of the Committee of Ministers to Member States on Combating Hate Speech' (Council of Europe 2022) CM/Rec(2022)16.

[72] 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 2008 (OJ L 328/5).

[73] ibid 1, para 1, lett (a). However, the Framework Decision also clarifies that Member States may decide to subject the possibility of punishing such instances of hate speech under the condition that such conducts are "carried out in a manner likely to disturb public order or which is threatening, abusive or insulting", thus leaving to national jurisdictions quite a relevant margin of discretion as regards the limits of criminalization of the phenomenon. Quite interestingly, the 2016 Code of Conduct on Illegal Hate Speech drafted by the European Commission together with a range of IT Companies, refers directly to the Framework Decision 2008/913/JHA, recognizing that "hate speech" should be defined 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

Additionally, the Framework Decision also makes it mandatory to define as criminal offences the conducts of public condonement, denial and gross trivialization of the crimes of genocide, of crimes against humanity, and of war crimes as defined within the Statute of the International Criminal Court[74] as well as of the crimes defined in Article 6 of the Charter of the International Military Tribunal appended to the London Agreement of 1945.[75] The document thus formally aims at introducing within the legal framework of all EU Member States the crime of denialism,[76] quite in line, altogether, with the case law of the ECtHR on Holocaust denial. In such cases, however, the conduct should only be punishable under the condition that the condonement, denial, or gross trivialization concerns crimes that have been established by a final decision of a domestic or international court.[77]

A striking aspect of the Framework Decision is that, evidently, it only encompasses forms of hate speech on grounds of racial and religious discrimination, thus leaving behind many other potential targets. The reason behind this is connected to the rules on EU competences which generally exclude the field of criminal law. In order to be able to enact the Framework Decision, and thus in order to be able to impose upon Member States the duty to make hate speech conducts punishable, its drafters built upon the old Article 29 of the Treaty on the European Union (TEU),[78] conflated today within Article 67 of the Treaty on the Functioning of the European Union (TFEU), pursuant to which the EU "shall endeavour to ensure a high level of security through measures to prevent and combat crime, racism and xenophobia … if necessary, through the approximation of criminal laws".[79] The specific reference to racism and xenophobia thus made it impossible for the EU lawmakers to also include, within the Framework Decision, also other forms of discrimination.[80]

---

to race, colour, religion, descent or national or ethnic origin". Code of Conduct on Countering Illegal Hate Speech Online 2016.

[74] Rome Statute of the International Criminal Court 1998 arts 6–8.

[75] Charter of the International Military Tribunal appended to the Agreement by the government of the United Kingdom of Great Britain and Northern Ireland, the government of the United States of America, the provisional government of the French Republic and the government of the Union of Soviet Socialist Republics for the prosecution and punishment of the major war criminals of the European Axis (UN Treaty Series No 251) 284, art 6.

[76] Paolo Lobba, 'From Introduction to Implementation: First Steps of the EU Framework Decision 2008/913/JHA against Racism and Xenophobia' in Paul Behrens, Nicholas Terry and Olaf Jensen (eds), *Holocaust and Genocide Denial* (Routledge 2017).

[77] Also with regard to this point, the Framework Decision is arguably consistent with the general approach of the ECtHR in this matter. Indeed, the case law of the Strasbourg Court clearly indicates that the abuse clause of art 17 ECHR only applies to those cases of denialism where the genocide or war crime or crime against humanity represents an historically ascertained fact. See, in particular, *Lehideux and Isorni v France* [1998] ECtHR [GC] 24662/94, Reports 1998-VII; *Perinçek v Switzerland* (n 56).

[78] Treaty on the European Union (consolidated version of 2006).

[79] Treaty on the Functioning of the European Union art 67, para 3.

[80] Indeed, as clarified within the document's text itself, also the reference to "religion" should be interpreted restrictively, as it is "intended to cover, *at least,* conduct *which is a pretext* for directing acts against a group of persons or a member of such a group defined by reference to race, colour, descent, or national or ethnic origin" (emphasis added), meaning that Member States, although they may decide to extend the criminal protection required by the Framework Decision also to all cases of religious discrimination, are only required to do so inasmuch as religion constitutes in the case at hand a proxy for racial discrimination. Framework Decision 2008/913/JHA art 1, para 3.

To address such limitations, as well as to develop a more efficient and unitary action against the spread of the phenomenon, especially via the Internet, the European Commission adopted in December 2021 a Communication[81] prompting a Council decision to extend the current list of "EU crimes" under Article 83, paragraph 1, TFEU[82] to include also hate crimes and hate speech. The extension of the scope of action of such a provision would allow the harmonization of Member States' criminal regulation of hate speech, namely through the establishment of minimum rules on its definition and the sanctions connected to it, and would thus open the doors to the possibility, stressed by the Commission, of protecting also people targeted based on other grounds of discrimination, including, in particular, "sex, sexual orientation, age and disability".[83] However, the Council, although a majority expressed its favour towards the proposal in March 2022, has until now failed to adopt the suggested decision unanimously, as lamented by the Parliament's Committee on Civil Liberties, Justice and Home Affairs in the report adopted at the end of November 2023: on this occasion, the Committee suggested *inter alia* to activate the so-called "*passerelle* clause", with a view to making Article 83 "subject to reinforced qualified majority rather than the current required unanimity".[84]

Besides, the width of the scope of the notion of "hate speech" under EU law is sensitively different when moving from the field of criminal law to other fields of the law. For instance, Belavusau has highlighted how the CJEU has delivered a range of decisions under labour law recognizing as illicit pursuant to the equality directives the utterance by employers of public statements disparaging protected categories and declaring the

[81] European Commission, 'Communication from the Commission to the European Parliament and the Council, A More Inclusive and Protective Europe: Extending the List of EU Crimes to Hate Speech and Hate Crime' COM(2021) 777 final. The proposal builds notably on the following documents: European Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. A Union of Equality: Gender Equality Strategy 2020-2025' COM(2020) 152 final; European Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Union of Equality: LGBTIQ Equality Strategy 2020-2025' COM(2020) 698 final; European Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Union of Equality: Strategy for the Rights of Persons with Disabilities 2021-2030' COM(2021) 101 final.

[82] Nina Peršak, 'Criminalising Hate Crime and Hate Speech at EU Level: Extending the List of Eurocrimes Under Article 83(1) TFEU' (2022) 33 Criminal Law Forum 85.

[83] European Commission, 'Communication on Extending the List of EU Crimes to Hate Speech and Hate Crime' (n 81) Annex, recital 6. Quite regrettably, the text of the proposal does not mention either "gender" nor "gender identity", as it refers directly to the grounds of discrimination mentioned within art 19, para 1, TFEU (i.e., "sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation"). Although this would not prevent the EU from taking actions against transphobic and gender-based hate speech, the lack of inclusion of such grounds within recital 6 of the proposal is seemingly an inadequate response to the calls for protection of the LGBTQIA+ community. As highlighted by Peršak, "in line with societal developments and the EU objective of social inclusiveness or fighting social exclusion … the inclusion of gender or gender identity – already employed, for example, by ECRI – in addition to (the more biological category of) sex would be appropriate". Peršak (n 82) 98. Besides, such an approach would be more in line with European Commission, 'LGBTIQ Equality Strategy 2020-2025' (n 81).

[84] European Parliament, 'Report on Extending the List of EU Crimes to Hate Speech and Hate Crime' (2023) 2023/2068(INI) point 7.

intention not to employ members of such categories.[85] Such an approach, according to Belavusau, represents an important tool to fight hate speech also through private, rather than criminal, EU law.[86] With respect to media law, the Audiovisual Media Services Directive (AVMSD),[87] as subsequently amended by Directive (EU) 2018/1808 (AVMSD Refit Directive),[88] requires that both providers of audiovisual media services and providers of online video-sharing platforms put in place measures to reduce the presence and dissemination of content amounting to "incitement to violence or hatred" based on any of the grounds referred to in Article 21 CFREU,[89] the latter expressly prohibiting all discrimination based on "sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation".[90]

2.2.4. Interim *conclusions*

The normative frameworks offer an insight into the inherent issues connected to the definition of the phenomenon itself and, consequently, to the building of regulatory responses at the international and regional level. "Hate speech" is, inherently, an umbrella term encompassing multiple and multi-faceted utterances, which jurisdictions may address in different and often conflicting ways.[91] However, although the drafting of a universally accepted definition of "hate speech" may thus amount to an insurmountable challenge, some patterns can be identified.

The term "speech" can include a wide range of different types of utterances, to which different forms of regulatory response may correspond. Alexander Brown, amongst others, identifies at least ten clusters of regulatory approaches worldwide, including the adoption of measures against incitement to hatred, against the denial of genocide or other crimes against humanity or war crimes, and measures against simple negative stereotyping or stigmatization.[92] Speech, moreover, does not simply include verbal language, but

[85] Case C-54/07, *Centrum voor gelijkheid van kansen en voor racismebestrijding v Firma Feryn NV* [2008] ECLI:EU:C:2008:397; Case C-81/12, *Asociaţia Accept v Consiliul Naţional pentru Combaterea Discriminării* [2013] ECLI:EU:C:2013:275; Case C-507/18, *NH v Associazione Avvocatura per i diritti LGBTI - Rete Lenford* [2020] ECLI:EU:C:2020:289.

[86] Uladzislau Belavusau, 'Fighting Hate Speech through EU Law' (2012) 4 Amsterdam Law Forum 20; Uladzislau Belavusau, 'The *NH* Case: On the "Wings of Words" in EU Anti-Discrimination Law' (2020) 5 European Papers 1001.

[87] 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), OJ L 95/1.

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

[89] AVMSD arts 6, 28b. See, in this respect, Philippe Jougleux, *Facebook and the (EU) Law: How the Social Network Reshaped the Legal Framework* (Springer 2022) 198–201; Oreste Pollicino, Marco Bassini and Giovanni De Gregorio, *Internet Law and Protection of Fundamental Rights* (Bocconi University Press 2022) 147–166.

[90] Charter of Fundamental Rights of the European Union OJ C 364/1 2000 art 21.

[91] As clearly shown by the landmark judicial saga of *LICRA v Yahoo!*. See *infra*, §2.4.2.3.

[92] Brown, *Hate Speech Law* (n 1).

may also include non-verbal forms of expression such as, for instance, the burning of a cross.[93] In this respect, the ample and multi-faceted definition contained within ECRI's General Policy Recommendation No. 15 is, arguably, comprehensive and valid as it identifies an extremely wide range of speech forms and utterances: it is no coincidence if, in presenting its Communication for extending the list of EU crimes, the European Commission still referred to it to describe the phenomenon the proposal aims to confront.[94]

Besides, all these forms of speech are similar in that their ultimate goal or effect consists of conveying, disseminating, and perpetrating the systemic discrimination of people or groups of people defined by specific common features. The intensity of such an intent, as well as the likelihood of that goal being achieved, represent the variables identifying the specific form of hate speech to be addressed and, as such, should be taken into account when developing any regulatory strategy to face the phenomenon. Thus, for instance, serious cases of direct incitement to violence may warrant severe action, including the use of criminal sanctions; whereas simpler cases of negative stereotyping may require more limited (if any) intervention of the law. The catchphrase "hate speech", in this sense, has a sociological, rather than strictly legal, validity, as it includes phenomena which the law must inevitably treat differently.

Be that as it may, the usefulness of such an expression is still relevant for the literature and for policymakers precisely because it captures the essence of all the different forms of speech mentioned above, that is, their role in the perpetration of traditional dynamics of power between categories and "classes" of people.[95] It is precisely in this sense that the term "hate speech" will be intended in the course of the present work, thus focusing on the common character of "hate" rather than upon the multiple possible meanings of "speech".

In this respect, however, a further *caveat* is essential, as the word "hate" can be subjected itself to a multiplicity of different interpretations.[96] In the present context, moreover, the concept of hate is strictly interconnected with that of discrimination. In this sense, the OSCE practical guide to hate crime laws highlighted how in many cases hate crimes and hate speech can be performed by agents who do not, in fact, necessarily feel the sentiment of "hate" and, for this reason, the guide suggests referring to "bias motive", rather than "hate" motive, in order to stress the nature of these phenomena as intrinsically

[93] Alexander Tsesis, 'Dignity and Speech: The Regulation of Hate Speech in a Democracy Articles & Essays' (2009) 44 Wake Forest Law Review 497, 501. See, for example, the notable case of *RAV v City of St. Paul* (n 12).

[94] European Commission, 'Communication on Extending the List of EU Crimes to Hate Speech and Hate Crime' (n 81) 6. See also Patricia Ypma and others, *Study to Support the Preparation of the European Commission's Initiative to Extend the List of EU Crimes in Article 83 of the Treaty on the Functioning of the EU to Hate Speech and Hate Crime: Final Report* (Publications Office of the European Union 2021) 38 <https://data.europa.eu/doi/10.2838/04029> accessed 9 April 2024.

[95] Thus, with specific respect to racist speech, Mari J Matsuda argues that it "is best treated as a sui generis category, presenting an idea so historically untenable, so dangerous, and *so tied to perpetuation of violence and degradation of the very classes of human beings who are least equipped to respond* that it is properly treated as outside the realm of protected discourse" (emphasis added). Matsuda (n 34) 2357.

[96] See, among others, Brown, 'What Is Hate Speech?' (n 1).

discriminatory against very specific protected grounds.[97] This perspective is also apparently welcomed by the European Commission, according to which, for both hate speech and hate crime, "it is the bias motivation that triggers the perpetrator's action".[98]

2.3. The transatlantic debate on hate speech regulation

The phenomenon of hate speech as described in the previous section, i.e., a wide range of speech utterances commonly characterized by their inherent goal of perpetrating forms of discrimination based on certain grounds, has triggered strikingly different legal reactions across the globe. Indeed, the choice to adopt measures restricting and/or punishing hate speech touches directly on the constitutional nerve of any jurisdiction, as it necessarily entails a curtailment of that fundamental pillar of democracy represented by freedom of expression: a dramatic choice which echoes the paradox of tolerance famously described in 1945 by philosopher Karl Popper.[99] In this respect, the clearest dichotomy, at least among Western democracies, is the one between the model of the "tolerant democracy", symbolized by the United States, and that of the "militant democracy", promoted namely by most European countries as well as by the already described EU and ECHR frameworks.[100]

2.3.1. *The liberal approach: the US model of the free marketplace of ideas*

Building on Popper's paradox of tolerance, Bollinger famously described the American constitutional landscape as representing a model of "tolerant society",[101] characterized by

[97] "Taken literally, the phrases 'hate crimes' or 'hate motive' can be misleading. Many crimes which are motivated by hatred are not categorized as hate crimes. Murders, for instance, are often motivated by hatred, but these are not 'hate crimes' unless the victim's protected characteristics were targeted. Conversely, a crime where the perpetrator does not feel 'hate' towards the particular victim can still be considered a hate crime. Hate is a very specific and intense emotional state, which may not properly describe most hate crimes … Rather, the perpetrator is motivated by their stereotypes, preconceived ideas or intolerance towards a particular group of people and the protected characteristic(s) they share". Office for Democratic Institutions and Human Rights, 'Hate Crime Laws: A Practical Report' (2nd edn, OSCE 2022) 17 <https://www.osce.org/files/f/documents/1/4/523940.pdf> accessed 9 January 2023. Although the quoted paragraph is notably focused on hate crimes, the argument also applies, clearly, to hate speech. On the distinction between hate crimes and hate speech, see Walker (n 3) 9.

[98] European Commission, 'Communication on Extending the List of EU Crimes to Hate Speech and Hate Crime' (n 81) 7.

[99] "Unlimited tolerance must lead to the disappearance of tolerance. If we extend unlimited tolerance even to those who are intolerant, if we are not prepared to defend a tolerant society against the onslaught of the intolerant, then the tolerant will be destroyed, and tolerance with them. – In this formulation, I do not imply, for instance, that we should always suppress the utterance of intolerant philosophies … But we should claim the *right* even to suppress them … We should therefore claim, in the name of tolerance, the right not to tolerate the intolerant". Karl Popper, *The Open Society and Its Enemies*, vol I: *The Spell of Plato* (Routledge 1945) 226.

[100] Pitruzzella and Pollicino (n 59) 54.

[101] Lee C Bollinger, *The Tolerant Society: Freedom of Speech and Extremist Speech in America* (Oxford University Press 1988).

such an inherent primacy of the First Amendment[102] that "the free speech idea … remains one of [the US'] foremost *cultural* symbols".[103] Indeed, as US constitutional law rejects any form of "content" or "viewpoint discrimination",[104] meaning any legislation imposing restrictions and limitations or punishing speech based on the content or viewpoint expressed by the speaker, the idea of adopting hate speech regulation is generally considered to be at odds with the First Amendment.[105]

In fact, contemporary US jurisprudence on free speech took its first steps at the end of the 1910s, when the Supreme Court had to deal with a series of cases concerning the Espionage Act 1917. At first, based on the "bad tendency test",[106] the justices had upheld a number of convictions under the statute concerning cases of individuals advocating against the participation of the US in World War I. Subsequently, however, the SCOTUS changed drastically its approach. Thus, in 1919, *Schenck v United States* abandoned the bad tendency test in favour of the "clear and present danger test",[107] while *Abrams v United States* bears one of the most well-known excerpts of US free speech history, that is, Justice Holmes' dissenting opinion containing the notorious metaphor of free speech as a "free marketplace of ideas":

> But when men have realized that time has upset many fighting faiths, they may come to believe even more than they believe the very foundations of their own conduct that the ultimate good desired is better reached by free trade in ideas – that the best test of truth is the power of the thought to get itself accepted in the competition of the market, and that truth is the only ground upon which their wishes safely can be carried out. That, at any rate, is the theory of our Constitution … I think that we should be eternally vigilant against attempts to check the expression of opinions that we loathe and believe to be fraught with death, unless they so imminently threaten immediate interference with the lawful and pressing purposes of the law that an immediate check is required to save the country.[108]

These words, today, are engraved in the American mindset: Holmes' position, originally expressed in dissent, eventually became predominant.

Thus, according to the US constitutional tradition, truth is considered to be more likely to prevail through open discussion than through the adoption of legal measures aiming at curtailing and eradicating falsehoods outright.[109] This applies, of course, to almost any form of "toxic" speech. Clearly, the metaphor of speech as a free marketplace of ideas is

---

[102] "Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances".

[103] Bollinger (n 101) 7. On the cultural and legal significance of free speech in the US, as well as on its uniqueness within the international landscape, see Frederick Schauer, 'The Exceptional First Amendment' in Michael Ignatieff (ed), *American Exceptionalism and Human Rights* (Princeton University Press 2005).

[104] See *supra*, §2.2.1.

[105] See, *ex multis*, *Brandenburg v Ohio* (n 5); *RAV v City of St. Paul* (n 12); *Matal v Tam* 582 US __ (2017).

[106] That is, speech could be subjected to regulation (including criminal prosecution) when it had the tendency to cause or incite illegal activity. For an overview of the development of the bad tendency test under the Espionage Act 1917, see Geoffrey R Stone, 'The Origins of the Bad Tendency Test: Free Speech in Wartime' (2002) 2002 Supreme Court Review 411.

[107] *Schenck v United States* 249 US 47 (1919).

[108] *Abrams v United States* 250 US 616 (1919) 630.

[109] Rosenfeld (n 15) 1534.

inspired by neoclassical economics, according to which, in a market economy, (rational) consumers are drawn to choose the products that are best suitable to their needs and interests so that, at the aggregate level, the best product will end up being the predominant one within the market. Similarly, in the marketplace of ideas, truth and the best opinions, thoughts, and ideologies for society will end up being chosen by the vast majority of (rational) individuals.[110]

Therefore, the response to phenomena like hate speech should not be the adoption of legal measures to restrict its utterance and spread but, rather, the protection of speech itself and the fostering of "more speech". In fact, limiting speech through law would be counterproductive, as it could easily backfire.[111] As argued by Justice Brandeis in his concurring opinion for the case of *Whitney v. California*, "order cannot be secured merely through fear of punishment for its infraction" because "fear breeds repression" and "repression breeds hate": therefore, "the path of safety lies in the opportunity to discuss freely supposed grievances and proposed remedies, and … the fitting remedy for evil counsels is good ones".[112] As a result, it is essential to avoid coercing silence through law, which Brandeis considers to be the expression of "the argument of force in its worst form".[113]

Nonetheless, the mainstream US liberal approach towards hate speech and its relationship with free speech and the First Amendment have been put into question by several American scholars, not fully content with the choice of granting equal protection to all speech, including that expressing, to quote Anthony Lewis, the "thoughts that we hate".[114] These authors, many of whom take a critical race theory approach to hate speech,[115] have most notably highlighted the inherent power dynamics[116] entailed by it and have stressed that such power dynamics often prevent members of minorities or marginalized or discriminated groups from being able to counter racist and hate speech through "more speech":

> The idea that talking back is safe for the victim or educative for the racist simply does not correspond with reality. It ignores the power dimension to racist remarks, forces minorities to run very real risks, and treats a hateful attempt to force the victim outside the human

[110] "Thus ideas and opinions compete with each other, and each of us has the possibility to evaluate them, weigh them in a discussion, and then choose the ones we prefer. As rational consumers of ideas, we will choose the best among many. Just as poor products are expelled from the market due to lack of demand and good products have success determined by the growth of demand for them, good ideas should prevail and bad ideas should be marginalized by market competition". Pitruzzella and Pollicino (n 59) 33.

[111] Rodney A Smolla, 'The Meaning of the "Marketplace of Ideas" in First Amendment Law' (2019) 24 Communication Law and Policy 437, 438.

[112] *Whitney v California* 274 US 357 (1927) 375.

[113] ibid 376.

[114] Anthony Lewis, *Freedom for the Thought That We Hate* (Basic Books 2008).

[115] Richard Delgado and Jean Stefancic, *Critical Race Theory: An Introduction* (3rd edn, New York University Press 2017); Mari J Matsuda and others (eds), *Words That Wound: Critical Race Theory, Assaultive Speech, And The First Amendment* (Westview Press 1993).

[116] Matsuda (n 34); Richard Delgado, 'Words That Wound: A Tort Action for Racial Insults, Epithets, and Name-Calling' (1982) 17 Harvard Civil Rights-Civil Liberties Law Review 133.

> community as an invitation for discussion. Even when successful, talking back is a burden.[117]

Critical race theory authors also stress how hate speech directly affects the psychological and physical well-being of its targets, who are generally at a higher risk of isolation, mental illness and psychosomatic diseases (including depression, high blood pressure, or strokes), and can lead to addiction to alcohol and drugs.[118] Additionally, hate speech can also represent, in their opinion, a danger for society as a whole, namely because discrimination represents itself "a breach of the ideal of egalitarianism, that 'all men are equal' and each person is an equal moral agent".[119]

2.3.2. *The militant approach: the case of Europe*

The liberal and "tolerant" approach of the US with respect to "the thoughts we hate" does not represent a common standard across the world. As described in section 2.1, the prohibition of hate speech is in fact foreseen by international human rights law, both at the global and regional level, and many jurisdictions, such as European countries but also Canada, Australia, Japan, South Africa, as well as many South American states, have indeed enacted forms of restriction of such phenomena.[120]

These jurisdictions thus follow a more "militant" approach, as they put in place measures and limitations to the absolute enjoyment of the fundamental right to free speech and freedom of expression with the goal of actively ensuring the actual protection of core democratic and constitutional principles.[121] In this respect, the European perspective on hate speech represents one of the clearest and most notable examples of such a "militant" strategy and has thus been frequently approached by comparative law as the main term of comparison with US First Amendment jurisprudence on the subject: a comparison which, however, has often had to face the risks of an inherent incommunicability between the two systems,[122] a sort of legal "lost in translation".

The main rationale behind the "militant" approach of Europe can be found first and foremost within the case law of the ECtHR which, in the 2003 judgment of *Gunduz v Turkey*, emphasized that

> tolerance and respect for the equal dignity of all human beings constitute the foundations of a democratic, pluralistic society. That being so, as a matter of principle it may be considered necessary in certain democratic societies to sanction or even prevent all forms of

---

[117] Richard Delgado and Jean Stefancic, *Must We Defend Nazis? Why the First Amendment Should Not Protect Hate Speech and White Supremacy* (New York University Press 2018) 69.

[118] ibid 9–10; Richard Delgado and Jean Stefancic, 'Four Observations about Hate Speech' (2009) 44 Wake Forest Law Review 353, 362.

[119] Delgado (n 116) 140.

[120] Rosenfeld (n 15); Brown and Sinclair (n 1); Spigno (n 1). See *infra*, §4.

[121] Karl Loewenstein, 'Militant Democracy and Fundamental Rights, I' (1937) 31 The American Political Science Review 417.

[122] Eric Heinze, 'Wild-West Cowboys versus Cheese-Eating Surrender Monkeys: Some Problems in Comparative Approaches to Hate Speech' in Ivan Hare and James Weinstein (eds), *Extreme Speech and Democracy* (Oxford University Press 2009); Roger Kiska, 'Hate Speech: A Comparison between the European Court of Human Rights and the United States Supreme Court Jurisprudence' (2012) 25 Regent University Law Review 107.

> expression which spread, incite, promote or justify hatred based on intolerance (including religious intolerance).[123]

According to such reasoning, which the ECtHR has repeatedly confirmed in subsequent case law,[124] hate speech poses a threat to the foundations of paramount constitutional values and principles, namely those connected to the protection of democracy and of pluralism, and for this reason states parties to the Council of Europe may well decide to adopt measures against its spread – including criminal actions – without this constituting a violation of the right to freedom of expression and information as protected by Article 10 ECHR. As a matter of fact, because hate speech affects the possibility for its targets to actively participate in the public debate, it is considered to represent a threat itself to the full protection of the freedom of expression of discriminated groups as well as of the public's right to freedom of information, intended as a right to receive and impart pluralistic and diverse information.[125]

The incompatibility of hate speech with the constitutional framework and the democratic value system of the Council of Europe was recently confirmed by the Committee of Ministers in its already mentioned Recommendation No. R (2022) 16, the Preamble to which argues that

> hate speech negatively affects individuals, groups and societies in a variety of ways and with different degrees of severity, including by instilling fear in and causing humiliation to those it targets and by having a chilling effect on participation in public debate, which is detrimental to democracy.[126]

This approach, besides, is also echoed by EU institutions. Namely, the Commission's Communication on extending the list of EU crimes to hate speech and hate crimes states that these phenomena "are a threat to democratic values, social stability and peace",[127] that they weaken "the mutual understanding and respect for diversity on which pluralistic and democratic societies are built"[128] and that they negate the affected individuals' right to participate in the political or social life, which represents a core principle on which the Union itself is founded.[129]

In this respect, the "militant" viewpoint of the framework of the Council of Europe is in stark contrast to the "tolerant" one of the United States. Whereas the former perceives hate speech as an assault on the democratic tenets of society, including equality and dignity but also freedom of expression itself, the latter considers it as an inevitable facet of the paramount value of free speech and sees any attempt at regulation as an impermissible violation of the First Amendment. In other words, while hate speech regulation on the

---

[123] *Gunduz v Turkey* [2003] ECtHR 35071/97, ECHR 2003-XI [40].

[124] See, *ex multis, Erbakan v Turkey* [2006] ECtHR 59405/00 [56]; *Féret v Belgium* [2009] ECtHR 15615/07 [64]; *Sanchez v France* [2021] ECtHR 45581/15 [84].

[125] Oreste Pollicino, 'Fake News, Internet and Metaphors (to Be Handled Carefully)' (2017) 1 Rivista di Diritto dei Media 23; Pitruzzella and Pollicino (n 59) 91.

[126] Committee of Ministers of the Council of Europe, 'CM/Rec(2022)16' (n 71), Preamble.

[127] European Commission, 'Communication on Extending the List of EU Crimes to Hate Speech and Hate Crime' (n 81) 9.

[128] ibid 1.

[129] ibid 7.

Western side of the Atlantic is treated as a threat to free speech, on the Eastern side of the Ocean it is presented as a tool for the promotion of the equal enjoyment of freedom of expression and information itself in conditions of equality.

The protection of core societal values, of democracy, and of freedom of expression, protected not only as an individual tool of personal autonomy and self-expression but also (and especially) as a collective instrument for the fostering of democracy itself,[130] thus represents the essence of the rationale behind the European restrictive strategy against hate speech and signals a mindset strongly oriented towards the promotion of constitutional-driven principles.

At the same time, however, hate speech is also perceived as being inherently harmful to the personal lives of the individuals affected. Namely, recent case law from the ECtHR, mostly dealing with forms of anti-LGBTQIA+ speech, has increasingly underlined how such forms of expression represent an assault on persons' right to the protection of private and family life as enshrined within Article 8 of the Convention, in conjunction with Article 14 on the prohibition of discrimination.[131] In particular, the Court noted that hateful comments affect the targets' "psychological well-being and dignity",[132] which represent essential components of the right protected by Article 8. Quite interestingly, those cases have even suggested that contracting states may be subject to positive obligations to guarantee that individuals are protected against such assaults[133] and that, while the choice concerning the legal measures to be adopted lies within states' margin of appreciation, "effective deterrence against grave acts where essential aspects of private life are at stake requires efficient criminal-law provisions".[134]

The provision of legal restrictions on hate speech in Europe is thus motivated by the aim of protecting a number of constitutionally relevant values and principles which are considered to be particularly worthy of protection under the ECHR and CFREU fundamental rights systems. These interests pertain both to the collective sphere and to the individual sphere. Hate speech regulation, indeed, aims at preventing the personal harms that can affect the single persons who are contingently targeted by the hateful speech, the harms that affect their group of membership as a whole, and the harms that hate speech produces to society as a whole.[135]

---

[130] On the multiple functions of freedom of expression, both from an individualistic and collective perspective, see among others Rosenfeld (n 15) 1530–1536.

[131] *Beizaras and Levickas v Lithuania* [2020] ECtHR 41288/15; *Association Accept and Others v Romania* [2021] ECtHR 19237/16. With respect to the first case, see Ingrida Milkaite, 'A Picture of a Same-Sex Kiss on Facebook Wreaks Havoc: Beizaras and Levickas v. Lithuania' (*Strasbourg Observers*, 7 February 2020) <https://strasbourgobservers.com/2020/02/07/a-picture-of-a-same-sex-kiss-on-facebook-wreaks-havoc-beizaras-and-levickas-v-lithuania/> accessed 16 January 2023.

[132] *Beizaras and Levickas v Lithuania* (n 131) para 117.

[133] "Positive obligations on the State are inherent in the right to effective respect for private life under Article 8, these obligations may involve the adoption of measures even in the sphere of the relations of individuals between themselves … The Court reiterates its finding that comments that amount to hate speech and incitement to violence, and are thus clearly unlawful on their face, may in principle require the States to take certain positive measures". ibid 110, 125.

[134] ibid 110. Likewise, *Association Accept and Others v Romania* (n 131) para 101.

[135] In this respect, the European approach resembles in many ways that proposed by the critical race theory in the US. See *supra*, §2.3.1.

Although recognizing that freedom of expression also covers those utterances that "offend, shock or disturb",[136] the European viewpoint is that hate speech is not simply a form of expression that "offends" its targets. Rather, hate speech is perceived as a phenomenon that is intrinsically at odds with the democratic functioning of society, as it violates and debases at its core the equal dignity of its victims: an act which has detrimental effects both on single persons and on the social tissue. In this sense, the sensitivity of the Old Continent resonates, curiously, with the words of US author Jeremy Waldron:

> Dignity … is precisely what hate speech laws are designed to protect – not dignity in the sense of any particular level of honor or esteem (or self-esteem), but dignity in the sense of a person's basic entitlement to be regarded as a member of society in good standing, as someone whose membership of a minority group does not disqualify him or her from ordinary social interaction. That is what hate speech attacks, and that is what laws suppressing hate speech aim to protect.[137]

2.4. Hate speech and the Internet

2.4.1. *Free speech and information in the digital age*

Freedom of expression in the twenty-first century has undergone significant transformations following the digital revolution, which has made widely available new technologies "that make it easy to copy, modify, annotate, collate, transmit, and distribute content by storing it in digital form",[138] and following the rise of the "algorithmic society",[139] which features most notably the advent of social media platforms and the increasing use of AI systems as a means of speech governance. The rise and consolidation of the Internet, in particular, has deeply affected the way individuals experience and enjoy freedom of expression and freedom of information as human rights.

The lowering of the costs connected to producing, copying and distributing content have expanded the possibilities for individuals to express and disseminate their ideas, opinions, points of view, and art, by setting aside the issues connected to the traditional scarcity of the means of communication and mass communication. As a result, many commentators saluted the new digital and online sphere as a de-centralizing architecture thanks to which anyone would be given a space and a voice without the need to rely and depend on the will of the private owners of traditional broadcasting infrastructures. This democratizing force of the Internet represents the core of what Yochai Benkler famously defined as the "wealth of networks".[140] As noted by Benkler, information production is

---

[136] *Handyside v the United Kingdom* (n 45).

[137] Jeremy Waldron, *The Harm in Hate Speech* (Harvard University Press 2012) 105.

[138] Jack M Balkin, 'Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society' (2004) 79 New York University Law Review 1, 6.

[139] Jack M Balkin, 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation' (2018) 51 U.C. Davis Law Review 1149. On the impact of artificial intelligence on freedom of expression, see Giovanni De Gregorio and Pietro Dunn, 'Artificial Intelligence and Freedom of Expression' in Alberto Quintavalla and Jeroen Temperman (eds), *Artificial Intelligence and Human Rights* (Oxford University Press 2023).

[140] Yochai Benkler, *The Wealth of Networks: How Social Production Transforms Markets and Freedom* (Yale University Press 2006).

not an exclusive prerogative of professionals anymore, as individual users of the Internet, operating in a more informal and cooperative manner, can today contribute to "peer-produce" information themselves.[141]

This – rather optimistic – viewpoint on the Internet as a tool with an incredibly expansive potential for free speech also emerged in the historical US Supreme Court decision of *Reno v ACLU* (1997).[142] Indeed, in finding the recently enacted Communications Decency Act (CDA),[143] which introduced measures to protect minors from "indecent" and "patently offensive" digital communications (i.e., pornography), unconstitutional under the First Amendment because excessively vague and restrictive, the SCOTUS explicitly recognized the Internet as a new fundamental avenue of the free marketplace of ideas:

> The dramatic expansion of this new marketplace of ideas contradicts the factual basis of this contention. The record demonstrates that the growth of the Internet has been and continues to be phenomenal. As a matter of constitutional tradition, in the absence of evidence to the contrary, we presume that governmental regulation of the content of speech is more likely to interfere with the free exchange of ideas than to encourage it. The interest in encouraging freedom of expression in a democratic society outweighs any theoretical but unproven benefit of censorship.[144]

Nonetheless, it is also worth noticing that such a process of de-centralization of the means of mass communication has in turn led to another type of scarcity, that is, that of the attention of audiences.[145] Indeed, because of the democratization and multiplication of the sources of content, audiences are generally not capable of processing the information overload characterizing the Internet. The inevitable consequence of this process has been the substitution of the old, traditional, gatekeepers of information (newspapers, editors, television broadcasters, radios etc.) with the new "Internet information gatekeepers",[146] that is, precisely, those "large, multinational social media platforms that sit between traditional nation states and ordinary individuals"[147] that select and filter the contents to be included upon and disseminated through their digital infrastructures.

These corporations act as intermediaries between the producers and the receivers of information, structuring the provision of content based on the needs and interests of Internet users themselves.[148] Content moderation, broadly intended as the set of practices and measures adopted to govern the dissemination of speech through a specific Internet

---

[141] "In liberal democracies, the primary effect of the Internet runs through the emergence of the networked information economy. We are seeing the emergence to much greater significance of nonmarket, individual and cooperative peer-production efforts to produce universal intake of observations and opinions about the state of the world and what might and ought to be done about it". ibid 271.

[142] *Reno v American Civil Liberties Union* 521 US 844 (1997).

[143] Communications Decency Act 1996.

[144] *Reno v ACLU* (n 142) 885. On the US approach towards freedom of expression on the Internet, with a focus on the matter of intermediary liability, see *infra*, §4.4.

[145] Balkin, 'Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society' (n 138) 7; Massimo Durante, *Computational Power: The Impact of ICT on Law, Society and Knowledge* (Routledge 2021).

[146] Emily B Laidlaw, *Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibility* (Cambridge University Press 2015).

[147] Balkin, 'Free Speech in the Algorithmic Society' (n 139) 1151.

[148] See, *ex multis*, Nicolas P Suzor, *Lawless: The Secret Rules That Govern Our Digital Lives* (Cambridge University Press 2019).

infrastructure, is in fact the actual commodity offered by platforms, as it allows them to offer "a better experience of all this information and sociality".[149] Besides, from a practical point of view, these actors generally employ algorithms largely based on machine-learning systems,[150] the functioning of which acts as a "black box"[151] for users (and, oftentimes, for programmers themselves).[152] This clearly raises questions about the quality and diversity of information users are exposed to.

Most notably, the migration of the information market to the online infrastructures of privately-owned platforms has led to the consolidation of content management practices, based on the use of AI, that are focused on ensuring the maximization of users' engagement and fidelity towards the platforms themselves, mostly through the profiling of customers and the consequent customization of the information transmitted. This way, the new gatekeepers of information contribute to the construction of a digital space that has been effectively defined by Cass Sunstein as the "Daily Me".[153] However, on the one hand, the engagement-oriented governance of online speech, as well as the "Daily Me", can affect the quality of journalistic sources and of the media and the press in general, inevitably pushed to adjust to the algorithms created by private oligopolists governing the Internet.[154] On the other hand, the customization of online content impacts the possibility for individuals to being truly exposed to pluralistic information, as Internet users end up being locked within echo chambers and filter bubbles.[155] The result of this is also, in turn,

[149] Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press 2018) 13. As highlighted by Sunstein, "in the face of dramatic increases in communications options, there is an omnipresent risk of information overload", so much so that "filtering, often in the form of narrowing, is inevitable in order to avoid overload and impose some order on an overwhelming number of sources of information". Cass R Sunstein, *#Republic: Divided Democracy in the Age of Social Media* (Princeton University Press 2017) 63. Additionally, Wilson and Land note: "Moderation of uncomfortable speech … is part and parcel of the service that social media companies offer". Richard Wilson and Molly Land, 'Hate Speech on Social Media: Content Moderation in Context' (2021) 52 Connecticut Law Review 1029, 1054.

[150] Cambridge Consultants, 'Use of AI in Online Content Moderation' (Ofcom 2019) <https://www.ofcom.org.uk/research-and-data/online-research/online-content-moderation> accessed 30 August 2023; Giovanni Sartor and Andrea Loreggia, 'The Impact of Algorithms for Online Content Filtering or Moderation. "Upload Filters"' (European Parliament 2020) JURI Committee PE 657.101.

[151] Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2015).

[152] Jenna Burrell, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms' (2016) 3 Big Data & Society 2053951715622512. On the use of AI in the context of content moderation (specifically, moderation of hate speech) see *infra*, §5.3.

[153] Sunstein (n 149).

[154] Julia Haas, 'Freedom of the Media and Artificial Intelligence' (Global Conference for Media Freedom, 16 November 2020) <https://www.international.gc.ca/world-monde/assets/pdfs/issues_development-enjeux_developpement/human_rights-droits_homme/policy-orientation-ai-ia-en.pdf> accessed 2 August 2022.

[155] Eli Pariser, *The Filter Bubble: What the Internet Is Hiding From You* (Penguin 2011); Sunstein (n 149); Frank Pasquale, *New Laws of Robotics: Defending Human Expertise in the Age of AI* (The Belknap Press of Harvard University Press 2020); Giovanni Pitruzzella, Oreste Pollicino and Stefano Quintarelli, *Parole e Potere: Libertà d'Espressione, Hate Speech e Fake News* (Egea 2017) 68; Matteo Cinelli and others, 'The Echo Chamber Effect on Social Media' (2021) 118 Proceedings of the National Academy of Sciences of the United States of America e2023301118.

a polarization of the public and political debate, characterized by the rise of disinformation, hate speech, and digital populist narratives.[156]

Moreover, the ambiguous nature of the Internet, and its potential aptness to raise new challenges, and even threats, to human rights and democratic values and principles, has been underscored in many judgments of the ECtHR, which, on the topic of the Internet, has indeed taken a view which is very different from that of the SCOTUS.[157] Though aware that the Internet offers "essential tools for participation in activities and discussions concerning political issues and issues of general interest",[158] the Strasbourg Court has nonetheless underscored how new digital forms of communications might in fact be the source of unprecedented dangers. Thus, for instance, in *KU v Finland*, in finding that Finland had not taken sufficient measures to ensure the protection of the right to private and family life of a minor whose identity had been stolen to create a profile on an adult online dating website, the ECtHR held that freedom of expression on the Internet must in some cases yield to other legitimate imperatives such as the prevention of disorder or crime and the protection of the rights and freedoms of others.[159] In *Editorial Board of Pravoye Delo and Shtekel v Ukraine*, the Strasbourg judges compared the Internet to the printed media, arguing that the former entails a higher risk of harm to the exercise and enjoyment of human rights and freedoms, notably the right to respect for private life, and may, therefore, call for policy measures more restrictive of freedom of expression.[160]

Also, in *Stoll v Switzerland*, the Court's Grand Chamber declared that the Internet increases journalists' duties in providing "reliable and precise" news exactly because, in the contemporary world, where individuals are faced with information overload, compliance with journalistic ethics has become fundamental to guarantee the public's right to being informed.[161] In other words, the Internet has made journalists even more responsible for their essential role as the "public watchdogs".[162]

The concerns of the Strasbourg Court are also shared by the institutions of the EU. Apart from the case law of the CJEU which, aware of the increased risks connected to the digital sphere, has significantly expanded the liability of Internet service providers (ISPs) for the dissemination of illegal information of the Internet since the beginning of the 2010s,[163] the adoption of a number of legislative acts showcases the Union's

---

[156] Oreste Pollicino and Giovanni De Gregorio, 'Constitutional Law in the Algorithmic Society' in Amnon Reichman and others (eds), *Constitutional Challenges in the Algorithmic Society* (Cambridge University Press 2021).

[157] On the different approaches taken by the two courts with respect to the enjoyment of fundamental rights on the Internet, and specifically with respect to the enjoyment of freedom of expression in the digital sphere, see most notably Oreste Pollicino, *Judicial Protection of Fundamental Rights on the Internet: A Road Towards Digital Constitutionalism?* (Hart 2021) 51–98.

[158] *Ahmet Yıldırım v Turkey* [2012] ECtHR 3111/10, ECHR 2012 [54].

[159] *KU v Finland* [2008] ECtHR 2872/02, ECHR 2008 [49].

[160] *Editorial Board of Pravoye Delo and Shtekel v Ukraine* [2011] ECtHR 33014/05, ECHR 2011 [63].

[161] *Stoll v Switzerland* [2007] ECtHR [GC] 69698/01, ECHR 2007-V [103–104].

[162] See, *ex multis*, *Observer and Guardian v the United Kingdom* [1991] ECtHR 13585/88, Series A 216 [59]; *Jersild v Denmark* [1994] ECtHR [GC] 15890/89, Series A 298 [31].

[163] Oreste Pollicino and Giovanni De Gregorio, 'A Constitutional-Driven Change of Heart: ISP Liability and Artificial Intelligence in the Digital Single Market' in Giuliana Ziccardi Capaldo (ed), *The Global*

preoccupations with respect to the possibility of "bad" and harmful information being disseminated through the Internet.[164] The European Commission has itself stressed repeatedly the inherent challenges that the online setting entails for the well-being of democracy. Most notably, although recognizing that the digital revolution has brought more opportunities for civic engagement, making access to information and participation in public life and the democratic debate easier, the Commission has stressed that it has also "opened up new vulnerabilities", affecting *inter alia* the integrity of elections, the protection of free and plural media, and the fight against disinformation and information manipulation.[165]

2.4.2. *Main characters of online hate speech*

The ambiguous nature of the Internet as both an enabler of freedom of expression and as a cause of enhanced risks for the protection of other fundamental rights and democratic values is especially relevant when it comes to the topic of online hate speech. Indeed, in this respect, the European Commission has precisely declared:

> The increase in internet and social media usage has also brought more hate speech online over the years … emotions and vulnerabilities have been increasingly used, including in public debate for political gain, to disseminate racist and xenophobic statements and attacks, amplified in many cases by social media.[166]

The increased risks connected to freedom of expression online are in fact quite relevant when it comes to the dissemination of hate speech content, as the specific characters of

Community Yearbook of International Law and Jurisprudence 2018 (Oxford University Press 2019); Giovanni De Gregorio, 'The Rise of Digital Constitutionalism in the European Union' (2021) 19 International Journal of Constitutional Law 41. With respect to the relationship between online freedom of expression and the protection of intellectual property see, *ex multis*, Joined Cases C-236/08, C-237/08 and C-238/08, *Google France SARL and Google Inc v Louis Vuitton Malletier SA*, *Google France SARL v Viaticum SA and Luteciel SARL* and *Google France SARL v Centre national de recherche en relations humaines (CNRRH) SARL and Others* [2010] ECLI:EU:C:2010:159; Case C-324/09, *L'Oréal SA and Others v eBay International AG and Others* [2011] ECLI:EU:C:2011:474; Case C-70/10, *Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM)* [2011] ECLI:EU:C:2011:771; Case C-360/10, *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV* [2012] ECLI:EU:C:2012:85. With respect to the protection of privacy rights, and specifically the right to be forgotten, see the seminal judgment Case C-131/12, *Google Spain SL and Google Inc v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González* [2014] ECLI:EU:C:2014:317. With regard to the protection of individuals from defamation, see Case C-18/18, *Eva Glawischnig-Piesczek v Facebook Ireland Limited* [2019] ECLI:EU:C:2019:821. For a more in-depth account of the evolution of the approach to ISP liability in the case law of the CJEU, see *infra*, §3.4.2.

[164] See, most notably, AVMSD Refit Directive; Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online, OJ L 172/79; Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act), OJ L 277/1. See more *infra*, §3.4.3.2.

[165] European Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions On the European Democracy Action Plan' COM(2020) 790 final 2.

[166] European Commission, 'Communication on Extending the List of EU Crimes to Hate Speech and Hate Crime' (n 81) 2.

online communication are held to have contributed significantly to its quantitative rise in the Internet ecosystem.[167]

Significant concerns have also been raised by the UN Special Rapporteur on minority issues who, in his 2021 Recommendations on "Hate speech, social media and minorities", defined the scale of hate speech targeting minorities on social media "overwhelming".[168] The diffusion of the hate speech phenomenon over the Internet has reportedly increased significantly especially in the aftermath of the breakout of the COVID-19 pandemic, targeting mainly, but not only, individuals of Asian descent.[169]

Many aspects contribute to making hate speech a particularly challenging phenomenon in the context of the Internet. Amongst these, at least four main features characterizing online hate speech, as opposed to its offline counterpart, have been underscored by literature and have therefore been considered to raise new, significant challenges: permanence, itinerancy, anonymity, and the inherently cross-jurisdictional character of Internet content.[170]

2.4.2.1. Permanence

"Permanence" refers to the ability of hateful content to thrive online and to be easily circulated, also thanks to the use of hyperlinking tools. Permanence often depends significantly on the architecture of platforms involved: thus, for instance, X's conversational structure, based on trending topics, can enable hate speech to spread quickly and widely.[171] Such a feature is especially relevant not only because it enhances the harm

[167] Matteo Cinelli and others, 'Dynamics of Online Hate and Misinformation' (2021) 11 Scientific Reports 22083; European Commission, TIPIK, and Spark Legal Network, *Study to Support the Preparation of the European Commission's Initiative to Extend the List of EU Crimes in Article 83 of the Treaty on the Functioning of the EU to Hate Speech and Hate Crime: Final Report* (Publications Office 2021) <https://data.europa.eu/doi/10.2838/04029> accessed 3 February 2023 Annex VII.

[168] Fernand de Varennes, 'Recommendations Made by the Forum on Minority Issues at Its Thirteenth Session on the Theme "Hate Speech, Social Media and Minorities"' (United Nations 2021) A/HRC/46/58 para 4. As a matter of fact, determining the exact scale of the phenomenon of hate speech on the Internet is not an easy task for various reasons, including the alleged under-reporting of the phenomenon to authorities and the consequent need to refer to platforms' transparency reports which, although giving some precious insights into the dynamics of online hatred, often lack important qualitative data. See Barbora Bukovská, 'The European Commission's Code of Conduct for Countering Illegal Hate Speech Online' (TWG 2019) <https://www.ivir.nl/publicaties/download/Bukovska.pdf> accessed 22 January 2023. As a result, as stressed by Siegel, there is still, at the state of the art, limited literature assessing systematically the scale of the phenomenon of online hate speech: see Alexandra A Siegel, 'Online Hate Speech' in Joshua A Tucker and Nathaniel Persily (eds), *Social Media and Democracy: The State of the Field, Prospects for Reform* (Cambridge University Press 2020). Nonetheless, its diffusion across the Internet has been underscored within the study to support the Commission's proposal of extending art 83(1) TFEU, highlighting the "pan-European" dimension that the issue has taken: see European Commission, TIPIK, and Spark Legal Network (n 167) Annex VII 13.

[169] United Nations, 'Countering COVID-19 Hate Speech' (*United Nations Secretary-General*, 2020) <https://www.un.org/sg/en/node/251827> accessed 15 December 2021; Shivang Agarwal and C Ravindranath Chowdary, 'Combating Hate Speech Using an Adaptive Ensemble Learning Model with a Case Study on COVID-19' (2021) 185 Expert Systems with Applications 115632. Cf. also Meta, 'Community Standards Enforcement: Hate Speech' (*Transparency Center*) <https://transparency.meta.com/reports/community-standards-enforcement/hate-speech/facebook/> accessed 28 April 2024.

[170] Iginio Gagliardone and others, *Countering Online Hate Speech* (UNESCO Publishing 2015).

[171] ibid 13–14.

inflicted on the targeted persons by making it more difficult to remove hate speech contents, thus amplifying significantly their de-humanizing and discriminatory effects,[172] but also because their longevity contributes to the development of what Leiter defined as "cyber-cesspools", that is, "places in cyberspace – chat rooms, websites, blogs, and often the comment sections of blogs – which are devoted in whole or in part to demeaning, harassing, and humiliating individuals: in short, to violating their 'dignity'".[173]

2.4.2.2. Itinerancy

The feature of "itinerancy", instead, consists of the ability of online content to be easily moved across the cyber-space. This way, "when content is removed, it may find expression elsewhere, possibly on the same platform under a different name or on different online spaces": even websites, in case they are shut down, can be immediately reopened by using less stringent web-hosting providers or by reallocating them in countries where hate speech tolerance is much higher.[174]

In this respect, the feature of itinerancy is strictly intertwined with that of permanence, as they both contribute to render the removal of online hate speech content much more difficult. Moreover, itinerancy, combined with permanence, can also contribute to making it easier for "poorly formulated thoughts that would have not found public expression and support in the past" to "land on spaces where they can be visible to large audiences".[175]

2.4.2.3. Anonymity

Anonymity represents both a fundamental asset of online freedom of expression and the cause of significant challenges. Indeed, on the one hand, the possibility of expressing one's views without disclosing one's personal identity represents an important tool of democracy, as it protects the speaker from backlash from private and public actors:[176] thus, anonymity on the Internet can represent an important tool for the enjoyment of freedom of expression especially within illiberal democracies. On the other hand, anonymity

---

[172] "The potential permanency of content made available online is also a relevant consideration when quantifying the nature and extent of the harms caused … Content remains traceable and largely retrievable after its original dissemination to an unprecedented extent when the dissemination takes place online … This means that there is a danger that victims of hate speech will continuously, or at least repeatedly, be confronted by the same instance of hate speech after their original articulation". McGonagle (n 60) 32.

[173] Brian Leiter, 'Cleaning Cyber-Cesspools: Google and Free Speech' in Saul Levmore and Martha C Nussbaum (eds), *The Offensive Internet: Privacy, Speech, and Reputation* (Harvard University Press 2010) 155.

[174] Gagliardone and others (n 170) 14.

[175] ibid.

[176] Dirk Voorhoof, 'Internet and the Right of Anonymity' in Jelena Surculija (ed), *Proceedings of the conference Regulating the Internet, Belgrade, 2010* (Center for Internet Development 2011); Giorgio Resta, 'Anonimato, Responsabilità, Identificazione: Prospettive Di Diritto Comparato' (2014) 2 Il diritto dell'informazione e dell'informatica 171; Giulio Enea Vigevani, 'Anonimato, Responsabilità e Trasparenza Nel Quadro Costituzionale Italiano' (2014) 2 Il diritto dell'informazione e dell'informatica 207; András Koltay, *New Media and Freedom of Expression: Rethinking the Constitutional Foundations of the Public Sphere* (Hart 2019) 202–204. In the US, the right to anonymous speech is considered to be generally protected under the First Amendment following *McIntyre v Ohio Elections Commission* 514 US 334 (1995).

increases the risk of dissemination of illegal and harmful content not only because it makes enforcement of content regulation more burdensome,[177] but also, and perhaps even more so, because it can lead individuals to feel hidden, and therefore secure, when uploading such materials.

In fact, most Internet users do not have the technological tools, nor the know-how to attain full anonymity.[178] Nonetheless, the anonymity perceived by users of the Internet can disinhibit them significantly and thus contribute to the rise of toxic content.[179] In this respect, Citron argues:

> Anonymity frees people to defy social norms. When individuals believe, rightly or wrongly, that their acts won't be attributed to them personally, they become less concerned about social conventions. Research has shown that people tend to ignore social norms when they are hidden in a group or behind a mask. Social psychologists call this condition *deindividuation*. People are more likely to act destructively if they do not perceive the threat of external sanction … People are more inclined to act on prejudices when they think they cannot be identified.[180]

It has correctly been noted that anonymity is, in truth, not always sought by purveyors of hate speech. In fact, many of them actively disclose their identity by making their names and surnames public as their main goal is precisely "to attract attention and consensus", whereas "acting anonymously would not provide recognition in the community in which they are active".[181] This holds true, especially, when hate is used as a political tool to gather followers.

Be that as it may, anonymity contributes sensitively to the increase of spontaneous and/or low-profile forms of hate speech. According to Citron, additionally, the tendency of anonymity to encourage and promote the dissemination of hate speech is often further intensified by the physical separation between speaker and target, as the distance makes the consequences of such utterances seem as if they are remote and affecting indistinct, and thus dehumanized, persons.[182] In other words, anonymity and physical distance, resulting in the invisibility of the target of hate speech and of its consequences, affect the capability of Internet users to exercise sympathy towards their digital interlocutors, thus

---

[177] Christopher D Van Blarcum, 'Internet Hate Speech: The European Framework and the Emerging American Haven' (2005) 62 Washington and Lee Law Review 781, 783.

[178] Graeme Horsman, 'The Challenges Surrounding the Regulation of Anonymous Communication Provision in the United Kingdom' (2016) 56 Computers & Security 151. In fact, the main hurdle that anonymity entails when it comes to the enforcement of hate speech regulation is that, because of the massive amount of unlawful content being posted on the Internet, public resources and finances are often insufficient to prosecute and identify all authors of such content: see Giovanni Ziccardi, *L'Odio Online: Violenza Verbale e Ossessioni in Rete* (Raffaello Cortina 2016) 95.

[179] "The fast sharing of hate speech through the digital word is eased by the online disinhibition effect, as the presumed anonymity on the internet and sense of impunity reduce people's inhibition to commit such offences … The internet provides a channel for increased and easily shared hate speech online. Perpetrators of hate speech online are triggered and disinhibited by a sense of anonymity and impunity on the internet, which increases the risk that they continue commit such offences". European Commission, 'Communication on Extending the List of EU Crimes to Hate Speech and Hate Crime' (n 81) 2, 16.

[180] Danielle Keats Citron, *Hate Crimes in Cyberspace* (Harvard University Press 2014) 58. Cf. Alexander Brown, 'What Is so Special about Online (as Compared to Offline) Hate Speech?' (2018) 18 Ethnicities 297.

[181] Ziccardi (n 40) 202.

[182] Citron (n 180) 59.

favouring psychological processes of "moral disengagement" by which they are able to avoid "the constraint of negative self-sanctions for conduct that violates one's moral standards"[183] and that generally contributes to shaping human beings' moral agency.[184]

2.4.2.4. Cross-jurisdictional nature of online content

The cross-jurisdictional character of online hate speech is problematic for at least two reasons. First, together with permanence and itinerancy, it enhances significantly the negative effects of such content, mainly because it amplifies enormously its reach and thus helps hate groups widen their audiences, especially to countries facing similar political or social situations.[185] Second, it raises important issues as regards international cooperation between jurisdictions. This second aspect is especially problematic precisely because the specific sensitivities of jurisdictions regarding hate speech regulation can be very different, as showcased by the gap described above between the European and US approaches.[186]

In this respect, the notorious *LICRA v Yahoo!* judicial saga is perhaps the most notable and symbolic example of the legal challenges entailed by the ability of Internet content to move across traditional state borders. [187] The episode concerned, namely, the auctioning of Nazi memorabilia upon websites which were stored on Yahoo!'s servers, located in the US, but were accessible worldwide. Because, however, the sale of such items was illegal and punished as a criminal offence under the French Criminal Code, the Paris *Tribunal de Grande Instance* (TGI) issued an order against Yahoo!, requiring it to adopt all means necessary to dissuade from and to block consultation of the abovementioned websites, as well as to pre-emptively inform Internet users of all risks involved in the consultation of such websites.[188]

As the order affected not only the subsidiary French company, but also the mother company, based in California, because of the location of the servers hosting those unlawful auctions, Yahoo!, arguing that the order represented an unacceptable interference on

---

[183] Albert Bandura, *Moral Disengagement: How People Do Harm and Live with Themselves* (Worth Publishers, Macmillan Learning 2016) 1.

[184] See Marta Lamanuzzi, 'Il "Lato Oscuro Della Rete": Odio e Pornografia Non Consensuale. Ruolo e Responsabilità Delle Piattaforme Social Oltre La *Net Neutrality*' (2021) 2 La Legislazione Penale 254, 260.

[185] European Commission, 'Communication on Extending the List of EU Crimes to Hate Speech and Hate Crime' (n 81) 16.

[186] Because of such a gap, Breckheimer lamented the risk of the US attracting "hate mongers" by offering them a "safe haven". Peter J II Breckheimer, 'A Haven for Hate: The Foreign and Domestic Implications of Protecting Internet Hate Speech under the First Amendment' (2001) 75 Southern California Law Review 1493. Similarly Tsesis: "Hate groups have found a haven in the United States for their Internet sites because the Supreme Court has significantly limited the government's ability to prohibit the distribution of racist, provocative materials". Alexander Tsesis, 'Hate in Cyberspace: Regulating Hate Speech on the Internet' (2001) 38 San Diego Law Review 817, 838.

[187] Joel R Reidenberg, 'Yahoo and Democracy on the Internet' (2001) 42 Jurimetrics 261; Marc H Greenberg, 'A Return to Lilliput: The *LICRA v. Yahoo!* Case and the Regulation of Online Content in the World Market' (2003) 18 Berkeley Technology Law Journal 1191; Pollicino, *Judicial protection of fundamental rights on the Internet* (n 157) 37–39; Marco Bassini, *Internet e Libertà Di Espressione: Prospettive Costituzionali e Sovranazionali* (Aracne 2019) 166–167.

[188] TGI Paris (22 May 2000) RG 00/05308, *Ligue internationale contre le racisme et l'antisémitisme et Union des étudiants juifs de France v Yahoo!, Inc et Yahoo! France*.

its First Amendment rights, referred the case to the US District Court for Northern California. The Court concluded indeed that the Parisian decision should not be enforced in the US, noting that "the French order's content and viewpoint-based regulation of the web pages and auction site … clearly would be inconsistent with the First Amendment if mandated by a court in the United States".[189] The District Court's decision was, nonetheless, subsequently reversed by the Court of Appeals of the Ninth Circuit[190] which acknowledged, on the one hand, that the French order would only prevent French users, and not US citizens, from accessing the discussed websites and, on the other hand, that refusing to enforce it would lead the US First Amendment to apply extraterritorially: a result quite controversial and debatable as potentially in contrast with the sovereignty of other countries.[191]

The *LICRA v Yahoo!* episode thus confirms the additional difficulties entailed by the contemporary regulation of online hate speech at the intersection with the issue of digital sovereignty against the Internet landscape.[192]

2.4.3. *The role of algorithmic content moderation and curation*

A significant aspect that requires attention when discussing the phenomenon of online hate speech is also represented by the impact that content governance practices have on its spread.

The set of these practices, from a terminological point of view, can be broadly included within the notion of "content moderation", which is defined, *lato sensu*, as "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse".[193] Such a notion thus entails an extremely wide range of techniques such as the exclusion of unwanted members from the community; norm-setting; and organization of the information flow. However, within this ample group, a distinction may be made between systems of "hard moderation" (moderation *stricto sensu*) and systems of "soft moderation" (curation). Whereas the former consist of decisions concerning the removal of content violating the law or a platform's terms and conditions and, consequently, the measures to be adopted against the accounts violating those rules, the latter govern the way content is presented to users, and thus consist of decisions concerning, rather, the design and architecture of a website, as well as of those techniques put in place to present

[189] *Yahoo! Inc v La Ligue Contre Le Racisme Et L'Antisemitisme* 169 FSupp2d 1181 (NDCal 2001) 1192.

[190] *Yahoo! Inc v La Ligue Contre Le Racisme Et L'Antisemitisme* 379 F3d 1120 (9th Cir 2004); *Yahoo! Inc v La Ligue Contre Le Racisme Et L'Antisemitisme* 433 F3d 1199 (9th Cir 2006).

[191] *Yahoo! Inc v La Ligue Contre Le Racisme Et L'Antisemitisme* (n 190) 1221–1222.

[192] Stephane Couture and Sophie Toupin, 'What Does the Notion of "Sovereignty" Mean When Referring to the Digital?' (2019) 21 New Media & Society 2305; Julia Pohle and Thorsten Thiel, 'Digital Sovereignty' (2020) 9 Internet Policy Review 1; Luciano Floridi, 'The Fight for Digital Sovereignty: What It Is, and Why It Matters, Especially for the EU' (2020) 33 Philosophy & Technology 369.

[193] James Grimmelmann, 'The Virtues of Moderation' (2015) 17 Yale Journal of Law and Technology 42, 47.

users with tailored and customized information.[194] As mentioned above, today these activities are based consistently on the use of algorithmic systems, with important consequences with regard to the governance of hate speech on the Internet.[195]

First, with respect to *stricto sensu* content moderation, these systems are subject to significant margins of error,[196] with a high risk of legitimate content being unwarrantedly removed or, conversely, of hate speech content escaping detection.[197] Margins of error are inherent to any form of online content moderation, but are especially significant when the type of "information bad"[198] to be detected requires a significant amount of contextual elements to be taken into account, as is the case of hate speech.[199] Automated systems of hate speech detection often fail indeed to grapple with the intention behind a post or behind the use of a specific word, and thus wrongly categorize a specific piece of content.[200] Additionally, automated systems can replicate, often involuntarily, human biases and prejudice, often leading to a discriminatory enforcement of moderation strategies, with the collateral effect of removing oftentimes content produced by minority, discriminated, or marginalized communities.[201] This silencing effect, far from contributing to the fight against the phenomenon of hate speech, has precisely the effect of replicating the dynamics of domination and subordination it entails.[202]

Second, the way content is algorithmically curated within the online digital sphere is also extremely relevant. Content curation plays indeed an essential role in determining what is actually seen and what remains hidden on the Internet. This is mostly done, today, through the implementation of recommender systems,[203] which collect and process user data to develop a profile reflecting their interests, likes and dislikes and subsequently

---

[194] Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7 Big Data & Society 2053951719897945, 3; Emma Llansó and others, 'Artificial Intelligence, Content Moderation, and Freedom of Expression' (TWG 2020) <https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf> accessed 13 December 2021. Tim Wu refers to "negative speech control", to indicate the "removing and taking down [of] disfavored, illegal, or banned content, and [the] punishing or removing [of] users", whereas he defines as "affirmative speech control" the act of "choosing what is brought to the attention of the user". Tim Wu, 'Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-Ordering Systems' (2019) 119 Columbia Law Review 2001, 2014.

[195] Pietro Dunn, 'Moderazione Automatizzata e Discriminazione Algoritmica: Il Caso dell'*Hate Speech*' in Laura Abba, Adriana Lazzaroni and Marina Pietrangelo (eds), *La Internet Governance e le Sfide della Trasformazione Digitale* (Editoriale Scientifica 2022); De Gregorio and Dunn (n 139).

[196] Evelyn Douek, 'Governing Online Speech: From "Posts-as-Trumps" to Proportionality and Probability' (2021) 121 Columbia Law Review 759. See more *infra*, §5.4.

[197] Sartor and Loreggia (n 150) 45.

[198] ibid 17.

[199] On the contextual elements to take into account when evaluating when a specific utterance amounts to hate speech, see Weber (n 48).

[200] Machines, indeed, although endowed with extraordinary computational and syntactic capacities, are often still rather dysfunctional as far as semantic understanding is concerned. Luciano Floridi, *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality* (Oxford University Press 2014); Durante (n 145). The understanding of the semantic meaning of online content is particularly complex when it comes to multi-modal forms of expression such as, for example, memes: see *infra*, §5.3.4.1.

[201] See *infra*, §5.3.4.2.

[202] See *infra*, §2.5.1.

[203] Silvia Milano, Mariarosaria Taddeo and Luciano Floridi, 'Recommender Systems and Their Ethical Challenges' (2020) 35 AI & Society 957.

compute a similarity score between that profile and the content items published online so as to be able to suggest relevant content to the Internauts.[204] Automated content curation, however, is driven notably by the purpose of maximizing the engagement of users. Therefore, the content promoted is not necessarily the best content available. Since highly controversial pieces of information tend to trigger people's emotions and, therefore, tend to spark reactions and draw interest, it is often those items that recommender systems tend to offer to users. This is the case, for instance, of disinformation as well as of hate speech. Because these forms of communication are often designed in such a way as to excite the feelings and capture the attention of audiences, recommender systems can often be moved to contribute to their spread.[205] Besides, details about the algorithmic functioning of platforms' recommender systems are usually not disclosed due to proprietary concerns, so that there is a lack of transparency both for the public and for research purposes on this point: the meaning itself of "relevance", that is, the precise methodology associated to the understanding of what is the "best" content for their clients, is in fact often not clear, nor do platforms tend to indicate what they mean by it.[206]

Far from being a mere theoretical issue, the "negative externalities" of the use of – biased – automated content moderation and curation systems have reportedly had significant repercussions in recent years. The most notable – and tragic – example is represented by the genocide of the Rohingya population in Myanmar, which reached its apex in 2016-2017, with the perpetration of military violences against the minority group. In that instance, Facebook, one of the most popular and important sources of information in the country, came under fire for its failure, on the one hand, to detect and remove hate speech utterances inciting to violence against the Rohingya people and, on the other hand, for the automated removal of content posted by Rohingya activists denouncing publicly the violences perpetrated against them.[207] The algorithms used by Facebook were considered, in fact, to be actively responsible for the dissemination and virality of hatred against the persecuted group.[208] This triggered, ultimately, the initiation of coordinated lawsuits

[204] Llansó and others (n 194).

[205] Maria Romana Allegri, *Ubi Social, Ibi Ius: Fondamenti Costituzionali Dei* Social Network *e Profili Giuridici Della Responsabilità Dei* Provider (Franco Angeli 2018) 188; Llansó and others (n 194); Lamanuzzi (n 184) 264.

[206] Llansó and others (n 194). The lack of transparency as regards the processes of customization of content and the targeting of users represents a significant issue at the intersection between the right to freedom of expression and information and the right to privacy and data protection and demonstrate how, in the context of the "algorithmic age", the protection of these fundamental interests has undergone a process of convergence: on this aspect, see namely Giovanni De Gregorio, *Digital Constitutionalism in Europe: Reframing Rights and Powers in the Algorithmic Society* (Cambridge University Press 2022). See also De Gregorio and Dunn (n 139) 81–82.

[207] Suzor (n 148) 128–129. As highlighted by De Gregorio and Stremlau, hate speech detection systems of online social media platforms are more than often not sufficiently (or not at all) trained to deal with non-Western languages, such as African or Asian languages. In these cases, the margin of error increases significantly. Giovanni De Gregorio and Nicole Stremlau, 'Platform Governance at the Periphery: Moderation, Shutdowns and Intervention' in Judit Bayer and others (eds), *Perspectives on Platform Regulation. Concepts and Models of Social Media Governance Across the Globe* (Nomos 2021).

[208] Paul Mozur, 'A Genocide Incited on Facebook, With Posts From Myanmar's Military' *The New York Times* (15 October 2018) <https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-

against Meta aiming at seeking redress for its platform's negligence in combatting the diffusion of dangerous and violent narratives.[209]

2.5. Anti-discrimination perspectives on hate speech: a substantive equality approach

2.5.1. *Hate speech as domination: some takeaways from speech act theory*

As highlighted throughout the previous sections, hate speech, both offline and online, can affect significantly the fundamental rights of target individuals and groups, as well as society as a whole. For instance, hate speech can excite the audiences it reaches and provoke them to perpetrate acts of violence and/or of discrimination against the members of groups traditionally subject to marginalization and victimization. Additionally, hate speech can cause direct effects on the well-being of those people, who may suffer important psychological and psychosomatic damage. More in general, however, hate speech affects the dignity of targeted subjects as human beings, by denying their equal standing in society and relegating them to further conditions of isolation.

In other words, hate speech represents an instrument for perpetuating traditional dynamics of power and domination characterizing the relationship between different segments of the population. In this respect, the philosophical branch of speech act theory, first inaugurated by John Langshaw Austin[210] and by his pupil John Rogers Searle,[211] can offer some relevant insights.[212]

According to Austin, there are cases where utterances can be "performative", meaning that "there is something which is *at the moment of uttering being done by the person*

genocide.html> accessed 26 January 2023; Tom Miles, 'U.N. Investigators Cite Facebook Role in Myanmar Crisis' *Reuters* (12 March 2018) <https://www.reuters.com/article/us-myanmar-rohingya-facebook-idUSKCN1GO2PN> accessed 26 January 2023; Natasha Lomas, 'Meta Urged to Pay Reparations for Facebook's Role in Rohingya Genocide' (*TechCrunch*, 29 September 2022) <https://techcrunch.com/2022/09/29/amnesty-report-facebook-rohingya-reparations/> accessed 26 January 2023. See also the dedicated report by Amnesty International: Amnesty International, 'The Social Atrocity: Meta and the Right to Remedy for the Rohingya' (Amnesty International 2022) ASA 16/5933/2022 <https://www.amnesty.org/en/wp-content/uploads/2022/09/ASA1659332022ENGLISH.pdf> accessed 26 January 2023.

[209] Elizabeth Culliford, 'Rohingya Refugees Sue Facebook for $150 Billion over Myanmar Violence' *Reuters* (8 December 2021) <https://www.reuters.com/world/asia-pacific/rohingya-refugees-sue-facebook-150-billion-over-myanmar-violence-2021-12-07/> accessed 26 January 2023; Dan Milmo, 'Rohingya Sue Facebook for £150bn over Myanmar Genocide' *The Guardian* (6 December 2021) <https://www.theguardian.com/technology/2021/dec/06/rohingya-sue-facebook-myanmar-genocide-us-uk-legal-action-social-media-violence> accessed 26 January 2023.

[210] John L Austin, *How to Do Things with Words: The William James Lectures Delivered at Harvard University in 1955* (Clarendon Press, Oxford University Press 1962).

[211] John R Searle, 'What Is a Speech Act?' in Maurice Black (ed), *Philosophy in America* (Allen and Unwin 1965); John R Searle, 'Austin on Locutionary and Illocutionary Acts' (1968) 77 The Philosophical Review 405; John R Searle, *Expression and Meaning: Studies in the Theory of Speech Acts* (Cambridge University Press 1979).

[212] On the relationship between speech act theory and hate speech, see most notably Alessandro Di Rosa, Hate Speech *e Discriminazione: Un'analisi Performativa Tra Diritti Umani e Teorie Della Libertà* (Mucchi Editore 2020).

uttering".[213] Performative utterances are distinguished from "constative utterances", where no additional act is performed apart from the act of speaking. For example, when saying "He is running", the speaker merely describes a situation which is happening externally and upon which, therefore, they do not actively intervene; whereas, when saying "I apologize", the speaker actually performs an action, that is, that of apologizing. According to Austin, therefore, the distinction between constative and performative utterances equals to that between saying and doing.[214] In other words, there are cases where to speak is, in fact, to do. Besides, Austin actually warns that in most cases constative utterances also entail performative results, so that the actual barrier between speech and action is not always that clear: speech is more often than not an act.[215]

Based on this premise, Austin further develops a taxonomy by distinguishing between locutionary acts, illocutionary acts, and perlocutionary acts:

> We first distinguished a group of things we do in saying something, which … we summed up by saying we perform a *locutionary act*, which is roughly equivalent to uttering a certain sentence with a certain sense and reference, which again is roughly equivalent to 'meaning' in the traditional sense. Second, we said that we also perform *illocutionary acts* such as informing, ordering, warning, undertaking …, i.e. utterances which have a certain (conventional) force. Thirdly, we may also perform *perlocutionary acts*: what we bring about or achieve by saying something, such as convincing, persuading, deterring, and even, say, surprising or misleading.

In practice, locutionary acts consist of the material acts "of" speaking (thus including constative utterances). These acts generally entail a certain conventional force such as to transform the sentence into an act of "doing" something: in this sense, the illocutionary act is precisely that act which is put in place "in" saying something. Finally, perlocutionary acts refer to the material consequences of speaking. For instance, the sentence "Shoot her!" represents a locutionary act in the sense that enunciating it represents *per se* an act; it is at the same time an illocutionary act because it has a conventional force in that it entails the act of ordering someone to do something (conventional consequence), i.e., to shoot a person; finally, it represents a perlocutionary act if that sentence is capable of persuading the person receiving that order, thus leading them to shoot (material consequence).[216]

Searle, in continuing Austin's work, actually criticized the distinction between locutionary acts and illocutionary acts, arguing that "the meaning of the sentence, which is supposed to determine the locutionary act, is already sufficient to fix a certain range of illocutionary act", so that it is not possible to "distinguish between meaning and force, because force is already part of the meaning of the sentence".[217] Conversely, Searle argues that the distinction between the illocutionary and the perlocutionary is essential, as

[213] Austin (n 210) 60.
[214] ibid 47.
[215] "Perhaps indeed there is no great distinction between statements and performative utterances". ibid 52.
[216] John R Searle, 'J.L. Austin (1911-1960)' in Aloysius Patrick Martinich and David Sosa (eds), *A Companion to Analytic Philosophy* (Blackwell 2001) 220–221; Di Rosa (n 212) 120.
[217] Searle, 'J.L. Austin' (n 216) 221.

it is essential for the purposes of identifying the capability of speech of performing *per se* an act "regardless of the subsequent effects on the hearers".[218]

The distinction between illocutionary and perlocutionary acts is not uninfluential in the debate over the harms of hate speech and, consequently, over the regulation of hate speech itself. As a matter of fact, hate speech can amount to a perlocutionary act: for instance, this is the case where the speaker is able to convince their audiences and to push them to physically and materially commit acts of violence or discrimination against persons based on protected features. At the same time, however, it has been argued that hate speech represents an illocutionary act, as it is capable, "in" being uttered, of putting in place a direct act of subordination of the targeted subjects.[219]

In other words, even in those cases where it does not lead audiences to take material actions against the people and groups it aims to attack, hate speech is nonetheless capable of performing an illocutionary act by which its simple existence entails the establishment of a dominator-dominated relationship between social groups and demographics:

> Hate speech is a kind of … oppressive speech: letters "persecute" and "degrade" … with assault-like hate speech; Nazi editorials "incite" and "promote" hatred against Jews with propaganda-like hate speech. But there may be other kinds of … oppressive speech: a court says slaves are "incapable of performing civil acts", are "things, not persons"; a proprietor says "Whites Only". Speech like this is not, or not solely, assault-like or propaganda-like … Its point is to enact, or help enact, a system of … oppression: it authoritatively *ranks* a certain group as inferior, *deprives* them of powers and rights, legitimates discrimination against them. Speech that does these things has, perhaps, the illocutionary force of subordination.[220]

An illocutionary approach to hate speech thus reveals the inherent potential for harmfulness of the phenomenon, regardless of its direct "material" consequences, as the utterance of hate speech discourses constitutes an act of subordination *per se*.[221] Of course, not all hate speech acts are identical, as their conventional force, and therefore their capability to constitute subordination, also depends on a variety of extra-verbal and contextual

---

[218] ibid.

[219] Rae Langton, Sally Haslanger and Luvell Anderson, 'Language and Race' in Gillian Russell and Delia Graff Fara (eds), *The Routledge Companion to Philosophy of Language* (Routledge 2012); Rae Langton, 'Beyond Belief: Pragmatics in Hate Speech and Pornography' in Ishani Maitra and Mary Kate McGowan (eds), *Speech & Harm: Controversies Over Free Speech* (Oxford University Press 2012); Ishani Maitra, 'Subordinating Speech' in Ishani Maitra and Mary Kate McGowan (eds), *Speech & Harm: Controversies Over Free Speech* (Oxford University Press 2012); Rae Langton, 'The Authority of Hate Speech' in John Gardner, Leslie Green and Brian Leiter (eds), *Oxford Studies in Philosophy of Law*, vol 3 (Oxford University Press 2018). According to MacKinnon, a similar role in promoting subordination (of women) is performed by pornography: Catharine A MacKinnon, 'Pornography as Defamation and Discrimination' (1991) 71 Boston University Law Review 793.

[220] Langton, Haslanger and Anderson (n 219) 759.

[221] The direct impact of hate speech as an illocutionary act having the power of perpetrating long-standing relations of domination, subordination, and marginalization is well portrayed by Mari J Matsuda with respect to cross burning, a symbolic gesture typical of the Ku Klux Klan: "All of this is what we see when a cross burns on a suburban lawn. The cross is chosen because it carries with it in an instant 400 years' worth of terror. However we analyze the speech of cross burning the embodied experience of life under a reign of terror must inform the conversation". Mari J Matsuda, 'Dissent in a Crowded Theater' (2019) 72 SMU Law Review 441, 454.

elements, as well as upon the position of authority of the speaker.[222] Nevertheless, the conventional force that hate speech has in structuring society and in building a hierarchy between different demographics represents an essential aspect that the law must take into account.

As has been noted,[223] the distinction between illocutionary and perlocutionary is seemingly reflected by and helps explain the different legal approaches taken, most notably, by the US and by Europe. Whereas, following *Brandenburg v Ohio*, hate speech in the US may be subject to limitation only when it "is directed to inciting or producing imminent lawless action and is likely to incite or produce such an action"[224] so that the focus is, clearly, on the "material" consequences of the speech act (perlocutionary), European jurisdictions tend to extend the scope of hate speech regulation, so as to prohibit more generally speech acts that have the effect, for the simple fact of being uttered, of dehumanizing and attacking the dignity of people as (equal) members of society.

The European approach, therefore, ultimately aims to remedy the structural power dynamics of domination and subordination that characterize the relationship between segments of the population, as these dynamics poison fundamental and constitutional democratic values and affect directly the freedom of the groups targeted by hate speech.[225] It is thus no coincidence that the European Commission, in its Communication on extending the list of EU crimes to hate crimes and hate speech, expressly argued for a common regulatory approach as those phenomena lead "to the devaluation of and threat to the human dignity of a person or a group", namely by negating "their equal footing as members of the society, including their right to participate in the political or social life".[226]

2.5.2. *Substantive equality as a lodestar for hate speech governance*

2.5.2.1. The concept of substantive equality

Interpreting hate speech as an illocutionary act, inherently capable of perpetrating and perpetuating societal dynamics of domination and subordination, and thus identifying hate speech regulation as a possible tool to remedy the resulting imbalances between demographics, leads to conclude that such forms of regulation ultimately aim, at least in the European context, at fostering and protecting the principle of equality, interpreted not so much under its formal acceptation but, rather, under its "substantive" one.

---

[222] Maitra (n 219).

[223] Di Rosa (n 212).

[224] *Brandenburg v Ohio* (n 5) 447.

[225] The term "freedom" is hereby intended, following the neo-republican meaning of the term, as "non-domination" (by others). Such an interpretation of freedom, which is different from the libertarian one focusing on "non-interference" (by the state) is inherently egalitarian, requiring, according to Pettit, the pursuit of structural egalitarianism in society: "For all practical purposes, the goal which we set for ourselves in espousing the republican ideal of freedom is the promotion of equally intense non-domination. The general presumption can be that non-domination will not be furthered unless there is an increase in the equality with which the intensity of non-domination is enjoyed". Philip Pettit, *Republicanism: A Theory of Freedom and Government* (Clarendon Press, Oxford University Press 1997) 116.

[226] European Commission, 'Communication on Extending the List of EU Crimes to Hate Speech and Hate Crime' (n 81) 7.

Whereas the concept of formal equality, grounded in the Aristotelian postulate that – unless there is an objective reason not to do so – similar cases should be treated alike and different cases should be treated differently (equal treatment principle), tends to apply symmetrically to all individuals, irrespective of their gender, ethnic background, sexual orientation, gender identity, age, (dis)ability, etc.,[227] substantive equality takes into account that "disadvantage persists, and this disadvantage tends to be concentrated in groups with a particular status, such as women, people with disabilities, ethnic minorities and others".[228] As a result, substantive equality requires that such disadvantages, inherent within society and often the product of historical forms of discrimination, marginalization, and victimization, are directly confronted by the law, often through asymmetric features.

In fact, the concept of substantive equality is not unitary across jurisdictions nor across literature. In this respect, Sandra Fredman identifies at least three conceptions which may be referred to under the umbrella term of "substantive equality". The first approach focuses on results rather than on treatment (equality of results), meaning that the law, instead of treating all individuals the same way, takes affirmative steps and "preferential" treatments in order to actively distribute benefits in a fairer way: a practical example of such a system is the adoption of quota systems for occupational purposes (e.g., reserving a specific percentage of work positions to women).[229] The second approach is that focusing on "equality of opportunity", meaning that the law, rather than redistributing in a top-down fashion all benefits, should make efforts to ensure that all individuals are put in the same condition by removing pre-existing disadvantages – thus, the metaphor is that of the competitors of a race, who must be all brought to the same starting point: once this goal has been attained, individuals should be treated equally.[230] The third approach focuses on the promotion of the fundamental core of the right to equality, identified in the principle of human dignity.[231]

In opposition to such perspectives, that reduce the notion of substantive equality to one, specific, meaning, Fredman argues for a "four-dimensional concept",[232] which has the advantage of allowing for a more holistic approach in responding to the real social wrongs connected to inequality and addressing its many facets. The first point consists of redressing the disadvantages to which certain groups and categories are subjected, tackling the detrimental consequences attached to a specific social status (redistributive dimension). Second, enforcing substantive equality requires addressing stigma, stereotyping, and humiliation, which have the effect of denying the humanity of targeted individuals: by responding to such actions, the law can protect victims' societal "recognition",

[227] Evelyn Ellis and Philippa Watson, *EU Anti-Discrimination Law* (2nd edn, Oxford University Press 2012) 5–6.

[228] Sandra Fredman, 'Substantive Equality Revisited' (2016) 14 International Journal of Constitutional Law 712, 712.

[229] ibid 720–723.

[230] ibid 723–724.

[231] ibid 724–727.

[232] ibid 727. See also Sandra Fredman, 'Emerging from the Shadows: Substantive Equality and Article 14 of the European Convention on Human Rights' (2016) 16 Human Rights Law Review 273, 281–284.

which "refers to the central importance of inter-personal affirmation to [the] sense of who we are"[233] (recognition dimension). Third, substantive equality should focus on promoting social inclusion and on making sure that disadvantaged individuals are given a political voice (participative dimension). Fourth, substantive equality must respect and accommodate differences among humans, meaning that "existing social structures must be changed to accommodate difference rather than requiring members of out-groups to conform to the dominant norm" (transformative dimension).[234]

In a critique to Fredman's work, Catharine MacKinnon argues that her proposal for a four-dimensional approach to substantive equality fails, in fact, to recognize the "single principle" unifying all those facets of substantive equality: that is, its inherent mission to address social hierarchy as the core principle of social inequality.[235] Indeed, according to MacKinnon,

> The essence of inequality is the misanthropic notion … that some are intrinsically more worthy than others, hence justly belong elevated over them, because of the group of which they are (or are perceived to be) a member. The substance of each inequality, hence the domain in which it operates as a hierarchy, is distinctive to each one, but it is hierarchy that makes it an inequality.[236]

Both positions offer, nonetheless, important insights into how a substantive equality approach can invest the discourse over hate speech governance, both in the online and in the offline dimension. On the one hand, if substantive equality, as stated by MacKinnon, aims at addressing those social inequalities that rest upon historical hierarchies of groups and individuals, and if hate speech as an illocutionary act has the power of creating, structuring, and creating domination and subordination dynamics, then hate speech regulation can (and should) represent a direct instrument to address those social hierarchies.

On the other hand, Fredman's architecture of the principle of substantive equality can offer important indications for the purposes of creating a roadmap for hate speech governance. Most notably, an effective approach to such a phenomenon should focus not only on tackling, and punishing, the stigma, stereotyping, and humiliation hate speech entails (recognition dimension), but should also ensure the full protection and fostering of the fundamental rights – including, namely, freedom of expression and, in general, all fundamental rights and liberties that are conditional for the participation in the public and political life – of targeted groups and categories of people (participative dimension). In other words, a substantive equality approach to hate speech governance equally entails a

---

[233] Fredman (n 228) 730–731.

[234] ibid 733. "For example, working hours have always been patterned on the assumption that childcare takes place outside the labor market. Women who wish to participate in the paid labor market must conform to this paradigm, either by forgoing having children, or leaving their children with paid child-carers or family members. Substantive equality aims to change such institutions so that participative parenting is possible for both mothers and fathers in the labor market. Similarly, the built environment must be adapted to accommodate the needs of disabled people, and dress codes and holidays must accommodate ethnic and religious minorities".

[235] Catharine A MacKinnon, 'Substantive Equality Revisited: A Reply to Sandra Fredman' (2016) 14 International Journal of Constitutional Law 739, 740.

[236] Catharine A MacKinnon, 'Substantive Equality: A Perspective' (2011) 96 Minnesota Law Review 1, 12.

"negative" facet, consisting of the prohibition and punishment of hate speech acts, and a "positive" facet, consisting of the promotion of the voices of minorities and of historically dominated groups.

2.5.2.2. Substantive equality and hate speech in the European multi-level human rights protection system

A substantive equality approach to hate speech governance appears to be quite consistent with today's European multi-level system of human rights protection. However, in this respect, it is important to acknowledge that the European approach to the right to equality and non-discrimination has undergone significant developments since the turn of the new millennium. In fact, at least in the beginning, the right to non-discrimination, both under Article 14 ECHR and under EU law, was rather subject to a formalistic interpretation.

Most notably, equality law scholars lamented for many years the ECtHR's tendency to treat Article 14 as a "Cinderella provision".[237] Indeed, the principle of non-discrimination was only applied *de juncto* with other rights set forth within the Convention. In other words, the right to equality did not have, in the interpretation of the Court, an equal standing with other rights but had, rather, a "parasitic" nature as it simply prohibited discrimination in the enjoyment of other rights.[238] Additionally, the ECtHR was criticized for failing to develop an approach to equality capable of recognizing and considering as a relevant factor the systematic subjection of a certain group to disadvantage, discrimination, exclusion, and oppression.[239]

More recently, however, Strasbourg case law on Article 14 ECHR has significantly evolved, progressively acknowledging the insufficiency of previous approaches to non-discrimination and thus accepting, albeit often implicitly, multiple features resonating with the principle of substantive equality.[240] Namely, the Court has begun accepting that equal treatment before the law may ultimately have the effect of causing forms of indirect discrimination and that, therefore, there may be cases where contracting states are

[237] Rory O'Connell, 'Cinderella Comes to the Ball: Art 14 and the Right to Non-Discrimination in the ECHR' (2009) 29 Legal Studies 211.

[238] Fredman (n 232) 273.

[239] In his partly dissenting opinion for the 2002 judgement of *Anguelova v Bulgaria*, Judge Bonello commented significantly criticized the ECtHR "colour-blind" approach: "I consider it particularly disturbing that the Court, in over fifty years of pertinacious judicial scrutiny, has not, to date, found one single instance of violation of the right to life (Article 2) or the right not to be subjected to torture or to other degrading or inhuman treatment or punishment (Article 3) induced by the race, colour or place of origin of the victim … Frequently and regularly the Court acknowledges that members of vulnerable minorities are deprived of life or subjected to appalling treatment in violation of Article 3; but not once has the Court found that this happens to be linked to their ethnicity. Kurds, coloured, Muslims, Roma and others are again and again killed, tortured or maimed, but the Court is not persuaded that their race, colour, nationality or place of origin has anything to do with it. Misfortunes punctually visit disadvantaged minority groups, but only as the result of well-disposed coincidence" *Anguelova v Bulgaria* [2002] ECtHR 38361/97, ECHR 2002-IV, Partly Dissenting Opinion of Judge Bonello [2-3].

[240] O'Connell (n 237); Fredman (n 232).

required to actively treat individuals differently, taking positive actions to remove societal hurdles, when this is necessary to address situations of objective unfairness.[241]

This progressive shift from a formalistic to a more substantive protection of the right to non-discrimination can also be traced within the case law concerning, namely, anti-LGBTQIA+ hate speech. Indeed, a relatively small but highly significant development emerges if one compares the 2012 judgment of *Vejdeland and others v Sweden*[242] with the already mentioned decisions of *Beizaras and Levickas v Lithuania* and *Association Accept and others v Romania*[243]. In the former case, the ECtHR addressed the legitimacy of the criminal sanctions enacted by Sweden against a group of people who had entered a high school and distributed leaflets – leaving many of them in pupils' lockers – containing serious accusations against homosexual people and associating homosexuality with HIV/AIDS and paedophilia. The Court, on that occasion, had recognized for the first time that criminal persecution of anti-LGBTQIA+ speech could be consistent with Article 10 ECHR. Nevertheless, the decision did not argue in favour of a criminalization of such a phenomenon across states that are party to the Council of Europe.[244]

Conversely, although both *Beizaras and Levickas* and *Association Accept* implicitly recognize states a wide margin of appreciation with respect to such criminalization, they nonetheless stress, as has already been mentioned above,[245] the need to comply with the "positive obligation to secure the effective enjoyment of these rights and freedoms under the Convention", arguing that "this obligation is of particular importance for persons … belonging to minorities, because they are more vulnerable to victimisation".[246] The focus on the actual existence of "positive obligations" to support those groups that are at risk of victimization, rather than on the mere acceptability of implementing measures against forms of hate speech, represents an important step further and, arguably, an implicit recognition of the ultimate goal of hate speech regulation of confronting structural hierarchies of power in society and thus of promoting forms of substantive equality.[247] Such a perspective was later confirmed at the beginning of 2023 in *Valaitis v Lithuania*.[248]

---

[241] For instance, in the case of *Taddeucci and McCall v Italy*, which concerned the refusal of Italian authorities to grant a residence permit for family reasons to a New Zealander citizen who was in a same-sex relationship with an Italian citizen, based on the fact that the two were not married, the Court underlined that, because at the time of the facts Italy did not provide for the recognition of same-sex marriage nor same-sex civil unions, "by deciding to treat homosexual couples … in the same way as heterosexual couples who had not regularized their situation the State infringed the applicants' right not to be discriminated against on grounds of sexual orientation in the enjoyment of their rights under Article 8 of the Convention". *Taddeucci and McCall v Italy* [2016] ECtHR 51362/09 [98].

[242] *Vejdeland and others v Sweden* [2012] ECtHR 1813/07, ECHR 2012.

[243] *Beizaras and Levickas v Lithuania* (n 131); *Association Accept and Others v Romania* (n 131).

[244] Mia Caielli, 'Punire l'omofobia: (Non) Ce Lo Chiede l'Europa. Riflessioni Sulle Incertezze Giurisprudenziali e Normative in Tema Di *Hate Speech*' (2015) 1 GenIUS 54.

[245] See *supra*, §2.3.2.

[246] *Beizaras and Levickas v Lithuania* (n 131) para 108.

[247] Besides, a substantive equality approach, namely in its participative dimension, seemingly emerges in the reference to the chilling effect of hate speech on targeted groups made in Committee of Ministers of the Council of Europe, 'CM/Rec(2022)16' (n 71).

[248] *Valaitis v Lithuania* [2023] ECtHR 39375/19. In that case, however, the Court found that Lithuania had not, in fact, violated the applicant's rights under Article 13 (right to an effective remedy), precisely

Simultaneously, the EU approach to equality and non-discrimination has also undergone significant developments. As is well known, the European Communities were, originally, mainly focused on the promotion of economic and market interests, so that the notion of equality was initially interpreted under a strict formalistic acceptation. Indeed, non-discrimination was inherently seen as being "instrumental for the economic purpose of free movement of people, services, goods, and capital" and thus "primarily serve[d] economic integration and [was] therefore naturally nonprescriptive in substance".[249] Subsequently, however, the EU has turned more and more towards a human rights- and constitutional-oriented paradigm: in particular, the Court of Justice has played an essential role in the evolution of anti-discrimination law.[250]

Thus, for instance, the 1974 judgment of *Sotgiu*[251] already recognized that apparently neutral provisions and rules can have the effect of leading to unfair consequences when applied to different demographics, concluding that rules regarding equality of treatment "forbid not only overt discrimination by reason of nationality but also all covert forms of discrimination which, by the application of other criteria of differentiation, lead in fact to the same result".[252] Hence, the Luxembourg judges introduced the concept of what would later be identified and defined by EU equality directives as "indirect discrimination".[253] As has been noted, the concept itself of indirect discrimination is, at its core, representative of an inherently substantive goal of EU anti-discrimination law as in many cases it may be necessary, in order to avoid liability for indirect discriminatory practices, to actively accommodate group differences, so that "a limited duty of preventive positive action is … implicit in the prohibition of indirect discrimination".[254]

---

because, following the previous holding of *Beizaras and Levickas*, authorities had in fact fulfilled their positive obligation to protect homosexual people from hate speech.

[249] Marc De Vos, 'Substantive Formal Equality in EU Non-Discrimination Law' in Thomas Giegerich (ed), *The European Union as Protector and Promoter of Equality* (Springer 2020) 247.

[250] As a matter of fact, scholars have highlighted that, although the ECtHR has traditionally taken the leading role in the development of human rights principles within Europe, the right to equality and non-discrimination represents an exception, as the CJEU has historically set landmark principles. See Janneke Gerards, 'Non-Discrimination, the European Court of Justice and the European Court of Human Rights: Who Takes the Lead?' in Thomas Giegerich (ed), *The European Union as Protector and Promoter of Equality* (Springer 2020) 138.

[251] Case C-152/73, *Giovanni Maria Sotgiu v Deutsche Bundespost* [1974] ECLI:EU:C:1974:13.

[252] ibid 11.

[253] Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin (Racial Equality Directive), OJ L 180/22 art 2(2)(b); Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation (General Framework for Equal Treatment Directive), OJ L 303/16 art 2(2)(b); Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services (Directive on Gender Equality in Goods and Services), OJ L 373/37 art 2(b); Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast) (Recast Equal Treatment Directive), OJ L 204/23 art 2(1)(b).

[254] Marc De Vos, 'The European Court of Justice and the March towards Substantive Equality in European Union Anti-Discrimination Law' (2020) 20 International Journal of Discrimination and the Law 62, 71. In the case of *Achbita*, an internal rule of a private undertaking, G4S, prohibited the undertaking's employees to wear an Islamic headscarf, leading to the dismissal of Ms Achbita who refused to comply

Moreover, in addition to having helped introduce the notion of indirect discrimination, the CJEU has also addressed the matter of positive actions, explicitly recognized as legitimate under the equality directives.[255] In this respect, in *Milkova*, where the referring court had brought up questions concerning the appropriateness of legislation favouring the employment of disabled people, the Court underscored the consistency of the adoption of such measures with the general goal of the directives themselves, arguing as follows:

> Thus, such a distinction in favour of people with disabilities contributes to achieving the aim of Directive 2000/78 … that is to say, the fight against discrimination, in the present case, based on disability as regards employment and occupation … The purpose of Article 7(2) of Directive 2000/78 is to authorise specific measures aimed at effectively eliminating or reducing actual instances of inequality affecting people with disabilities, which may exist in their social lives and, in particular, their professional lives, *and to achieve substantive, rather than, formal equality by reducing those inequalities*.[256]

Admittedly, as highlighted by De Vos, EU law and the CJEU still lay upon a bedrock of formality with respect to the right to equality. Nevertheless, the CJEU has built upon such a bedrock a significant body of case law through which it has been able to associate with it important substantive equality goals.[257] Thus, overall, the multiplicity of values connected to the principle and the promotion of substantive equality seems to be consistent with CJEU case law and with the EU human rights model.

A substantive equality approach to hate speech governance thus appears to be fully compatible not only with the ECHR framework, but also with that of the EU. With respect to this point, moreover, the policy documents delivered by the European Commission on this matter seem to go precisely in that direction. These include, in particular, the Communication on the European democracy action plan[258] and, even more, the already mentioned Communication on extending the list of EU crimes to hate speech and hate crime. Indeed, the Commission has proven to be especially invested in the need to address the direct silencing effect of hate speech, the utterance of which often results in members of discriminated groups refraining from engaging in public debate precisely because of the

---

with such a rule. As the prohibition was meant to showcase the neutrality of G4S, the CJEU held that the referring court should evaluate if, in the case at hand, it would have been possible for G4S to offer Ms Achbita a post not involving any visual contact with customers. In other words, the Court concluded that the undertaking should have taken, where possible, positive actions to avoid the discriminatory effects of the internal rule. Case C-157/15, *Samira Achbita and Centrum voor gelijkheid van kansen en voor racismebestrijding v G4S Secure Solutions NV* [2017] ECLI:EU:C:2017:203 [43]. Thus Ellis and Watson: "The rule against indirect discrimination ... represents an attempt to provide a greater degree of substantive equality, in particular equality of opportunity". Ellis and Watson (n 227) 142–143.

[255] Racial Equality Directive art 5; General Framework for Equal Treatment Directive art 7; Directive on Gender Equality in Goods and Services art 6; Recast Equal Treatment Directive art 3.

[256] Case C-406/15, *Petya Milkova v Izpalnitelen direktor na Agentsiata za privatizatsia i sledprivatizatsionen kontrol* [2017] ECLI:EU:C:2017:198 [46–47] (emphasis added).

[257] "It is no exaggeration to state that the Court of Justice has retooled formal EU equality law towards substantive equality aims, redefining piecemeal the overarching purpose of EU equality law in the process. Its practical effects in real life may well frustrate the engaged observer or activist, but non-discrimination law can never shape the course of society on its own. What should be acknowledged from a legal perspective, however, is that the pragmatic flexibility of the CJEU in furthering substantive equality goes hand in hand with judicial discretion. Substantive equality stands for outcomes". De Vos (n 254) 82.

[258] European Commission, 'Communication on the European Democracy Action Plan' (n 165).

hatred they are afraid to being subjected to, and on the need to promote, therefore, what Fredman defined as the participative dimension of substantive equality.[259]

2.5.3. *Hate speech governance and substantive equality in the world of bits*

As argued above, the principle and value of substantive equality has become increasingly relevant within the European multi-level human rights protection system, and has also invested, even if implicitly, the debate concerning the governance of hate speech. This has important and significant impact on the regulation and governance of the phenomenon in the context of the Internet.

Regulation of content in the "world of bits"[260] necessarily requires to be adapted to the new triangular scheme characterizing freedom of expression today, where the dynamics of speech regulation do not invest anymore only the relationship between the individual speaker and the state, but have to deal with a third new actor: the private corporate owners of digital infrastructures, that is, ISPs, including namely social media and social network platforms.[261] This has led many jurisdictions to move from "old-school" approaches to speech regulation, generally employing forms of control over individual speakers and publishers – including the adoption of criminal penalties, civil damages, and injunctions against them – to "new-school" techniques, which instead exercise forms of control that are aimed precisely at those private owners of digital infrastructures, often by providing for forms of liability for the presence of unlawful content upon them.[262] As will be highlighted throughout the next Chapter, this has been, precisely, the privileged approach of the EU with respect to online content regulation throughout the last decade and, especially, from the middle 2010s onwards.

Providing for increased forms of legal liability and accountability for ISPs with respect to the presence of illegal and harmful content on the Internet represents, indeed, an essential instrument to promote a safer digital sphere, as, from a technological point of view, these actors are generally better equipped than state authorities for the purposes of enforcing the respect of rules by Internet users. In most cases, thanks to the use of AI systems and algorithms for content moderation, ISPs are even capable of taking proactive and preventive measures against the dissemination of specific items and can thus contribute enormously to limit the existence and spread of unwarranted content. Besides, as mentioned above, ISPs, notably social media and social network platforms, tend to adopt autonomously rules and measures meant specifically to improving users' experiences by protecting them from exposure to unpleasant material.[263]

---

[259] European Commission, 'Communication on Extending the List of EU Crimes to Hate Speech and Hate Crime' (n 81) 7, 9–10.

[260] Oreste Pollicino, 'Judicial Protection of Fundamental Rights in the Transition from the World of Atoms to the World of Bits: The Case of Freedom of Speech' (2019) 25 European Law Journal 155.

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

[262] Jack M Balkin, 'Old-School/New-School Speech Regulation' (2014) 127 Harvard Law Review 2296, 2298.

[263] Gillespie (n 149); Wilson and Land (n 149). See *infra*, §5.2.

Nonetheless, the obvious drawbacks of such an approach should not be ignored. Vesting in practice private corporations with the power to govern individuals' freedom of online expression inherently raises significant questions and concerns as regards the protection of such a fundamental right and pillar of democratic society.[264] In this respect, David Kaye, former Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, has warned against the risks connected to the rise of a "platform law",[265] that is, the set of rules privately defined by the providers of intermediary services (notably, hosting services) within their terms and conditions. Thus, the increased para-constitutional role played by ISPs has led to the recent rise of calls for the development of new forms of "digital constitutionalism".[266]

Furthermore, the use of automated systems for content moderation is still often subject to significant error rates, especially when it comes to targeting forms of "toxic" or "hate" speech, the existence of which generally requires a qualitative assessment of the contextual background of the specific utterance. Notably, research has shown how hate speech detection systems can adversely impact precisely those speakers who are particularly vulnerable to being victimized by such a phenomenon.[267] "New-school" speech regulation systems could have the effect of encouraging significantly the use of these tools and, therefore, of enhancing the risks for errors and biased results.

A substantive equality approach to hate speech governance, however, requires addressing directly these issues. Most notably, the participative dimension of substantive equality, which is aimed precisely at promoting and giving strength to the voices of those individuals that are systematically targeted by hate speech, is incompatible with the silencing impact that automated systems of content detection and moderation can have, paradoxically, precisely on them. Such an inconsistency raises important challenges to the governance of the hate speech phenomenon at the intersection of AI fairness[268] in the context of the European Union and of Europe in general.[269]

---

[264] See *infra*, §3.2.2.

[265] David Kaye, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (Human Rights Council 2018) A/HRC/38/35 para 1.

[266] Nicolas P Suzor, 'Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms' (2018) 4 Social Media + Society 2056305118787812; Suzor (n 148); Pollicino, *Judicial protection of fundamental rights on the Internet* (n 157); Giovanni De Gregorio, 'From Constitutional Freedoms to the Power of the Platforms: Protecting Fundamental Rights Online in the Algorithmic Society' (2019) 11 European Journal of Legal Studies 65; Giovanni De Gregorio, 'Democratising Online Content Moderation: A Constitutional Framework' (2020) 36 Computer Law & Security Review 105374. See more *infra*, §5.5.

[267] European Union Agency for Fundamental Rights, *Bias in Algorithms: Artificial Intelligence and Discrimination* (Publications Office 2022) 49–72 <https://data.europa.eu/doi/10.2811/25847> accessed 3 February 2023.

[268] With respect to the relationship between AI (namely, machine-learning) and the promotion of substantive equality in the context of the EU, see most notably Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-Discrimination Law' (2020) 123 West Virginia Law Review 735.

[269] See *infra*, §5.4.

2.6. Conclusions

The present Chapter has addressed the definition of what hate speech is under the law, focusing both on the international and European framework, and has offered some insights into the rationales connected to the adoption (or rejection) of legal measures aimed at limiting and punishing the utterance and spread of the phenomenon.

Most notably, the previous sections have highlighted what the harms of hate speech can be both in the offline and in the online context and have stressed the deep connection between hate speech and the persistence of societal dynamics of power, domination, discrimination, and subordination. For these reasons, the Chapter has argued that the paramount goal of law, in addressing such a phenomenon, should be that of offering a remedy to such dynamics, namely by promoting and fostering the values of substantive equality (namely under its participative dimension).

The next Chapter, in analysing the developments in EU law as regards the regulation of online content and, especially, the liability of ISPs for the presence and spread of hate speech in the context of the Internet, will take precisely this perspective, arguing for a substantive equality-oriented approach to speech governance.

3.

Hate Speech and Intermediary Liability: The European Framework

Summary: 3.1. Introduction. – 3.2. Internet intermediaries and the triangular model of online speech regulation. – 3.2.1. Internet intermediaries. – 3.2.2. New-school speech regulation and constitutional challenges. – 3.3. Intermediary liability and hate speech: case law from the ECtHR – 3.3.1. The case of *Delfi AS v Estonia*. – 3.3.2. The legacy of *Delfi*. – 3.3.2.1. *MTE and Index.hu v Hungary*. – 3.3.2.2. Subsequent developments. – 3.4. Intermediary liability and hate speech: the framework of the EU. – 3.4.1. Intermediary (non)liability at the turn of the millennium: the e-Commerce Directive. – 3.4.2. Judicial activism of the Luxembourg Court. – 3.4.3. A new phase for the EU. – 3.4.3.1. The "new season" of content moderation regulation. – 3.4.3.2. The new sectoral framework on illegal content. – 3.4.3.3. The Code of Conduct on Illegal Hate Speech. – 3.5. The Digital Services Act. – 3.5.1. The Digital Services Package. – 3.5.2. The rules on the liability of providers of intermediary services. – 3.5.3. The new due diligence obligations for a transparent and safe online environment. – 3.5.3.1. Provisions applicable to all providers of intermediary services. – 3.5.3.2. Provisions applicable to providers of hosting services. – 3.5.3.3. Provisions applicable to providers of online platforms. – 3.5.3.4. Obligations for providers of very large online platforms and of very large online search engines to manage systemic risks. – 3.5.3.5. Standards, codes of conduct, and crisis protocols. – 3.5.4. DSA and hate speech moderation. – 3.5.4.1. Applicability of the DSA to hate speech moderation. – 3.5.4.2. Hate speech moderation and equality in the DSA. – 3.6. Conclusions.

3.1. Introduction

Having explored in Chapter 2 the main features characterizing the phenomenon of hate speech both offline and online, and having thus highlighted the main rationales and goals that may guide the law in regulating, and even banning, hate speech, the present Chapter delves into the evolution of the intermediary liability regime for third-party content within the European context and the effects of such evolution on the governance of hate speech. As anticipated in Chapter 2, recent legislative approaches towards the governance of speech in the digital landscape have turned increasingly towards forms of "new-school" speech regulation, building on the new triadic dynamics of speech on the Internet. Both within the ECHR and EU systems, the legal framework has in this respect undergone important developments since the turn of the millennium.

First of all, Section 3.2 examines the notion of "Internet intermediaries" and further investigates the effects of their rise in the context of the regulation of speech on the Internet, highlighting some concerns and challenges particularly relevant under the lens of constitutional and human rights law.

Section 3.3 addresses major ECtHR case law on intermediary liability with a specific eye on hate speech, focusing namely on the landmark judgment of *Delfi AS v Estonia* (§3.3.1) and its legacy (§3.3.2): the Section discusses, notably, how the ECtHR case law has in this respect established a rather exceptional approach towards intermediary liability for third-party hate speech content as opposed to other types of unlawful material.

Section 3.4, instead, addresses the extraordinary evolution of the EU framework, moving from its original liberal phase – symbolized by the e-Commerce Directive – (§3.4.1), investigating the active role of the CJEU in adapting the interpretation of the Directive in the light of the evolving technological paradigm (§3.4.2), and, finally, offering an overview of the most recent legislative trends characterizing EU policy strategies on content moderation from the end of the 2010s (§3.4.3). In this respect, the work critically assesses the characters of the developing framework, including the challenges arising from a constitutional perspective.

Section 3.5 explores the latest, and possibly most relevant, piece of the developing EU framework on content moderation. The Digital Services Act, finally adopted in October 2022, operates a general and horizontally applicable reform of the system established in 2000 by the e-Commerce Directive. This section, in particular, explores the context of the adoption of the Regulation, part of a twofold package together with the Digital Markets Act (§3.5.1), and describes its content, focusing upon the intermediary liability regime (§3.5.2) and upon the new and complex set of rules on providers' due diligence obligations "for a transparent and safe online environment" (§3.5.3), while also investigating the relationship between the new Act and the challenge of hate speech moderation (§3.5.4). The Digital Services Act represents in many ways a revolutionary piece of legislation complementing the EU body of laws on online speech governance, notably by introducing a "horizontal" framework that sets a baseline discipline for all providers of intermediary services. The Section critically analyses the content of the new Regulation and discusses the implications connected to the adoption of such a legislative model. Moreover, the problematic relationship between the new Regulation and the governance of hate speech represents a core thread of the subsection, highlighting in particular the interpretive issues arising from the adoption of a general and abstract notion of "illegal content" and, therefore, the role that may well be played by complementary sectoral instruments that could be adopted in the future.

Finally, Section 3.6 contains some conclusions and serves as a bridge for the remainder of the work, underlining most notably the challenges represented by the relationship between the DSA and non-EU legal frameworks – including both those of Member States and those of extra-EU jurisdictions – and the need for any tools complementary to the DSA to ensure the promotion of the right to substantive equality in the application of the

new framework, especially *vis-à-vis* the increasing resort to AI systems for content moderation.

3.2. Internet intermediaries and the triangular model of online speech regulation

3.2.1. *Internet intermediaries*

The expression "Internet intermediary" represents an umbrella term encompassing many providers of services. A well-known definition provided by the OECD clarifies that their role is to "bring together or facilitate transactions between third parties on the Internet", namely by "giv[ing] access to, host[ing], transmit[ting] and index[ing] content products and services originated by third parties on the Internet" or by "provid[ing] Internet-based services to third parties".[1] Since a characteristic feature of intermediaries is that of being positioned among a number of parties between whom the specific content, service or product is exchanged, content producers are excluded from such a category – although, clearly, hybrid cases also exist.[2]

At the same time, intermediaries include a variety of actors, such as access providers, data processing and web hosting providers, search engines and online portals, e-commerce intermediaries, Internet payment systems, and "participative networking platforms".[3] Although, admittedly, part of the literature on the subject identifies, from a technical point of view, the notion of access providers with that of Internet service providers, thus considering ISPs as that specific sub-group of Internet intermediaries that allow recipients to access the Internet materially, the present work, in line with existing legal scholarship,[4] tends to refer to ISPs more broadly, as including, within the scope of the term, the generality of Internet intermediaries. "ISPs" and "intermediaries", therefore, will generally be adopted as synonymic terms.

Besides, it is worth mentioning that, in the specific context of EU law, recent legislation has clarified the scope of the relevant terms used. Most notably, EU law refers to "information society services" when dealing with "any service normally provided for remuneration, at a distance, by electronic means and at the individual request of a recipient

[1] Karine Perset, 'The Economic and Social Role of Internet Intermediaries' (OECD 2010) 9 <https://www.oecd-ilibrary.org/content/paper/5kmh79zzs8vb-en> accessed 13 April 2023.

[2] Think, for instance, of a newspaper portal that also offers readers the opportunity to comment on news and exchange views.

[3] Perset (n 1) 9. See also Rebecca MacKinnon and others, *Fostering Freedom Online: The Role of Internet Intermediaries* (UNESCO Publishing 2014) 19–20; Sabine A Einwiller and Sora Kim, 'How Online Content Providers Moderate User-Generated Content to Prevent Harmful Online Communication: An Analysis of Policies and Their Implementation' (2020) 12 Policy & Internet 184, 186.

[4] Oreste Pollicino, Marco Bassini and Giovanni De Gregorio, *Internet Law and Protection of Fundamental Rights* (Bocconi University Press 2022); Mariarosaria Taddeo and Luciano Floridi (eds), *The Responsibilities of Online Service Providers* (Springer 2017).

of services".[5] These include, for example, interpersonal communications services, software applications stores, as well as what the Digital Services Act defines as "intermediary services", that is, mere-conduit, caching, and hosting services. As will be highlighted below, the Digital Services Act also defines "online platforms" as a special category of providers of hosting services having the goal of disseminating the content provided by users.[6]

As highlighted in Chapter 2,[7] intermediaries raise important challenges to the governance of freedom of expression and speech across the digital landscape, as the structure of the services they offer to recipients, as well as their transnational reach, are able to affect – and encourage – the spread and dissemination of illegal and harmful content, including hate speech. A clear example of this is represented by the already discussed *LICRA v Yahoo!* case,[8] where, in fact, Yahoo! did not actively sell Nazi memorabilia but was, rather, the intermediary allowing for the transactions to take place. Therefore, it should not come as a surprise that policies and laws addressing the governance of speech online, including the dissemination of illegal and harmful content such as hate speech, have moved from a paradigm focused on the relationship between the state and the individual to an approach aimed, conversely, at regulating the action of intermediaries themselves. Moreover, depending on the type of service provided, intermediaries play different roles in the dissemination of content and, thus, of hate speech as well. In this respect, among Internet intermediaries, increasing importance has been acquired by hosting providers, offering recipients the possibility to store information provided by them, and, most notably, by social media and social networking sites.

A product of the birth and expansion of the so-called "Web 2.0",[9] social media build on the creation and exchange of user-generated content (UGC) by the recipients of those services themselves.[10] Social networking sites can be seen as representing a sub-set of social media, characterized by the inherent goal of transposing and translating into the digital sphere the relational networks defining society.[11] In other words, the goal of social

[5] Directive (EU) 2015/1535 of the European Parliament and of the Council of 9 September 2015 laying down a procedure for the provision of information in the field of technical regulations and of rules on Information Society services (codification), OJ L 241/1 art 1, para 1, lett (b).

[6] See *infra*, §3.5.2.

[7] See *supra*, §2.4.1, §2.5.3.

[8] See *supra*, §2.4.2.4.

[9] Tim O'Reilly, 'What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software' (2007) 1 Communications & Strategies 17.

[10] Kaplan and Haenlein thus include in the notion of social media a variety of service providers, including: collaborative projects which enable the joint and simultaneous creation of content by many end-users (e.g., Wikipedia); blogs; content communities, whose goal is to share content between users (e.g., YouTube); social networking sites, which allow users to connect through the creation of personal information profiles accessible to friends and colleagues; virtual game worlds; virtual social worlds. See Andreas M Kaplan and Michael Haenlein, 'Users of the World, Unite! The Challenges and Opportunities of Social Media' (2010) 53 Business Horizons 59.

[11] Indeed, the concept itself of "social network" finds its origins in the work of Australian anthropologist John Arundel Barnes who, in 1954, argued: "Each person is, as it were, in touch with a number of other people, some of whom are directly in touch with each other and some of whom are not. Similarly each person has a number of friends, and these friends have their own friends; some of any one's person's friends

networking sites is to host and favour digital social bonds, namely through web-based services allowing people to build a public or semi-public online profile, to articulate a list of other users with whom they share a connection, and to extend their connections with other individuals that are party to the system.[12]

3.2.2. *New-school speech regulation and constitutional challenges*

The specific features characterizing social media in general and social networking sites in particular, aimed notably at hosting and disseminating content generated by their recipients across the Internet, render them particularly exposed to the risk of enhancing the presence of unwarranted material within the digital landscape.[13] Such a risk is further augmented by the reliance – which is growing exponentially – upon automated and algorithm-driven strategies of content moderation and content curation. The latter being oriented towards the maximization of the capture of recipients' interests and attention, they might in fact end up bringing to the fore highly controversial content, which is more likely to trigger debates and, therefore, user engagement.[14]

As a result, regulation in Europe has increasingly become focused upon vesting such intermediaries with duties and responsibilities aimed at promoting a safer online space. These new-school forms of speech regulation[15] build upon the specific features characterizing contemporary speech governance dynamics, as opposed to older, traditional models of regulation of freedom of expression. As a matter of fact, whereas "the twentieth century featured a *dualist* or *dyadic* system of speech regulation", where the relevant players were the nation-states and the speakers, the latter being subjected to the rules set by the former, "the twenty-first-century model is *pluralist*, with many different players" and has thus been compared by Jack Balkin to a triangle whose new, third corner consists of Internet-infrastructure companies.[16]

---

know each other, others do not. I find it convenient to talk of a social field of this kind as a *network*. The image I have is of a set of points some of which are joined by lines … We can of course think of the whole of social life as generating a network of this kind". John Arundel Barnes, 'Class and Committees in a Norwegian Island Parish' (1954) 7 Human Relations 39, 43.

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

[13] With respect to hate speech, for instance, Citron and Norton observed as early as in 2011: "The greatest increase in digital hate has occurred on social media sites. Examples include the *How to Kill a Beaner* video posted on YouTube, which allowed players to kill Latinos while shouting racial slurs, and the Facebook group *Kick a Ginger Day,* which inspired physical attacks on students with red hair. Facebook has hosted groups such as *Hitting Women, Holocaust Is a Holohoax,* and *Join if you hate homosexuals*". Danielle Keats Citron and Helen Norton, 'Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age' (2011) 91 Boston University Law Review 1435, 1437.

[14] Emma Llansó and others, 'Artificial Intelligence, Content Moderation, and Freedom of Expression' (TWG 2020) 15 <https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf> accessed 13 December 2021.

[15] See *supra*, §2.5.3.

[16] Jack M Balkin, 'Free Speech Is a Triangle' (2018) 118 Columbia Law Review 2011, 2013–2014. See also, with specific respect to the case of *Delfi AS v Estonia* (see *infra*, §3.3.1.), Robert Alexy, 'The Responsibility of Internet Portal Providers for Readers' Comments. Argumentation and Balancing in the Case of Delfi AS v. Estonia' in María Elósegui, Alina Miron and Iulia Motoc (eds), *The Rule of Law in Europe: Recent Challenges and Judicial Responses* (Springer 2021) 207.

In the contemporary context, traditional tools have indeed become insufficient when it comes to the enforcement of public strategies. Conversely, the private owners of digital infrastructures where speech flourishes today, through their moderation practices powered by their technical and economic capacity as well as by the availability of large quantities of data at their disposal, are in general better positioned to actively control, govern, and regulate the uploading of content to the Internet. Therefore, online platforms have been famously described as the "new governors" of speech in the digital landscape.[17] Building on this, governments, especially in Europe, have increasingly begun to adopt forms of public-private cooperation or co-optation with a view to pushing intermediaries to do their bidding as much as possible.[18]

Clearly, the adoption of such strategies for the governance of online speech is not without consequences from a constitutional and human rights law perspective, namely because it directly entails the result of vesting private actors with the task of supervising over the freedom of expression of the recipients of their services. In particular, new-school strategies often foster forms of "collateral censorship", which arises "whenever a nation-state puts pressure on digital-infrastructure companies to block, take down, and censor content by end users".[19] In many cases, this entails a significant drawback for the protection of freedom of expression, as intermediaries may choose to adopt moderation strategies that are particularly stringent so as to avoid any risk of liability for UGC and third-party content. In general, the delegation to private actors of speech surveillance tasks represents a significant challenge to the promise of a democracy-oriented Internet. Requiring intermediaries to "patrol" the Internet, indeed, implies giving them the duty – and power – to strike a balance between the (constitutional) interests at stake, namely freedom of expression, on the one hand, and the pursuit of public policies, on the other hand. Such a private enforcement of public interests – which is, besides, operated in practice through the adoption and implementation (also through automated systems of moderation) of private terms and conditions of service – inevitably conflates with private business-oriented interests.

Issues regarding the human rights sustainability of platform governance models, especially in the light of the principles of democracy and legitimacy,[20] are in many cases independent of the adoption of forms of new-school speech regulation models, as platforms tend to adopt governance and content moderation strategies irrespective of the imposition upon them of regulatory obligations. Content moderation is, in fact, an integral part of the

[17] Kate Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech' (2017) 131 Harvard Law Review 1598.

[18] Balkin (n 16) 2019–2021. Thus, "new-school regulation often emphasizes ex ante prevention rather than ex post punishment, and complicated forms of public/private cooperation": Jack M Balkin, 'Old-School/New-School Speech Regulation' (2014) 127 Harvard Law Review 2296, 2306.

[19] Balkin (n 16) 2030.

[20] Nicolas P Suzor, 'Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms' (2018) 4 Social Media + Society 2056305118787812; Hannah Bloch-Wehba, 'Global Platform Governance: Private Power in the Shadow of the State' (2019) 72 SMU Law Review 27; Blayne Haggart and Clara Iglesias Keller, 'Democratic Legitimacy in Global Platform Governance' (2021) 45 Telecommunications Policy 102152.

service offered to users and, therefore, an integral part of those intermediaries' business models.[21] In fact, in some cases, platforms have tentatively strived to address themselves such issues, as demonstrated for example by Meta's choice to create an (at least allegedly) independent Oversight Board, mainly composed of notorious international academics, activists, and politicians specialized in digital rights and freedom of expression, whose main task is that of ensuring the adequacy of Meta platforms' content moderation practices and their consistency with fundamental democratic principles.[22]

Nonetheless, it is inevitable that the adoption of regulatory strategies enhancing intermediary liability in the field of content moderation feeds those concerns and thus increases the constitutional challenges of platform governance. This holds true in all sectors pertaining to the regulation of online freedom of expression but is especially relevant in the context of hate speech moderation, the operationalization of which represents a particularly sensitive activity in the light of the need to consider all relevant contextual aspects and of the concrete risks of discriminatory and biased outcomes driven by the significant implementation of dedicated AI systems.

The following sections will explore how the legal framework on intermediary liability has evolved in Europe since the turn of the millennium and inquire how such developments can impact the governance of hate speech across digital platforms.

3.3. Intermediary liability and hate speech: case law from the ECtHR

3.3.1. *The case of* Delfi AS v Estonia

With respect to intermediary liability in the European context, especially with respect to the liability of ISPs for the failure to remove third-party hate speech, the ECtHR has delivered some significant case law. Namely, in the notorious decision of *Delfi AS v Estonia*,[23] the Grand Chamber of the Strasbourg Court upheld the decision of the Estonian Supreme Court to sentence a news portal to compensate damages for having failed to remove third-party comments that were of a "clearly unlawful nature".[24]

The applicant was an information portal that had published an article concerning a business company, SLK, triggering the audience to publish in the comments section a significant number of anonymous insults and defamatory and offending remarks. After several weeks, SLK requested Delfi AS to remove such comments and claimed contextually compensation for damages. Upon such notice, Delfi had, in fact, immediately removed those comments, but refused to pay compensation. Eventually, the Estonian Supreme Court concluded that, because Delfi usually put in place some forms of moderation

[21] Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press 2018).

[22] On the Facebook Oversight Board see, notably, Kate Klonick, 'The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression' (2020) 129 Yale Law Journal 2418; David Wong and Luciano Floridi, 'Meta's Oversight Board: A Review and Critical Assessment' (2023) 33 Minds and Machines 261. See more *infra*, §5.2.1.2.

[23] *Delfi AS v Estonia* [2015] ECtHR [GC] 64569/09, ECHR 2015.

[24] ibid 140.

practices, it could not be considered as acting as a merely neutral, automatic, and passive actor and, therefore, it should be considered liable for the damages caused by the presence of the defamatory comments. The Supreme Court thus confirmed the County Court's award of 5,000 kroons (approximately 320 euros) in favour of SLK's majority shareholder as compensation for non-pecuniary damages. Delfi, as a result, filed an application to the ECtHR arguing that the award represented an infringement of its right to freedom of expression as enshrined within Article 10 ECHR.

The ECtHR, however, upheld the Estonian Supreme Court's award. On the one hand, it accepted the characterization of Delfi – under Estonian law and consistent case law – as a publisher that offered its media services for economic purposes, rather than as a merely passive hosting provider.[25] On the other hand, the Strasbourg judges concluded that the measure imposed, that is, the sentencing to the payment of non-pecuniary damages for a sum of approximately 320 euros, was in fact proportionate and thus "necessary in a democratic society" as required by Article 10, paragraph 2, ECHR. In order to operate such an assessment the Court stressed, *inter alia*, the role played by the medium used for the establishment of the degree of responsibility of a journalistic actor such as Delfi: citing notably the previous judgment of *Editorial Board of Pravoye Delo and Shtekel v Ukraine*,[26] the judges argued that "the risk of harm posed by content and communications on the Internet to the exercise and enjoyment of human rights and freedoms … is certainly higher than that posed by the press".[27] In the light of such a risk, a business-oriented news portal provider should take extra care to ensure that no such content, including hate speech, is spread across its infrastructures. The Grand Chamber's decision, in practice, recognized for the first time as consistent with the ECHR framework on freedom of expression the possibility for a state to hold the provider of a computer service accountable for the failure to remove immediately third-party comments.

As a matter of fact, the choice of the ECtHR to accept the Estonian Supreme Court's argument that Delfi should be considered as a content provider – rather than as a hosting provider – is itself rather debatable,[28] as the moderation practices of the news portal do not appear to be of such a relevant entity as to warrant the conclusion that it is, in fact, the direct purveyor of the content produced by users. More in general, however, the ultimate outcome of *Delfi*, opening *de facto* the doors to the possibility for governments to punish providers of online services for third-party content (namely, third-party hate

[25] ibid 128–129.

[26] *Editorial Board of Pravoye Delo and Shtekel v Ukraine* [2011] ECtHR 33014/05, ECHR 2011. See *supra*, §2.4.1.

[27] *Delfi AS v Estonia* (n 23) para 133.

[28] With respect to the previous Chamber decision – which was basically confirmed by the Grand Chamber – see among others Dirk Voorhoof, 'Qualification of News Portal as Publisher of Users' Comment May Have Far-Reaching Consequences for Online Freedom of Expression: Delfi AS v. Estonia' (*Strasbourg Observers*, 25 October 2013) <https://strasbourgobservers.com/2013/10/25/qualification-of-news-portal-as-publisher-of-users-comment-may-have-far-reaching-consequences-for-online-freedom-of-expression-delfi-as-v-estonia/> accessed 26 April 2023. However, in this respect, Robert Spano justifies the Court's conclusion by arguing that "the commenting environment was … an integral part of [Delfi's] commercial activity". See Robert Spano, 'Intermediary Liability for Online User Comments under the European Convention on Human Rights' (2017) 17 Human Rights Law Review 665, 676.

speech), raised concerns about the inherent risk it entailed of promoting forms of private and collateral censorship.[29]

In this respect, the decision of the majority was criticized by judges Sajó and Tsotsoria in their joint dissenting opinion, where they argued that the approval of a liability system requiring "constructive knowledge on active Internet intermediaries"[30] may well represent a significant hurdle to the enjoyment of online freedom of expression in Europe, because it may ultimately lead to "deliberate overbreadth; limited procedural protections … and shifting of the burden of error costs", as "the entity in charge of filtering will err on the side of protecting its own liability, rather than protecting freedom of expression".[31]

Additionally, the judgment was also criticized for its apparent failure to develop an ECtHR case law consistent and coherent with the EU framework, in particular with respect to Directive 2000/31/EC, i.e., the "e-Commerce Directive", and related CJEU case law.[32]

3.3.2. *The legacy of* Delfi

Also following the concerns and criticisms raised by the *Delfi* judgment, subsequent ECtHR cases went on to develop a case law which, although maintaining the Grand Camber's decision as a valid and applicable precedent, clarified nonetheless the extent to which providers of online services may in fact be held liable for third-party content, narrowing down sensitively the scope of applicability of that decision.

3.3.2.1. *MTE and Index.hu v Hungary*

In the case of *MTE and Index.hu v Hungary*,[33] the ECtHR had to face a case similar to *Delfi*. The facts concerned the self-regulatory body of Hungarian Internet content providers, MTE, and an Internet news portal, Index.hu, which had published pieces criticizing harshly two real estate management websites, owned by the same company, basically accused of scamming consumers. This led, once again, to triggering readers into publishing anonymous or pseudonymous comments against the company. Eventually, the Hungarian Kúria awarded the company operating the websites compensation for the damages suffered for failure to promptly remove those user comments, even though, in fact, both MTE and Index.hu had immediately taken them down as soon as the lawsuit had been brought against them.

---

[29] See, among others, Dirk Voorhoof, 'Delfi AS v. Estonia: Grand Chamber Confirms Liability of Online News Portal for Offensive Comments Posted by Its Readers' (*Strasbourg Observers*, 18 June 2015) <https://strasbourgobservers.com/2015/06/18/delfi-as-v-estonia-grand-chamber-confirms-liability-of-online-news-portal-for-offensive-comments-posted-by-its-readers/> accessed 26 April 2023; Lisl Brunner, 'The Liability of an Online Intermediary for Third Party Content: The Watchdog Becomes the Monitor: Intermediary Liability after Delfi v Estonia' (2016) 16 Human Rights Law Review 163, 172; Marco Bassini, 'Fundamental Rights and Private Enforcement in the Digital Age' (2019) 25 European Law Journal 182, 192.

[30] *Delfi AS v Estonia* (n 23) joint dissenting opinion of judges Sajó and Tsotsoria para 1.

[31] ibid 2.

[32] See *infra*, §3.4.1.

[33] *Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v Hungary* [2016] ECtHR 22947/13.

In a manner similar to *Delfi*, the ECtHR accepted the national courts' conclusion that the applicants, under the Hungarian Civil Code, could be reasonably treated as content providers (rather than as intermediaries) with respect to third-party anonymous or pseudonymous comments.[34] Additionally, throughout its entire decision, the Strasbourg Court cited *Delfi* rather frequently, thus confirming the validity of the Grand Chamber's decision as a landmark precedent. Most notably, in *MTE*, the Fourth Section stated:

> The Court reiterates in this regard that although not publishers of the comments in the traditional sense, Internet news portals must, in principle, assume duties and responsibilities. Because of the particular nature of the Internet, those duties and responsibilities may differ to some degree from those of a traditional publisher, notably as regards third-party contents.[35]

Nonetheless, with respect to its outcome, *MTE* departed significantly from *Delfi*, as it recognized that the applicants' right to freedom of expression had in fact been breached in violation of Article 10 ECHR.

Indeed, in the case of *MTE*, the Court operated a thorough assessment of all relevant contextual elements, as well as of the content itself of the applicants' publications and of the third-party anonymous contents, and eventually concluded that the imposition of liability upon MTE and Index.hu was not at all proportionate to the purposes sought, so that the measure could not be recognized as "necessary in a democratic society".[36] For instance, the ECtHR stressed that, in the case at hand, at least the first applicant, MTE, was not a business actor, as it was, in fact, a self-regulatory body representing ISPs; whereas the second applicant, Index.hu, should enjoy additional protection as a press outlet since it "provided forum for the exercise of expression rights, enabling the public to impart information and ideas".[37] Moreover, the Court considered that the article published could not "be considered to be devoid of a factual basis or provoking gratuitously offensive comments".[38]

Even more interestingly, the ECtHR held that it should reach a different conclusion from that expressed in *Delfi* because, "although offensive and vulgar, … the incriminated comments did not constitute clearly unlawful speech; and they certainly did not amount to hate speech or incitement to violence".[39] In other words, the ECtHR distinguished the two cases not so much based on an inquiry of the position and role of the intermediary in the dissemination of the content impugned but, rather, based on the type of the illegal content spread. By focusing on this specific aspect, that is, the severity of the comments themselves, the ECtHR was able to uphold the Grand Chamber's previous decision while taking a decision responsive to the many concerns and criticisms that had followed *Delfi*. Such distinguishing between the two cases, nevertheless, appears to be slightly far-fetched and forced, precisely because it shifts the focus of attention from the critical and

[34] ibid 51.
[35] ibid 62.
[36] Convention for the Protection of Human Rights and Fundamental Freedoms 1950 art 10, para 2.
[37] *MTE and Index.hu v Hungary* (n 33) para 61.
[38] ibid 72.
[39] ibid 64.

technical assessment of the degree of liability and accountability of the intermediary involved towards the evaluation of the nature of the third-party content discussed. Besides, in doing so, the Court attaches different liability regimes based on the classification of the content as "clearly unlawful speech" or "hate speech": however, in doing so, it does not clearly define the criteria differentiating such clearly unlawful speech from merely offensive speech.

Overall, *MTE* thus appears to showcase a more careful approach on the part of the Court of Strasbourg, especially if compared to its landmark precedent in *Delfi*. Indeed, the ECtHR, by finding that the Hungarian courts had violated the applicants' right to freedom of expression, implicitly warned that measures entailing an enhanced liability of providers of online services should only be taken in rather serious and extreme situations. Nonetheless, although aimed at narrowing in general terms the acceptability of intermediary liability across the Internet, *MTE* still confirms, with specific respect to the countering of hate speech, the Court's conviction that such content is of such a foul nature as to allow for an increased severity in governmental repressive actions. In other words, though giving impulse to a new strand of case law which, while confirming *Delfi*, tends to be more lenient towards the rights and liberties of Internet actors and more attentive to the risks connected to the imposition of liability for third-party content, *MTE* did not extend such leniency to those cases where the Court believes that forms of hate speech have indeed been uttered. This is clearly confirmed by the Court's conclusions, according to which

> in cases where third-party user comments take the form of hate speech and direct threats to the physical integrity of individuals, the rights and interests of others and of the society as a whole might entitle Contracting States to impose liability on Internet news portals if they failed to take measures to remove clearly unlawful comments without delay, even without notice from the alleged victim or from third parties.[40]

3.3.2.2. Subsequent developments

Subsequent case law from the ECtHR confirmed the conclusion reached in *MTE*, thus upholding the validity of the *Delfi* precedent with respect to hate speech while taking, nonetheless, a rather cautious approach towards the protection of freedom of expression under Article 10 ECHR.[41]

In *Pihl v Sweden*,[42] the ECtHR's Third Section had to deal with a case of defamation concerning the publication of a blogpost – and the consequent uploading of an anonymous comment – upon the website of a small non-profit organization. Considering those

[40] ibid 91. A critical take on such a direction taken in *MTE* is expressed, namely, by Christina Angelopoulos, 'MTE v Hungary: A New ECtHR Judgment on Intermediary Liability and Freedom of Expression' (2016) 11 Journal of Intellectual Property Law & Practice 582.

[41] See, in this respect, Dirk Voorhoof, 'Blog Symposium "Strasbourg Observers Turns Ten" (2): The Court's Subtle Approach of Online Media Platforms' Liability for User-Generated Content since the "Delfi Oracle"' (*Strasbourg Observers*, 10 April 2020) <https://strasbourgobservers.com/2020/04/10/the-courts-subtle-approach-of-online-media-platforms-liability-for-user-generated-content-since-the-delfi-oracle/> accessed 6 May 2023.

[42] *Pihl v Sweden* (dec) [2017] ECtHR 74742/14.

contents to be highly offensive, the applicant had requested the organization to remove them, which was immediately done. The organization also added a post acknowledging the falsity of the information published and apologizing to the applicant. However, Mr. Pihl, having discovered that the defamatory blogpost could still be located by searching his name on the Internet, brought a lawsuit claiming damages. The domestic courts having found the defendant not liable for failing to remove sooner the anonymous comment, Mr. Pihl filed a complaint before the ECtHR, alleging that his rights under Article 8 ECHR had been infringed. The Strasbourg judges, however, underlined that the comment did not amount to hate speech nor to incitement to violence, so that the case required a particularly attentive balancing of the competing interests by domestic courts which, nonetheless, enjoyed in this task a rather wide margin of appreciation.[43] Ultimately, the ECtHR declared Mr. Pihl's application inadmissible, arguing as follows:

> In view of the above, and especially the fact that the comment, although offensive, did not amount to hate speech or incitement to violence and was posted on a small blog run by a non-profit association which took it down the day after the applicant's request and nine days after it had been posted, the Court finds that the domestic courts acted within their margin of appreciation and struck a fair balance between the applicant's rights under Article 8 and the association's opposing right to freedom of expression under Article 10.[44]

Similarly, in *Høiness v Norway*,[45] the applicant argued that her right to respect for private life had been infringed by the failure of the provider of an Internet news portal and forum to remove anonymous comments alleging that she had rather unethically convinced an elderly widow to leave her most of her inheritance in her will. Having failed to obtain compensation before Norwegian courts, Ms. Høiness filed an application before the Strasbourg Court which, however, once again dismissed the complaint of violation of Article 8 by acknowledging that the anonymous comments, while certainly defamatory, did not amount to hate speech.[46] Domestic courts thus "acted within their margin of appreciation when seeking to establish a balance between the applicants' rights under Article 8 and the news portal and host of the debate forums' opposing right to freedom of expression under Article 10".[47]

Both *Pihl* and *Høiness* thus confirmed the strand of case law inaugurated by *Delfi* and perfected by *MTE*, according to which, ultimately, intermediary liability for third-party content should generally be limited to particularly serious cases so as to avoid disproportionate restrictions of those actors' freedom of expression under Article 10,[48] those

[43] ibid 25–26.
[44] ibid 37.
[45] *Høiness v Norway* [2019] ECtHR 43624/14.
[46] ibid 69.
[47] ibid 75.
[48] In this respect, see also the case of *Jezior v Poland*, where the ECtHR, with regard to a local politician upon whose local forum offensive and defamatory comments – but not hate speech – had been published against his competitor, held that the Polish courts' findings against the applicant had represented a disproportionate restriction of his freedom of expression under art 10 ECHR: "*La Cour estime que, à la suite de l'application cumulative des mesures susmentionnées à son encontre, le requérant a subi une sanction susceptible d'avoir un effet inhibiteur sur quelqu'un qui, comme lui-même en l'espèce, administrait à titre*

particularly serious cases being identified in instances of hate speech or incitement to violence.

In the meantime, the 2021 judgment of *Standard Verlagsgesellschaft mbH v Austria (no. 3)*[49] addressed the different, although related, subject of the duty of ISPs – in this case, once again, a news portal allowing readers to post their comments and opinions – to provide information concerning the identity of users having published defamatory content anonymously. While rejecting the interpretation that such comments should be interpreted as journalistic sources, and thus rejecting the direct consequence that the identity of those users should be covered and protected by the guarantees related to journalistic secrecy,[50] the ECtHR concluded nevertheless that ordering the applicant to disclose information about the identity of its recipients would hamper the news portal's freedom of expression under Article 10 ECHR. Indeed, the Court underscored that "an obligation to disclose the data of authors of online comments could deter them from contributing to debate and therefore lead to a chilling effect among users posting in forums in general" and that this would, indirectly, also affect "the applicant company's right as a media company to freedom of the press".[51] Any court, prior to issuing such an order, should thus operate a careful balance between the fundamental rights involved (even though domestic courts may enjoy a significant degree of discretion in this respect): something which Austrian courts had however failed to do.

The *Standard* case, therefore, represents another important tile in the ECtHR case law on intermediary liability, by extending the reach of the value of ISPs' freedom of expression to also include a right to the anonymity of the recipients of their services. The Strasbourg Court, however, in line with *Delfi* and *MTE*, held once again that such a favourable finding may not apply to hate speech, incitement to violence, or other "clearly unlawful content":

---

entièrement gracieux un blog sur Internet sur des sujets importants pour la collectivité. Sur ce point, la Cour rappelle avoir dit que l'imputation d'une responsabilité relativement à des commentaires émanant de tiers peut avoir des conséquences négatives sur l'espace réservé aux commentaires d'un portail Internet et produire un effet dissuasif sur la liberté d'expression sur Internet … En conclusion, la Cour estime que les juridictions nationales ayant statué dans la procédure diligentée à l'encontre du requérant en vertu de la loi sur les élections locales n'ont pas ménagé un juste équilibre entre le droit à la liberté d'expression de l'intéressé et celui, concurrent, de B.K. au respect de sa réputation en tant que candidat aux élections locales. Leurs décisions s'analysant en une ingérence disproportionnée dans le droit à la liberté d'expression du requérant n'étaient donc pas nécessaires dans une société démocratique". *Jezior v Poland* [2020] ECtHR 31955/11 [60–61].

[49] *Standard Verlagsgesellschaft Mbh v Austria (no 3)* [2021] ECtHR 39378/15. With respect to this decision, see among others, Meri Baghdasaryan, 'Standard Verlagsgesellschaft MBH v. Austria (No. 3): Is the ECtHR Standing up for Anonymous Speech Online?' (*Strasbourg Observers*, 25 January 2022) <https://strasbourgobservers.com/2022/01/25/standard-verlagsgesellschaft-mbh-v-austria-no-3-is-the-ecthr-standing-up-for-anonymous-speech-online/> accessed 6 May 2023; Pietro Dunn, 'L'anonimato degli utenti quale forma mediata della libertà di stampa: Il caso *Standard Verlagsgesellschaft mbH c. Austria*' (2022) 1 Rivista di Diritto dei Media 291.

[50] "In the instant case, the Court concludes that the comments posted on the forum by readers of the news portal, while constituting opinions and therefore information in the sense of the Recommendation, were clearly addressed to the public rather than to a journalist". *Standard Verlagsgesellschaft Mbh v Austria (no. 3)* (n 49) para 71.

[51] ibid 74.

> However, even a prima facie examination requires some reasoning and balancing. In the instant case, the lack of any balancing between the opposing interests … overlooks the function of anonymity as a means of avoiding reprisals or unwanted attention and thus the role of anonymity in promoting the free flow of opinions, ideas and information, in particular if political speech is concerned *which is not hate speech or otherwise clearly unlawful*. In view of the fact that no visible weight was given to these aspects, the Court cannot agree with the Government's submission that the Supreme Court struck a fair balance between opposing interests in respect of the question of fundamental rights.[52]

In its 2023 Grand Chamber judgment for the *Sanchez v France*[53] case, the majority also confirmed the legitimacy under Article 10 ECHR of the imposition of a criminal pecuniary penalty upon a local politician for failing to promptly remove third-party hate speech comments that had been posted under a post published on his Facebook "wall": namely, those comments targeted the applicant's political opponent and his partner, as well as the Muslim community as a whole. The applicant had been convicted under French law as a "producer", that is, as the person who had "taken the initiative of creating an electronic communication service for the exchange of opinions and pre-defined topics".[54]

The judgment is especially interesting from at least two points of view. First, it confirmed, under the ECHR framework, the possibility of holding as liable, in a manner similar to a hosting provider, an individual (especially if that individual is a politician in the context of an electoral campaign) who has failed to promptly remove third-party content from their own individual Facebook "wall": a finding which is striking not only because it expands the scope of third-party content liability so as to encompass also the holders of a social networking account, but also because of the high regard traditionally granted by the ECtHR to political freedom of expression.[55] Second, the applicant was held liable even though, in the case at hand, the authors of the impugned comments were not anonymous and had, in fact, also been sentenced to the payment of a fine and to the compensation of damages.

Nonetheless, the ECtHR's Grand Chamber held that France's interference with the applicant's freedom of expression was proportionate and necessary in a democratic society, arguing, *inter alia*, as follows:

---

[52] ibid 95 (emphasis added). Additionally, the ECtHR had previously stressed how "the comments made about the plaintiffs … although offensive and lacking in respect, did not amount to hate speech or incitement to violence … nor were they otherwise clearly unlawful (compare and contrast *Delfi* …)". ibid 89.

[53] *Sanchez v France* [2023] ECtHR [GC] 45581/15, ECHR 2023. For a comment, see Jannika Jahn, 'Strong on Hate Speech, Too Strict on Political Debate: The ECtHR Rules on Politicians' Obligation to Delete Hate Speech on Facebook Page' (*Verfassungsblog*, 25 May 2023) <https://verfassungsblog.de/strong-on-hate-speech-too-strict-on-political-debate/> accessed 1 June 2023; Pietro Dunn, 'Carattere Eccezionale Dell'"Hate Speech" e Nuove Forme Di Responsabilità per Contenuti Di Terzi Nella Giurisprudenza EDU. Nota a C.Edu, Sanchez c. Francia, 15 Maggio 2023' (2023) 6 Osservatorio Costituzionale 238. The decision of the Grand Chamber had already been preceded by a Chamber judgment in *Sanchez v France* [2021] ECtHR 45581/15. On the first decision, see Marina Castellaneta, 'Responsabilità Del Politico per Commenti Altrui Su Facebook: Conforme Alla Convenzione Europea La "Tolleranza Zero" Nei Casi Di Messaggi d'odio' (2021) 3 Rivista di Diritto dei Media 311.

[54] *Sanchez v France* (n 53) para 38.

[55] The judgment itself refers to previous case law addressing the importance and role of political speech in the public debate, while clarifying the duties, obligations, and limits it should comply with: see ibid 146–153.

> The Court would, moreover, reiterate that in cases where third-party user comments take the form of hate speech, the rights and interests of others and of society as a whole may entitle Contracting States to impose liability on the relevant Internet news portals, without contravening Article 10 of the Convention, if they fail to take measures to remove clearly unlawful comments without delay, even without notice from the alleged victim or from third parties (see *Delfi AS* …). Even though the applicant's situation cannot be compared to that of an Internet news portal …, the Court sees no reason to hold otherwise in the present case.[56]

Overall, ECtHR case law concerning the liability and responsibilities of Internet intermediaries for third-party content has undergone important developments after the landmark decision of *Delfi*. Indeed, the Court has showcased a renewed concern for the collateral effects that imposing such forms of liabilities and duties might entail for freedom of expression and has thus become progressively more lenient towards ISPs and attentive to their needs. This approach, besides, is in line with the Council of Europe Committee of Ministers' Recommendation No. R (2018) 2 on the roles and responsibilities of Internet intermediaries, according to which states should ensure "that intermediaries are not held liable for third-party content which they merely give access to or which they transmit or store", although they may hold them "co-responsible … if they do not act expeditiously to restrict access to content or services as soon as they become aware of their illegal nature, including through notice-based procedures".[57]

At the same time, the ECtHR has maintained a rather rigid approach towards the dissemination of hate speech content through the Internet by explicitly recognizing that such a phenomenon may (and should) require the adoption of more stringent measures from states and, consequently, from ISPs themselves. Besides, also in this respect, such a specific consideration towards hate speech is again reflected by the Committee of Ministers, whose Recommendation No. R (2022) 16, recognizing the fundamental role of Internet intermediaries in countering the phenomenon, mentions that states should namely require them "to respect human rights, including the legislation on hate speech, to apply the principles of human rights due diligence throughout their operations, and to take measures in line with existing frameworks and procedures to combat hate speech"[58] and should establish by law that intermediaries "must take effective measures to fulfil duties and responsibilities not to make accessible or disseminate hate speech that is prohibited under criminal, civil or administrative law".[59] Arguably, such a trend is, moreover, consistent with

[56] ibid 140.

[57] Committee of Ministers of the Council of Europe, 'Recommendation No. R (2018) 2 of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries' (Council of Europe 2018) CM/Rec(2018)2 Appendix, para 1.3.7.

[58] Committee of Ministers of the Council of Europe, 'Recommendation No. R (2022) 16 of the Committee of Ministers to Member States on Combating Hate Speech' (Council of Europe 2022) CM/Rec(2022)16 Appendix, para 18.

[59] ibid 22. The Recommendation also adds: "Important elements for the fulfilment of this duty include: rapid processing of reports of such hate speech; removing such hate speech without delay; respecting privacy and data-protection requirements; securing evidence relating to hate speech prohibited under criminal law; reporting cases of such criminal hate speech to the authorities; transmitting to the law-enforcement services, on the basis of an order issued by the competent authority, evidence relating to criminal hate speech; referring unclear and complex cases requiring further assessment to competent self-regulatory or

the developing attitude of the ECtHR with regard to hate speech governance that has been described in Chapter 2: it appears, indeed, that the Court is moving progressively towards an increasingly restrictive perspective on this phenomenon,[60] as highlighted, *inter alia*, in decisions such as *Beizaras and Levickas*,[61] *Association Accept*,[62] and *Valaitis*.[63]

In conclusion, it is possible to identify at least three stages in the evolution and development of the ECtHR case law on intermediary liability and hate speech. The first stage consists of the landmark judgment of *Delfi AS v Estonia*, where the Grand Chamber confirmed that the imposition of liability upon Internet intermediaries for the dissemination of clearly unlawful content is fully consistent with Article 10 ECHR, thus opening the doors for the establishment of intermediary liability for third-party illegal content. The second stage is represented by the decision of *MTE and Index.hu v Hungary*, where the Court clarified that a distinction ought to be made between clearly unlawful content – notably hate speech and incitement to violence – triggering the liability of intermediaries, and other illegal materials, thus making hate speech a rather exceptional case in this respect.

The third stage, finally, is represented by the body of subsequent judgments which, in line with *Delfi* and *MTE*, confirmed such a differentiation in treatment, on the one hand by limiting the scope of intermediary liability to selected and limited cases, where the speech uttered was recognized as being not merely offensive but amounting in fact to hate speech (e.g., *Pihl v Sweden*, *Høiness v Norway*), and, on the other hand, by adopting an increasingly strict and severe approach towards hate speech: namely, it concluded that an order may be issued against an intermediary to disclose information about anonymous users (*Standard Verlagsgesellschaft v Austria*) and that natural persons, notably politicians, may be liable for third-party comments posted on their personal social media "walls" (*Sanchez v France*).

Such developments confirm the deep aversion of the ECHR system from the phenomenon of hate speech. The definition of such a diverse and rather more stringent and severe treatment of hate speech, as opposed to other types of illegal content, is seemingly the expression of a clear orientation and agenda of the ECtHR, rather than of a strictly legal and technical reflection, and has been criticized also in light of its possible inconsistency with other legal frameworks, including that of the EU. Besides, the ECtHR has not so clearly identified the parameters and borders of what is to be considered "clearly unlawful speech" and hate speech, thus giving rise, for the future, to a concrete risk for uncertainty.

co-regulatory institutions or authorities; and foreseeing the possibility of implementing, in unclear and complex cases, provisional measures such as deprioritisation or contextualization".

[60] See *supra*, §2.5.2.2.

[61] *Beizaras and Levickas v Lithuania* [2020] ECtHR 41288/15.

[62] *Association Accept and Others v Romania* [2021] ECtHR 19237/16.

[63] *Valaitis v Lithuania* [2023] ECtHR 39375/19.

3.4. Intermediary liability and hate speech: the framework of the EU

The legal regime concerning intermediary liability for third-party content underwent a parallel and different evolution within the EU framework. The following subsections aim to explore these developments, highlighting in particular the shift from a liberal approach, predominant in the early 2000s, to a progressively more interventionist one, typical of the current historical period, which has led, ultimately, to the adoption in October 2022 of the Digital Services Act.

3.4.1. *Intermediary (non)liability at the turn of the millennium: the e-Commerce Directive*

At the turn of the millennium, inspired by the liberal and techno-optimistic approach of the US, the EU adopted Directive 2000/31/EC, so-called "e-Commerce Directive" (ECD),[64] whose provisions came to represent the normative baseline for the Union's approach towards ISP liability in the twenty years to come. Namely, the ECD, following the model of the notorious Section 230 of the US CDA, introduced a "safe harbour" framework.

Section 230 exempts intermediaries from liability for transmitting or hosting illegal third-party content, even when the latter constitute criminal conducts,[65] while establishing, at the same time, that liability shall not arise even in those cases where providers of computer services engage, actively, in moderation activities aimed at reducing the spread of content they deem illegal, harmful, or, in general, unacceptable (so-called "Good Samaritan clause").[66] Similarly, the ECD offers intermediaries a shield from liability, upon condition that those providers of intermediary services, namely, mere-conduit,[67] caching,[68] and hosting services,[69] comply with certain rules. At the same time, the ECD

[64] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce'), OJ L 178/1. For an overview of the ECD, see namely Lilian Edwards (ed), *The New Legal Framework for E-Commerce in Europe* (Hart 2005); Georgios N Yannopoulos, 'The Immunity of Internet Intermediaries Reconsidered?' in Mariarosaria Taddeo and Luciano Floridi (eds), *The Responsibilities of Online Service Providers* (Springer 2017); Mark D Cole, Christina Etteldorf and Carsten Ullrich, *Cross-Border Dissemination of Online Content* (Nomos 2020) 169–220.

[65] For an overview of Section 230 CDA, see, among others, Eric Goldman, 'An Overview of the United States' Section 230 Internet Immunity' in Giancarlo Frosio (ed), *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020). See *infra*, §4.4.2.

[66] The adoption of the Good Samaritan clause was sparked, namely, by the decision rendered by the New York Supreme Court in *Stratton Oakmont, Inc. v. Prodigy Services Co.*, 23 Media L. Rep. 1794 N.Y. Sup. Ct. 1995).

[67] ECD art 12.

[68] ibid 13. Mere-conduit services consist of the transmission in a communication network of information provided by a recipient of the service, or the provision of access to a communication network.

[69] ibid 14. Caching services consist of the transmission in a communication network of information provided by a recipient of the service where the provider stores that information in an automatic, intermediate and temporary manner for the sole purpose of making more efficient or more secure the information's onward transmission to other recipients upon request.

prohibits EU Member States from imposing upon such providers any duty to conduct general monitoring activities aimed at assessing the presence of such illegal content or activities.[70]

Most notably, whereas in the case of mere-conduit and caching service providers immunity fundamentally depends on the provider not modifying the information transmitted nor interfering with the transmission thereof, providers of hosting services (that is, services consisting "of the storage of information provided by a recipient of the service")[71] are indirectly[72] required to establish notice and take down mechanisms, as the ECD rules that those providers shall maintain the exemption from liability as long as they do not have "actual knowledge of illegal activity or information and, as regards claims for damages, [are] not aware of facts or circumstances from which the illegal activity or information is apparent"[73] and as long as, "upon obtaining such knowledge or awareness", they act "expeditiously to remove or disable access to the information".[74] In other words, a notice to hosting providers concerning the presence of illegal content or the commission of illegal activities through their services would trigger their responsibility to take down those items, or to disable access to them, on penalty of incurring liability for such contents or activities. This strategy mirrors the one adopted by the US Digital Millennium Copyright Act (DMCA)[75] with regard to copyright infringement.[76]

In light of such a provision, indeed rather favourable towards the position of hosting providers, it may be easily understood why *Delfi* has been criticized also because of its apparent lack of coordination with the legal framework of the EU and, thus, for its potential capability of leading to the creation of an ECtHR case law at odds with the ECD. Indeed, if the *Delfi* case had been dealt with by the CJEU, the outcome might have been rather different, notably because, being in fact the applicant a hosting provider – at least with respect to the anonymous third-party comments – liability for such comments under the ECD should only have arisen in case Delfi had failed to respond promptly to SLK's notices. Conversely, the ECtHR considered that Delfi should have actively removed the hate speech content even prior to receiving those complaints.[77] Although it is true that

[70] ibid 15.

[71] ibid 14(1).

[72] Pablo Asbo Baistrocchi, 'Liability of Intermediary Service Providers in the EU Directive on Electronic Commerce' (2002) 19 Santa Clara Computer and High Technology Law Journal 111, 123–124; Aleksandra Kuczerawy, 'From "Notice and Takedown" to "Notice and Stay Down": Risks and Safeguards for Freedom of Expression' in Giancarlo Frosio (ed), *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020).

[73] ECD art 14(1)(a).

[74] ibid 14(1)(b).

[75] Digital Millennium Copyright Act 1998.

[76] Giovanni De Gregorio, *Digital Constitutionalism in Europe: Reframing Rights and Powers in the Algorithmic Society* (Cambridge University Press 2022) 45.

[77] In this respect, see most notably Brunner (n 29) 167–169. See also Oreste Pollicino and Marco Bassini, 'Free Speech, Defamation and the Limits to Freedom of Expression in the EU: A Comparative Analysis' in Andrej Savin and Jan Trzaskowski (eds), *Research Handbook on EU Internet Law* (Edward Elgar Publishing 2014).

"the two Courts work in different jurisdictions and operate with different semantics",[78] many commentators argued that there seemingly was, nevertheless, the concrete risk of creating two parallel and disjointed legal traditions within the same European continent: a risk which, although partly resolved by the leniency showcased by the Strasbourg judges in their subsequent judgments, may still be valid with regard to hate speech content.

Besides, the EU choice to foresee ample liability exemptions favourable to ISPs was in great part motivated, like in the US, by the will not to suffocate the economic and libertarian potential of the Internet, which was, at the time, still in its infancy. It is undeniable, however, that such an approach towards intermediary liability has had in the following years some important political and social consequences. Indeed, the resulting legal regime led most notably to entrusting online platforms with the power to autonomously decide whether to remove or block vast amounts of content: a choice often driven first and foremost by business interests.

The decision to remove illegal or harmful content, including hate speech, was thus mainly left to the discretion of private actors, without any significant safeguards for individual rights and democratic principles, such as the protection of users' right to freedom of expression in conditions of (substantive) equality. Additionally, the identification of what was to be considered as illegal and, therefore, subject to moderation, came to rely in great part upon providers' own – privately enacted – terms of services.[79] Private standards, in other words, progressively came to define the contours of what should and should not be subject to punitive measures. Against this backdrop, values such as the rule of law and the due process of law are clearly at stake.[80]

Moreover, the extraordinary success and spread of digital technologies and of the Internet, and thus of the increased capacities and role of ISPs themselves, have progressively led the scholarly literature, the CJEU and, eventually, the lawmakers of the EU to rethink the strategy to be followed with respect to intermediary liability. The following subsections will investigate precisely these developments in EU digital policies.

3.4.2. *Judicial activism of the Luxembourg Court*

The changing technological and societal landscape first triggered some important judicial reactions from the CJEU, which attempted, namely, to adapt the ECD framework to contemporary needs. With a view to overcoming the inertia of the EU lawmaker, the CJEU took a creative, if not manipulative,[81] approach towards the interpretation of the

[78] Marta Maroni, 'The Liability of Internet Intermediaries and the European Court of Human Rights' in Bilyana Petkova and Tuomas Ojanen (eds), *Fundamental Rights Protection Online: The Future Regulation of Intermediaries* (Edward Elgar Publishing 2020) 268.

[79] Richard Wilson and Molly Land, 'Hate Speech on Social Media: Content Moderation in Context' (2021) 52 Connecticut Law Review 1029. See *infra*, §5.2.

[80] See *infra*, §5.4.3.

[81] Oreste Pollicino, *Judicial Protection of Fundamental Rights on the Internet: A Road Towards Digital Constitutionalism?* (Hart 2021) 13.

Directive's provisions on intermediary liability by clarifying the boundaries of the safe harbour system set therein.

The CJEU's "judicial activism",[82] with regard to intermediary liability, was most notably propelled by a series of cases involving the protection of intellectual property and copyright rights, where the Luxembourg Court came to interpret the provisions of the ECD, including Article 14 on hosting providers, in light of Recital 42:

> The exemptions from liability established in this Directive cover only cases where the activity of the information society service provider is limited to the technical process of operating and giving access to a communication network over which information made available by third parties is transmitted or temporarily stored, for the sole purpose of making the transmission more efficient; *this activity is of a mere technical, automatic and passive nature*, which implies that the information society service provider has neither knowledge of nor control over the information which is transmitted or stored.[83]

Moving from such wording, the CJEU held in the landmark judgment of *Google France*[84] that a necessary precondition for the applicability of the liability exemption under Article 14 is, precisely, that the hosting provider has acted in a merely technical, automatic and passive way, thus precluding the enjoyment of the prerogatives set by the safe harbour system to all those intermediaries that intervened actively in the organisation of the third-party contents: in other words, only "neutral" ISPs could benefit from the ECD's favourable provisions.[85]

With the assessment concerning the neutrality of ISPs being left to the discretion of national courts when applying the Directive, the CJEU tried to clarify what should be the key elements, aspects, and features under consideration when making such an evaluation. Thus, *Google France* excluded, for instance, that simply requiring the payment of a fee for the provision of referencing services should be considered sufficient to prove the non-neutrality of a provider and thus to deprive it of the exemption from liability set within the ECD.[86] However, other elements could contribute to such a conclusion, including the provider's active role in drafting a commercial message associated with the incriminated links and the active establishment and selection of keywords to be associated with such

[82] Giovanni De Gregorio, 'The Rise of Digital Constitutionalism in the European Union' (2021) 19 International Journal of Constitutional Law 41, 49.

[83] ECD rec 42 (emphasis added).

[84] Joined Cases C-236/08, C-237/08 and C-238/08, *Google France SARL and Google Inc v Louis Vuitton Malletier SA, Google France SARL v Viaticum SA and Luteciel SARL* and *Google France SARL v Centre national de recherche en relations humaines (CNRRH) SARL and Others* [2010] ECLI:EU:C:2010:159. In this decision, the CJEU addressed the issue of the liability of the provider of a referencing service with respect to the unlawful exploitation of keywords by third parties infringing trademarks. More specifically, Google had been sued in France by the owners of distinctive signs who complained that, by selecting keywords identical to trademarks, users were seeing advertisements for counterfeit or imitation products alongside original products. See Pollicino, Bassini and De Gregorio (n 4) 76–77.

[85] *Google France* (n 84) 114–125. According to Van Eecke, the CJEU' conclusions in this respect were mistaken, as "the actual content of the recital … clearly points to mere conduit and caching providers, and the discussion about these two services is continued in recital 43 … and recital 44". Patrick Van Eecke, 'Online Service Providers and Liability: A Plea for a Balanced Approach' (2011) 48 Common Market Law Review 1455, 1482.

[86] *Google France* (n 84) 116.

links.[87] Thus, the "non-neutral" character of an intermediary should be assessed by looking at further elements other than the simple request for compensation: rather, the ISP should take an active role in the actual promotion of certain products, services, or contents.

In *L'Oréal*,[88] the Luxembourg judges provided further elements of interpretation. Addressing the lawsuit brought by L'Oréal against eBay for the sale, through the latter's platform, of a number of products in violation of the former's trademark rights, the CJEU confirmed several of the points addressed in *Google France*, adding that

> where … [an] operator has provided assistance which entails, in particular, optimising the presentation of the offers for sale in question or promoting those offers, it must be considered not to have taken a neutral position between the customer-seller concerned and potential buyers but to have played an active role of such a kind as to give it knowledge of, or control over, the data relating to those offers for sale.[89]

As recently as in 2021, *YouTube and Cyando*[90] once again confirmed the principles set in *Google France* and *L'Oréal*, holding that providers of content-sharing platforms, in order to enjoy the safe harbour regime of the ECD, must behave as neutral actors.[91] The decision, however, is particularly interesting as it also addresses the question of whether the resort to AI systems for content moderation and curation should lead to the conclusion of excluding a provider from the enjoyment of the exemption from liability. In this respect, the Court clarified that the implementation of technological measures aimed at detecting illegal content, as well as the provision of automated indexing systems, of a search function, and/or of a recommender system suggesting content based on users' profiles or preferences are "not a sufficient ground for the conclusion that that operator has 'specific' knowledge of illegal activities carried out on that platform or of illegal information stored in it".[92]

With respect to the implementation of technical systems for moderation, the CJEU also rendered two "twin" landmark decisions interpreting the ECD prohibition to impose general monitoring obligations upon providers of intermediary services. Once again

[87] ibid 118.

[88] Case C-324/09, *L'Oréal SA and Others v eBay International AG and Others* [2011] ECLI:EU:C:2011:474.

[89] ibid 116. In the case at hand, specifically, eBay had actively organized the display of products to be sold, thus assisting and fostering transactions between its clients. Moreover, eBay had been notified by L'Oréal of the actual existence of transactions infringing the firm's property rights and had not taken action.

[90] Joined Cases C-682/18 and C-683/18, *Frank Peterson v Google LLC and Others* and *Elsevier Inc v Cyando AG* [2021] ECLI:EU:C:2021:503. The judgment concerned two separate cases involving the liability of providers of content-sharing platforms for copyright infringement: in *YouTube*, the plaintiff had brought action against the famous video-sharing platform after a number of videos reproducing singer Sarah Brightman's performances had been uploaded to the Internet in violation of their proprietary rights. In *Cyando*, Elsevier brought action against the operator of a file-hosting and file-sharing platform where several protected materials had been uploaded and made available for downloading.

[91] ibid 105.

[92] ibid 114.

addressing the field of copyright infringement, the *Scarlet*[93] and *Netlog*[94] judgments held that ordering ISPs to adopt preventive filtering systems to detect content circulated illegally is not a measure consistent with the Directive, as such an order would require "active observation of all communications conducted on the network of the ISP concerned and, consequently, would encompass all information to be transmitted and all customers using that network".[95] In the opinion of the Court, the primary issue, in this respect, concerned the proportionality of such an order. Indeed, an injunction of this type would overall fail to strike an adequate balance between the fundamental rights concerned, namely the protection of intellectual property on the one hand and the freedom to conduct business on the other,[96] as well as disproportionately affect users' rights to privacy and freedom of expression as protected under Articles 8 and 11 CFREU.[97]

Nonetheless, as also recognized by the CJEU,[98] the prohibition of general obligations to monitor does not prevent Member States from requiring from ISPs "the termination or prevention of any infringement, including the removal of illegal information or the disabling of access to it"[99], as Member States are only prevented from "imposing a monitoring obligation on service providers only with respect to obligations of a general nature".[100] Therefore, as long as an order is sufficiently substantiated and limited as to its scope, that is, as to the specific content which should be acted upon, that order shall be in compliance with EU law and the ECD. With respect to this point, the Luxembourg Court, moving this time from a case of defamation, offered some significant insights into the width of such a power when it rendered in 2019 the landmark judgment of *Glawischnig-Piesczek v Facebook.*[101]

The case concerned the publication by a Facebook user of a post where a thumbnail image portraying Eva Glawischnig-Piesczek, a representative of the Austrian Green Party, was associated with highly derogatory and insulting terms – "lousy traitor", "corrupt oaf", a member of a "fascist party". The Austrian Supreme Court referred some questions to the CJEU regarding, notably, the consistency with EU law of an order requiring a hosting provider such as Facebook to remove content declared to be illegal, as well as the territorial and material scope that such an injunction might have. With respect to the

[93] Case C-70/10, *Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM)* [2011] ECLI:EU:C:2011:771.

[94] Case C-360/10, *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV* [2012] ECLI:EU:C:2012:85.

[95] *Scarlet* (n 93) para 39. Similarly, *Netlog* (n 94) para 38.

[96] *Scarlet* (n 93) paras 49, 53; *Netlog* (n 94) paras 44–47.

[97] *Scarlet* (n 93) paras 50–53; *Netlog* (n 94) paras 48–51.

[98] *Scarlet* (n 93) paras 30–31; *Netlog* (n 94) paras 28–29.

[99] ECD rec 45.

[100] ibid rec 47.

[101] Case C-18/18, *Eva Glawischnig-Piesczek v Facebook Ireland Limited* [2019] ECLI:EU:C:2019:821. With respect to this decision, see, among others, Aleksandra Kuczerawy, 'General Monitoring Obligations: A New Cornerstone of Internet Regulation in the EU?' in Centre for IT & IP Law (ed), *Rethinking IT and IP law: Celebrating 30 years CiTiP* (Intersentia 2020); Daphne Keller, 'Facebook Filters, Fundamental Rights, and the CJEU's *Glawischnig-Piesczek* Ruling' (2020) 69 GRUR International 616; Giovanni De Gregorio, '*Google v. CNIL* and *Glawischnig-Piesczek v. Facebook*: content and data in the algorithmic society' (2020) 1 Rivista di Diritto dei Media 249.

territorial scope, the CJEU held that the ECD's prohibition of general monitoring obligations did not preclude domestic courts from ordering the removal of that illegal content on a global scale.[102] As regards the material scope, the Luxembourg Court concluded that a removal order may encompass not only the content that has been found to be illegal but also any content that is "identical" or "equivalent" to it, provided, in the last case, that

> the monitoring of and search for the information concerned by such an injunction are limited to information conveying a message the content of which remains essentially unchanged compared with the content which gave rise to the finding of illegality and containing the elements specified in the injunction, and provided that the differences in the wording of that equivalent content, compared with the wording characterising the information which was previously declared to be illegal, are not such as to require the host provider to carry out an independent assessment of that content.[103]

However, according to Keller, *Glawischnig-Piesczek v Facebook* fails to address in a satisfying manner the matters related to the implications that its findings may have upon the fundamental rights not only of hosting providers but, even more, of the users of the Internet themselves: namely, their rights to privacy and data protection; to freedom of

[102] "In order to answer that question, it must be observed that, as is apparent, notably from Article 18(1), Directive 2000/31 does not make provision in that regard for any limitation, including a territorial limitation, on the scope of the measures which Member States are entitled to adopt in accordance with that directive. Consequently, and also with reference to paragraphs 29 and 30 above, Directive 2000/31 does not preclude those injunction measures from producing effects worldwide. However, it is apparent from recitals 58 and 60 of that directive that, in view of the global dimension of electronic commerce, the EU legislature considered it necessary to ensure that EU rules in that area are consistent with the rules applicable at international level. It is up to Member States to ensure that the measures which they adopt and which produce effects worldwide take due account of those rules". *Glawischnig-Piesczek* (n 101) paras 49–52. Similarly, with respect to the so-called "right to be forgotten", see Case C-507/17, *Google LLC, successor in law to Google Inc, v Commission nationale de l'informatique et des libertés (CNIL)* [2019] ECLI:EU:C:2019:772 [72], where the CJEU concluded that, while such a right, under EU law, does not entail the obligation for a search engine to carry out the de-referencing of the personal data requested on all its versions, the authorities of the Member States concerned are not prevented from issuing de-referencing orders applicable also outside the Union. In fact, the CJEU's main focus, in *Google v CNIL*, upon the impossibility of recognizing an extraterritorial scope of action under EU law had led many commentators to argue that the two decisions, though close in time, were incoherent with each other. However, as highlighted, among others, by De Gregorio, both *Glawischnig-Piesczek* and *Google v CNIL* "lead to the same result, namely that EU law does not either impose or preclude national measures whose scope extends worldwide". De Gregorio, 'Google v. CNIL and Glawischnig-Piesczek v. Facebook' (n 101) 259. See also, on the links and connections between the two decisions, Oreste Pollicino, 'L'"Autunno Caldo" Della Corte Di Giustizia in Tema Di Tutela Dei Diritti Fondamentali in Rete e Le Sfide Del Costituzionalismo Alle Prese Con i Nuovi Poteri Privati in Ambito Digitale' (2019) 19 Federalismi.it 1.

[103] *Glawischnig-Piesczek* (n 101) para 55. The CJEU, in this respect, clarifies that "it is important that the equivalent information … contains specific elements which are properly identified in the injunction, such as the name of the person concerned by the infringement determined previously, the circumstances in which that infringement was determined and equivalent content to that which was declared to be illegal. Differences in the wording of that equivalent content, compared with the content which was declared to be illegal, must not, in any event, be such as to require the host provider concerned to carry out an independent assessment of that content. In those circumstances, an obligation such as the one described … on the one hand – in so far as it also extends to information with equivalent content – appears to be sufficiently effective for ensuring that the person targeted by the defamatory statements is protected. On the other hand, that protection is not provided by means of an excessive obligation being imposed on the host provider, in so far as the monitoring of and search for information which it requires are limited to information containing the elements specified in the injunction, and its defamatory content of an equivalent nature does not require the host provider to carry out an independent assessment, since the latter has recourse to automated search tools and technologies". ibid 45–46.

expression and information; to a fair trial and effective remedy; and, finally, to equality and non-discrimination, which may be affected by the use of biased and under-representative automated content filters.[104]

The CJEU case law referred to above showcases the definition of a roadmap promoted by the Court with respect to intermediary liability for third-party content, a roadmap indicating the Luxembourg judges' will to set aside the inherently liberal approach of the first years of the twenty-first century. Although such case law did not address, specifically, the subject of hate speech governance, the push for an update of the previous framework clearly has – and will likely have even more in the future – an impact also upon that area, by encouraging the EU lawmaker to draft new legislation holding ISPs accountable for the spread of illegal content: most notably, it set a guideline for the adoption of the Digital Services Act.[105]

Particularly relevant for the purposes of hate speech governance is, seemingly, the CJEU's judgment for *Glawischnig-Piesczek v Facebook*, as it opens up to the opportunity, in cases where a content is found to be unlawful hate speech, to issue an order requiring a hosting provider to remove all "equivalent" content. Such a conclusion may have both positive and negative consequences. Indeed, while it may well represent an additional instrument in the hands of domestic courts to counter the dissemination of hate speech across the Internet, this kind of injunction would likely cause ISPs to implement rather restrictive content moderation filters, with little regard, as noted by Keller, to the provision of guarantees ensuring the fundamental rights of the recipients of the service. The CJEU's holding may thus contribute, in the future, to a higher removal rate, in absolute terms, of hate speech content across the Internet: nonetheless, it may hinder a substantive equality-oriented strategy against hate speech, such as that suggested in Chapter 2, notably by risking impacting disproportionately upon minority and discriminated groups' participation in the online digital environment.

3.4.3. *A new phase for the EU*

3.4.3.1. The "new season" of content moderation regulation

Against this backdrop, the second half of the 2010s saw the beginning of a new season for content moderation regulation within the EU, with the adoption of a rather wide array of new pieces of legislation requiring intermediaries to comply with duties and obligations to moderate online content to prevent the spread of illegal and/or harmful material through the Internet.[106]

[104] Keller (n 101) 2.

[105] See *infra*, §5.

[106] Claudia E Haupt, 'Regulating Speech Online: Free Speech Values in Constitutional Frames' (2021) 99 Washington University Law Review 751, 760; De Gregorio, 'The Rise of Digital Constitutionalism in the European Union' (n 82).

Coherently with the European Digital Single Market Strategy (EDSM),[107] the EU Commission promoted the rise of an innovative framework pushing for a more active involvement of intermediaries in the promotion of a safe and trustworthy digital landscape, while fostering the adoption of procedural safeguards to increase the degree of transparency and accountability of content moderation practices.[108] According to the Commission, indeed, online platforms have taken the centre stage in the provision of access to information and content and, as a result, must take on "wider responsibility" with respect to ensuring a level playing field for comparable digital services, behaving responsibly to protect European core values, promoting transparency and fairness for maintaining user trust and safeguarding innovation, and fostering open and non-discriminatory markets in a data-driven economy.[109]

This political approach resulted in a new wave of self-regulatory, co-regulatory, and regulatory[110] tools whose objective is, namely, to regulate platforms' moderation practices. It has been argued that the increased popularity of such strategies marks a shift from a merely negligence-based "liability" scheme, such as that set by the ECD, towards an approach more focused on the "responsibility" to actively implement and pursue publicly relevant policies, thus making platforms, in practice, the watchdogs of the Internet.[111] Such a choice clearly entails both upsides and downsides, the former consisting of the possibility of exploiting the economic and computational power held by platforms themselves, the latter being represented by the inherent risks for the rule of law and due process values and principles that an increase in the private power over speech governance necessarily entails. Indeed,

> this development poses plenty of challenges. First, enforcement through private ordering and voluntary measures moves the adjudication of lawful and unlawful content out of public oversight. In addition, private ordering … does push an amorphous notion of responsibility that incentivizes intermediaries' self-intervention to police allegedly infringing activities in the Internet. Further, enforcement would be looking once again for an 'answer to the machine in the machine'. By enlisting online intermediaries as watchdogs, governments would de facto delegate online enforcement to algorithmic tools – with limited or no accountability. Finally, tightly connected to the points above, transferring regulation and adjudication of Internet rights to private actors highlights unescapable tensions with fundamental rights – such as freedom of information, freedom of expression,

[107] European Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. A Digital Single Market Strategy for Europe' COM(2015) 192 final.

[108] Giovanni De Gregorio, 'Expressions on Platforms: Freedom of Expression and ISP Liability in the European Digital Single Market' (2018) 2 European Competition and Regulatory Law Review 203.

[109] European Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Online Platforms and the Digital Single Market Opportunities and Challenges for Europe' COM(2016) 288 final.

[110] On the notions of self-regulation, co-regulation, and regulation see, notably, Ugo Pagallo, Pompeu Casanovas and Robert Madelin, 'The Middle-out Approach: Assessing Models of Legal Governance in Data Protection, Artificial Intelligence, and the Web of Data' (2019) 7 The Theory and Practice of Legislation 1.

[111] Giancarlo Frosio, 'Why Keep a Dog and Bark Yourself? From Intermediary Liability to Responsibility' (2018) 26 International Journal of Law and Information Technology 1.

> freedom of business or a fundamental right to Internet access – by limiting access to information, causing chilling effects, or curbing due process.[112]

With respect to the EU's renovated approach towards ISPs, at least two important phases can be identified. Indeed, while, at first, the EU put in place sectoral reforms concerning specific fields of content moderation governance,[113] the eventual adoption of the already mentioned Digital Services Act marked the shift towards a horizontal, all-encompassing, approach.[114]

A characterizing aspect of the EU's new regulatory season is its frequent resort to a "risk-based approach"[115] towards content moderation. Indeed, with a view to limiting the collateral effects that an increased intermediary liability for third-party content necessarily entails, both with regard to the market and to ISPs' freedom to conduct business, on the one hand, and with regard to the protection of users' fundamental rights, on the other hand, the EU has taken the path of such an approach with the goal of calibrating the new duties based on the actual risks entailed by the provision of a service. In other words, through the adoption of a risk-based approach, whereby the concept of risk is used as a proxy to scale the obligations imposed on platforms and thus avoid the imposition of unnecessary and disproportionate burdens, the strategy implemented by the EU seeks to establish an accountability regime which is more or less strict depending on the assessed risk of harms.

This way, smaller providers and providers that do not offer "dangerous" services should not be constrained by the same strict rules as bigger IT companies and providers of riskier services, with a beneficial effect both on the market actors and on the individuals potentially subjected to forms of collateral censorship.[116]

---

[112] Giancarlo Frosio and Martin Husovec, 'Accountability and Responsibility of Online Intermediaries' in Giancarlo Frosio (ed), *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020) 630.

[113] See *infra*, §3.4.3.2.

[114] Oreste Pollicino, 'Potere Digitale' in Marta Cartabia and Marco Ruotolo (eds), *Enciclopedia del Diritto*, vol. *Potere e Costituzione* (Giuffrè 2023) 410 439.

[115] On the concept of the risk-based approach, with a specific eye on privacy and data protection law, see, among others, Raphaël Gellert, 'Understanding the Notion of Risk in the General Data Protection Regulation' (2018) 34 Computer Law & Security Review 279; Raphaël Gellert, *The Risk-Based Approach to Data Protection* (Oxford University Press 2020); Milda Macenaite, 'The "Riskification" of European Data Protection Law through a Two-Fold Shift' (2017) 8 European Journal of Risk Regulation 506; Claudia Quelle, 'Enhancing Compliance under the General Data Protection Regulation: The Risky Upshot of the Accountability- and Risk-Based Approach' (2018) 9 European Journal of Risk Regulation 502; Maria Eduarda Gonçalves, 'The Risk-Based Approach under the New EU Data Protection Regulation: A Critical Perspective' (2020) 23 Journal of Risk Research 139; Jeroen van der Heijden, 'Risk as an Approach to Regulatory Governance: An Evidence Synthesis and Research Agenda' (2021) 11 SAGE Open <https://doi.org/10.1177/21582440211032202> accessed 11 April 2022; Bridget M Hutter, 'Risk, Regulation, and Management' in Peter Taylor-Gooby and Jens O Zinn (eds), *Risk in Social Science* (Oxford University Press 2006); Adrian Vermeule, *The Constitution of Risk* (Cambridge University Press 2013).

[116] See, on the characters of the risk-based approach in EU digital policies, also with respect to content moderation regulation, Giovanni De Gregorio and Pietro Dunn, 'The European Risk-Based Approaches: Connecting Constitutional Dots in the Digital Age' (2022) 59 Common Market Law Review 473, 483–488.

3.4.3.2. The new sectoral framework on illegal content

A first notable example of the new wave of legislation against the spread of illegal content is represented by the amendments made in 2018 to Directive 2010/13/EU, i.e., the Audiovisual Media Services Directive.[117] Indeed, as mentioned in Chapter 2,[118] Directive 2018/1808, so-called AVMSD Refit Directive,[119] introduced within the scope of the EU framework on the audiovisual market a new legal regime for providers of video-sharing platforms services (VSPs), that is, of services whose principal purpose or whose essential functionality is the provision to the general public of programmes and/or user-generated videos that are outside the platform's editorial responsibility but the organization of which is determined by the provider "including by automatic means or algorithms in particular by displaying, tagging and sequencing".[120]

Following the Refit, VSPs must, namely, take "appropriate measures"[121] to protect the public from material that is harmful to the full development of minors; that constitutes incitement to violence or hatred based on any of the grounds protected by Article 21 CFREU; or that amounts to serious criminal offences such as provocation to commit a terrorist offence, child pornography, and offences related to racism and xenophobia pursuant to Framework Decision 2008/913/JHA.[122] Under the new framework, therefore, the concerned ISPs must put in place, on penalty of a fine, strategies aimed at countering the spread not only of hate speech (both illegal and "simply" harmful under EU law) but also at countering a rather significant set of unwarranted content.[123] The reference to "appropriate measures", nonetheless, reflects precisely the choice of implementing a risk-based approach such as that mentioned in the previous subsection: indeed, it implies that VSPs are vested with the duty (and power) to assess the risks of harmful or illegal content being spread across their infrastructures and, subsequently, identify and implement a mitigating strategy that is effective and proportionate to such a risk. In this sense, the Directive clarifies that

> the appropriate measures shall be determined in light of the content in question, the harm it may cause, the characteristics of the category of persons to be protected as well as the rights and legitimate interests at stake, including those of the video-sharing platform

[117] 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), OJ L 95/1.

[118] See *supra*, §2.2.3.2.

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

[120] AVMSD art 1, para 1, lett (aa).

[121] ibid 28b, para 1.

[122] 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 2008 (OJ L 328/5). See *supra*, §2.2.3.2.

[123] On the legal regime concerning VSPs, following the AVMSD Refit Directive, see Ľuboš Kukliš, 'Video-Sharing Platforms in AVMSD: A New Kind of Content Regulation' in Pier Luigi Parcu and Elda Brogi (eds), *Research Handbook on EU Media Law and Policy* (Edward Elgar Publishing 2021).

> providers and the users having created or uploaded the content as well as the general public interest.[124]

A similar approach can be found in the field of copyright law, where Directive (EU) 2019/790 on copyright in the Digital Single Market (DSM Copyright Directive)[125] introduced a new provision concerning the liability for copyright infringement of providers of content-sharing platforms. Indeed, Article 17 of the Directive, commonly known as the "value-gap" provision, requires that such providers make best efforts to obtain an authorization from rightholders to communicate to the public or make available to the public works or other subject matter.[126] In case they fail to obtain such authorization, however, they must make, "in accordance with high industry standards of professional diligence, best efforts to ensure the unavailability of specific works and other subject matter for which the rightholders have provided the service providers with the relevant and necessary information",[127] as well as act expeditiously to disable access to, or to remove, content infringing copyright as soon as they have been notified and make "best efforts to prevent their future uploads".[128] Failure to do so will lead them to be directly liable for the copyright infringement, as the Directive clearly states that "an online content-sharing service provider performs an act of communication to the public or an act of making available to the public … when it gives the public access to copyright-protected works or other protected subject matter uploaded by its users".[129]

The reference to "best efforts" – and the understanding of what such a notion should entail – raised many concerns in the aftermath of the adoption of the Directive and in the context of its domestic implementation across Member States: namely, the question arose with respect to whether "best efforts" should be interpreted in absolute terms, that is, as requiring all providers of such services to implement the maximum efforts possible at the state of the art, or whether it should be interpreted under the lens of the principle of proportionality, thus also taking into account all specific contextual aspects – including, for instance, the dimensions and finances of the provider.[130] Many commentators, indeed, argued that such a rule might imply a preventive obligation to implement "upload filters",

[124] AVMSD 28b, para 3. The paragraph also features a list of suggested measures that VSPs may choose to adopt.

[125] 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 (DSM Copyright Directive), OJ L 130/92.

[126] ibid 17, paras 1, 4, lett (a).

[127] ibid 17, para 4, lett (b).

[128] ibid 17, para 4, lett (c).

[129] ibid 17, para 1. This represents an important novelty introduced by the DSM Copyright Directive, as the mere presence of content upon a content-sharing platforms did not constitute previously, *per se*, an act of communication to the public, unless the provider "contribute[d], beyond merely making that platform available, to giving access to such content to the public in breach of copyright", as clearly stated in *YouTube and Cyando* (n 90) para 102.

[130] Marco Bassini and Giovanni De Gregorio, 'The Implementation of the Copyright Directive in Italy and the Proper Understanding of the "best Efforts" Clause' (MediaLaws) <https://www.medialaws.eu/wp-content/uploads/2021/04/Policy-paper-ML-Article-17-and-best-efforts-5.pdf> accessed 18 May 2023.

potentially harmful to individual users' rights and liberties, and introduce a general monitoring duty incoherent with the overall liability regime as established within the ECD.[131]

It is worth mentioning, nonetheless, that paragraph 5 of Article 17 seemingly mitigates itself the risks of such collateral effects, by clarifying that specific consideration should be given to "the type, the audience and the size of the service and the type of works or other subject matter uploaded by the users of the service" as well as to "the availability of suitable and effective means and their cost for service providers", while mentioning explicitly the need to interpret the provision in light of the principle of proportionality. Thus, in *Poland v Parliament and Council*,[132] the CJEU rejected the action of annulment promoted by the Polish government against Article 17, holding that the value-gap provision did not entail a violation of the right to freedom of expression, as protected by Article 11 CFREU, precisely because the "best efforts" clause must be interpreted, coherently with the Charter, as allowing their adaptability to the particular circumstances of the various online content-sharing service providers and also to the development of industry practices and of available technologies.[133] Overall, the CJEU held that the legal regime envisaged by the impugned provision was coherent with the principle of proportionality, while stressing that, when transposing Article 17, "Members States must … take care to act on the basis of an interpretation of that provision which allows a fair balance to be struck between the various fundamental rights protected by the Charter" and that also domestic authorities and courts must "make sure that they do not act on the basis of an interpretation of the provision which would be in conflict with those fundamental rights or with the other general principles of EU law, such as the principle of proportionality".[134]

A third example of a sectoral EU intervention in the field of content moderation regulation is represented by Regulation (EU) 2021/784 on terrorist content online (TERREG)[135] which, on the one hand, recognizes the power of Member States to issue orders to remove or disable access to terrorist content to providers of hosting services[136] and, on

[131] Christina Angelopoulos and João Pedro Quintais, 'Fixing Copyright Reform' (2019) 10 Journal of Intellectual Property, information Technology and Electronic Commerce Law 147; Martin Senftleben, 'Bermuda Triangle: Licensing, Filtering and Privileging User-Generated Content Under the Directive on Copyright in the Digital Single Market' (2019) 41 European Intellectual Property Review 480; João Pedro Quintais, 'The New Copyright in the Digital Single Market Directive: A Critical Look' (2020) 42 European Intellectual Property Review 28.

[132] Case C-401/19, *Republic of Poland v European Parliament, Council of the European Union* [2022] ECLI:EU:C:2022:297.

[133] ibid 73. Additionally, the CJEU underscored that the first subparagraph of art 17, para 7, expressly states that the "cooperation between online content-sharing service providers and rightholders shall not result in the prevention of the availability of works or other subject matter uploaded by users, which do not infringe copyright and related rights, including where such works or other subject matter are covered by an exception or limitation" of those rights. According to the decision, the unambiguous wording of this sentence "is not limited to requiring online content-sharing service providers to make their 'best efforts' to that end, but prescribes a specific result to be achieved". ibid 78.

[134] *Poland v Parliament and Council* (n 132) para 99.

[135] Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online, OJ L 172/79. With respect to the topic of the regulation of terrorist content online, under a comparative perspective, see among others Eliza Bechtold, 'Terrorism, the Internet, and the Threat to Freedom of Expression: The Regulation of Digital Intermediaries in Europe and the United States' (2020) 12 Journal of Media Law 13.

[136] TERREG art 3.

the other hand, establishes a specific regime for those providers that national authorities, based on objective factors, designate as being "exposed to terrorist content".[137] The Regulation thus establishes a distinction between at least two tiers of risk, based on which only those providers that entail higher levels of danger are required to put in place costlier and more challenging mitigation systems. According to the lawmaker, additionally, the "specific measures" required shall be "targeted and proportionate" to the seriousness of the level of exposure, as well as to the "operational capabilities, financial strength, the number of users of the services … and the amount of content they provide".[138]

A fourth example is represented, additionally, by the proposal put forward by the Commission in May 2022 concerning the adoption of a Regulation on Child Sexual Abuse Material (CSAM).[139] This proposal is different from the pieces of legislation mentioned before as it is the first one addressing the field of content moderation to be published after the adoption of the Digital Services Act (DSA).[140] Rhe CSAM Regulation proposal is, in fact, self-declaredly intended precisely as a *lex specialis* to the DSA.[141] The proposal aims at harmonizing the EU legal framework on the prevention and fight against online child sexual abuse materials, as well as solicitation of children (grooming),[142] while balancing the pursuit of such a goal with the need to guarantee the full respect of all rights and freedoms under the CFREU that may be affected by the implementation of dedicated moderation systems and practices.[143]

The CSAM Regulation proposal, most notably, foresees two main sets of obligations for providers of hosting and interpersonal communication services. First, they must periodically assess their exposure to the danger of being misused and implement "reasonable mitigation measures, tailored to the risk identified … to minimise that risk".[144] These measures shall be proportionate and applied "in a diligent and non-discriminatory manner, having due regard, in all circumstances, to the potential consequences of the mitigation measures for the exercise of fundamental rights of all parties affected".[145] Second, those providers must also comply with the detection orders that national judicial or independent administrative authorities, on request of the local Coordinating Authority,[146] may

[137] ibid 5.

[138] ibid 5, para 3, lett (b).

[139] European Commission, 'Communication of 11 May 2022, Proposal for a Regulation of the European Parliament and of the Council Laying down Rules to Prevent and Combat Child Sexual Abuse' COM(2022) 209 final.

[140] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act), OJ L 277/1. See *infra*, §3.5.

[141] European Commission, 'CSAM Regulation Proposal' (n 139) rec 8.

[142] For this reason, the proposal introduces a range of duties and obligations affecting the providers of those online services that are considered to be the most vulnerable to being misused for those purposes: namely, hosting services, interpersonal communications services, software application stores, and Internet access services. The providers of the last two types of services, nonetheless, are required to comply with fewer obligations than the first two.

[143] European Commission, 'CSAM Regulation Proposal' (n 139) art 2, lett (f).

[144] ibid 4, para 1.

[145] ibid 4, para 2, lett (c).

[146] ibid 25.

decide to issue: when reached by such orders, providers of hosting and interpersonal communication services must put in place mechanisms to identify the dissemination to the public of known or new CSAM or the carrying out of activities constituting solicitation of children, namely through the installation and operation of dedicated technologies.[147] Additionally, they are also required to report any information they may have become aware of indicating the potential carrying out of online child sexual abuse through their services[148] and they may be reached by removal orders issued by national judicial or independent administrative authorities.[149]

As anticipated above, the legal strategies followed by these sectoral pieces of legislation all feature the resort to risk as a fundamental proxy to calibrate the measures to be actively adopted to fight illegal content. In so doing, the law tends to delegate the tasks of detecting and sanctioning the upload of illegal content directly to the affected providers of intermediary services. Indeed, while they often include suggestions as to the possible tools they might adopt – as well as with respect to the elements to be taken into account when assessing the concrete risk level connected to a certain service –, the AVMSD Refit Directive, DSM Copyright Directive, TERREG, and CSAM Regulation proposal, ultimately, all leave up to ISPs the choice as to what moderation strategies to implement. Thus, providers are made responsible and accountable for the choices made and may be held liable if the measures put in place prove to be ineffective.

Furthermore, with a view to mitigating the impact that such a delegation of power to private entities might have, indirectly, on the fundamental rights and freedoms of the recipients of intermediary services, including freedom of expression and the right to non-discrimination,[150] the EU lawmaker has, on the one hand, introduced a range of procedural countermeasures, allowing users to complain against unjust moderation decisions. These often include not only redress mechanisms internal to the intermediary itself, but also out-of-court solutions as well as the possibility of claiming action before the Member States' courts.[151] The CSAM Regulation proposal, in specifying the right of recipients to an effective judicial redress system,[152] also requires providers of intermediary services to clearly inform users of such a possibility.[153]

---

[147] ibid 7–11. In this respect, the new obligation of complying with detection orders appears to be particularly challenging, especially in the light of the different types of content that may constitute the material scope of the orders themselves. Indeed, known CSAM, new CSAM, and "grooming" all entail significantly different challenges for the purposes of detection. Whereas known CSAM may quite easily be detected and recognized through the use of relatively simple AI systems for content moderation, the issue of collateral censorship may be higher in the case of new CSAM. As for grooming, it is the proposal's Explanatory Memorandum itself that mentions the inherent challenges faced by its detection. These challenges do not only attain to the technical difficulties faced by AI in understanding, semantically, when an adult is actually engaging in acts of grooming, but also to the significant impact that its detection might entail on users' freedom to privacy and on their right to secrecy of communications.

[148] ibid 12–13.

[149] ibid 14–15.

[150] See *supra*, §2.5.3.

[151] DSM Copyright Directive art 17, para 9; AVMSD arts 28b, paras 7-8; TERREG art 10.

[152] European Commission, 'CSAM Regulation Proposal' (n 139) art 9, para 1.

[153] See, e.g., ibid 10, para 5, lett (c).

On the other hand, from a more substantial perspective, the law, as shown above, explicitly requires that the measures adopted by online intermediaries are calibrated not only based on the risk of illegal activity, but also on the specular evaluation of the risk of violating individuals' fundamental rights as a result of the moderation strategies deployed. Proportionality, in other words, is considered not only as a guarantee for platforms and providers against the obligation of implementing disproportionately costly measures, but also as a guarantee for users themselves. Such a perspective, seemingly, has gathered an increasing momentum in recent years. For instance, pursuant to the TERREG, the "specific measures" implemented by providers of hosting services exposed to terrorist content must be "applied in a manner that takes full account of the rights and legitimate interests of the users, in particular users' fundamental rights concerning freedom of expression and information, respect for private life and protection of personal data"[154] as well as "applied in a diligent and non-discriminatory manner".[155] Specific consideration is given to the impact that the use of automated systems could have on fundamental rights.[156]

Thus, while the TERREG, once again, states that the resort to technical measures calls for "appropriate and effective safeguards, in particular through human oversight and verification", the CSAM Regulation proposal envisages a range of guarantees tailored to each of the obligations it sets: when technologies are used by providers of hosting services and of interpersonal communication services to comply with detection orders, for example, the proposal requires that such technologies are "the least intrusive in terms of the impact on the users' right to private and family life, including the confidentiality of communication, and to the protection of personal data"[157] and "sufficiently reliable in that they limit to the maximum extent possible the rate of errors regarding the detection",[158] while mitigation measures adopted in the light of the results of their risk assessments should be "applied in a diligent and non-discriminatory manner, having due regard, in all circumstances, to the potential consequences of the mitigation measures for the exercise of fundamental rights of all parties affected".[159]

3.4.3.3. The Code of Conduct on Illegal Hate Speech

With specific respect to the countering of illegal hate speech online, there are currently no hard law acts such as the ones examined within the previous subsection. Conversely, in 2016, the European Commission, together with the representatives of a number of IT companies[160] and of civil society organizations, agreed upon a Code of Conduct (CoC)

[154] TERREG art 5, para 3, lett (c).
[155] ibid 5, para 3, lett (d).
[156] See *infra*, §5.
[157] European Commission, 'CSAM Regulation Proposal' (n 139) art 10, para 3, lett (c).
[158] ibid 10, para 3, lett (d).
[159] ibid 4, para 2, lett (c).
[160] Originally, the CoC was signed by Facebook, Microsoft, Twitter (today X) and YouTube. The agreement was joined by Instagram, Snapchat and Dailymotion in 2018; by Jeuxvideo.com in 2019; by TikTok in 2020; by LinkedIn in 2021; and, finally, by Rakuten, Viber and Twitch in 2022. See European

on Countering Illegal Hate Speech Online.[161] The CoC is a form of self-regulatory[162] instrument that IT companies may decide to adhere to on a voluntary basis: as such, it may be perceived as a "less intrusive"[163] form of intervention if compared to other top-down regulatory strategies.

As a matter of fact, since the early 2000s and increasingly throughout the following two decades, EU institutions have recognized the role of self-regulatory and co-regulatory tools of governance.[164] Thus, for instance, the Commission's Communication on a renewed EU strategy 2011-2014 for Corporate Social Responsibility had already cited self- and co-regulation schemes as "important means by which enterprises seek to meet their social responsibility".[165] More recently, the 2016 Interinstitutional Agreement on Better Law-Making underscored the need to avoid as much as possible overregulation and excessive administrative burdens, while promoting the participation of relevant stakeholder parties in policymaking and lawmaking.[166] Self- and co-regulation, in other words, have long been perceived as governance strategies capable of reducing the collateral and negative effects on the market of top-down regulation, as they tend to rely "on private entities to perform a variety of government functions while state authorities provide oversight and enforcement".[167] Consistently, the CoC on Countering Illegal Hate Speech Online reflects such a reliance and trust on private self-regulation and the will to reduce as much as possible the imposition of top-down burdens upon IT companies: an aspiration which, besides, similarly emerged from the 2018 EU Code of Practice on Disinformation.[168]

For its purposes, the CoC refers to the notion of hate speech as defined by Council Framework Decision 2008/913/JHA:[169] thus, "illegal hate speech" encompasses "all conduct publicly inciting to violence or hatred directed against a group of persons or a

---

Commission, 'The EU Code of Conduct on Countering Illegal Hate Speech Online' (*European Commission*) <https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en> accessed 30 May 2023.

[161] Code of Conduct on Countering Illegal Hate Speech Online 2016.

[162] Teresa Quintel and Carsten Ullrich, 'Self-Regulation of Fundamental Rights? The EU Code of Conduct on Hate Speech, Related Initiatives and Beyond' in Bilyana Petkova and Tuomas Ojanen (eds), *Fundamental Rights Protection Online: The Future Regulation of Intermediaries* (Edward Elgar Publishing 2020); Natalie Alkiviadou, 'Hate Speech on Social Media Networks: Towards a Regulatory Framework?' (2019) 28 Information & Communications Technology Law 19, 30–33; Frosio (n 111) 24–27.

[163] Barbora Bukovská, 'The European Commission's Code of Conduct for Countering Illegal Hate Speech Online' (TWG 2019) 2 <https://www.ivir.nl/publicaties/download/Bukovska.pdf> accessed 22 January 2023.

[164] See, among others, Linda AJ Senden and others, 'Mapping Self- and Co-Regulation Approaches in the EU Context: Explorative Study for the European Commission, DG Connect' (Utrecht University Repository, 2015) 5–11 <https://dspace.library.uu.nl/handle/1874/327305> accessed 10 July 2023.

[165] European Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, A Renewed EU Strategy 2011-14 for Corporate Social Responsibility' COM(2021) 681 final 5.

[166] Interinstitutional Agreement between the European Parliament, the Council of the European Union and the European Commission of 13 April 2016 on Better Law-Making, OJ L 123/1.

[167] Carl Vander Maelen, 'Hardly Law or Hard Law? Investigating the Dimensions of Functionality and Legislation of Codes of Conduct in Recent EU Legislation and the Normative Repercussions Thereof' (2022) 47 European Law Review 752, 754.

[168] Code of Practice on Disinformation 2018.

[169] Framework Decision 2008/913/JHA art 1, para 1, lett (a). See *supra*, §2.2.3.2.

member of such a group defined by reference to race, colour, religion, descent or national ethnic origin".[170] The agreement reads that the signatory IT companies recognize that they share "a collective responsibility and pride in promoting and facilitating freedom of expression throughout the online world".[171] Additionally, the CoC clearly states:

> The IT Companies support the European Commission and EU Member States in the effort to respond to the challenge of ensuring that online platforms do not offer opportunities for illegal online hate speech to spread virally. The spread of illegal hate speech online not only negatively affects the groups or individuals that it targets, it also negatively impacts those who speak out for freedom, tolerance and non-discrimination in our open societies and has a chilling effect on the democratic discourse on online platforms.[172]

The CoC, therefore, echoes the "militant" approach that characterizes the European point of view on hate speech.[173] Notably, its utterance is perceived as being intrinsically harmful not only for the direct targets of such hatred but, more in general, for democracy as a whole, as it often has the effect of quashing and silencing opposing voices.[174]

The main objective of the CoC is to offer an efficient guideline to IT companies and to grant them an instrument to share best practices to counter hate speech.[175] To this purpose, the Code envisages a set of commitments that signatory companies declare to undertake. Namely, these include the commitments to implement clear and effective processes to review notifications regarding illegal hate speech on their services, to review the majority of valid notifications for removal of illegal hate speech in less than 24 hours, to educate and raise awareness across users, to provide information on the procedures for submitting notices, as well as to intensify and foster cooperation amongst IT companies themselves and with Member States, civil society organizations, and the EU Commission.[176] In this respect, the CoC explicitly mentions the role of "trusted reporters" as important and trustworthy sources for the detection and removal of illegal hate speech content.[177]

---

[170] CoC on Illegal Hate Speech Online 1. Therefore, the CoC is quite narrow with regard to the grounds of discrimination considered, failing to consider, for example, sexual orientation, gender identity, sex, age or disability. As clarified *supra*, §2.2.3.2, such a limited scope of action is a consequence of the impossibility, for lack of EU competences, to adopt harmonizing criminal legislation concerning hate speech based on grounds of discrimination different from those addressed within the Framework Decision. Thus, "illegal" hate speech within the EU cannot encompass additional grounds of discrimination. Nevertheless, it could be argued that, being the CoC a self-regulatory instrument which does not affect the criminal treatment of purveyors of hate speech, its scope of action could have been significantly broader. Be that as it may, it appears from the monitoring reports of the EU Commission that the CoC may have had an indirect effect also on moderation practices concerning other forms of hate speech: see Didier Reynders, '7th Evaluation of the Code of Conduct' (European Commission 2022) 4 <https://commission.europa.eu/system/files/2022-12/Factsheet%20-%207th%20monitoring%20round%20of%20the%20Code%20of%20Conduct.pdf> accessed 30 May 2023.

[171] CoC on Illegal Hate Speech Online 1.

[172] ibid.

[173] See *supra*, §2.3.2.

[174] Such a perspective is also consistent with the alleged rationale behind most platform bans on hate speech: see more *infra*, §5.2.

[175] CoC on Illegal Hate Speech Online 2.

[176] ibid 2–3.

[177] ibid 3.

Additionally, consistently with the assumption that countering hate speech represents an essential task for the protection of democracy and (pluralistic) freedom of expression, the CoC states that the EU Commission and the signatory IT companies, acknowledging "the value of independent counter speech against hateful rhetoric and prejudice, aim to continue their work in identifying and promoting independent counter-narratives, new ideas and initiatives and supporting educational programs that encourage critical thinking".[178]

At the time of writing, since the implementation of the Code, the European Commission has conducted seven monitoring rounds concerning the practices of signatory IT companies. The data released by the latter showcase a general improvement over recent years concerning the average time taken by signatory companies in assessing notifications concerning the presence of online hate speech: whereas in December 2016, date of the first monitoring, only around 40% of the content flagged was reviewed within 24 hours, the rate had increased to 64.4% by the time of the seventh monitoring report, with an additional 12.7% dealt with within 48 hours, 21.5% within a week, and only 1.4% requiring over a week. However, the Commission's monitoring report shows at the same time how the last two years have seen a significant decrease in the rate of flagged content being reviewed within 24 hours: indeed, the fifth evaluation of June 2020 featured a rate of 90%, while the sixth evaluation of October 2021 reported a rate of 81%.[179] No clear explanation is given for such trends: it may be argued, however, that the peak of 2020-2021 was partly the result of the increased concerns about hate speech content being shared in the wake of the COVID-19 pandemic.

Such an increasing trend, characterized by a peak on the occasion of the fifth monitoring round, also emerges when looking at the rate of flagged content being removed. Namely, while the rate was of 71% in the 2020 report, the rate decreased to 63.6% in 2021 and to 62.5% in 2022.[180] Interestingly, the report explains that the removal rate is significantly higher for content flagged by "trusted flaggers" *vis-à-vis* content flagged by the general public:

> The divergence in removal rates between content reported using trusted reporting channels as compared to channels available to all users was 25.4 percentage points, much higher than the 13.5 percentage points observed in 2021. This seems to suggest that there is a growing difference of treatment between the notifications from general users and those sent through special channels for "trusted flaggers".[181]

---

[178] ibid.

[179] Reynders (n 170) 2. With respect to the performance of the various IT companies involved, the report clarifies: "TikTok assessed notifications in less than 24 hours in 91.7% of the cases and an additional 3.8% in less than 48 hours. The corresponding figures for YouTube are 83.3% and 7% and for Twitter 54.3% and 28.9%, respectively. Instagram had 56.9% and 5.9%, and Facebook 63.8% and 8.2%. Only TikTok had a better performance than in 2021, while all other platforms had a worse score than last year".

[180] ibid. Besides, the removal rate varies significantly from platform to platform, with YouTube being especially keen on removing flagged content: "YouTube removed 90.4% of the content flagged, Facebook 69.1%, TikTok 60.2%, Instagram 58.4% and Twitter 45.4%. Except for YouTube, all the other platforms had a lower removal rate than in 2021, although often with minor variations (for example, Facebook removed 70.2% of content in 2021 and Twitter 49.8%)".

[181] ibid.

The higher regard given to trusted flaggers' notifications, as opposed to single users' notifications, is also confirmed by data concerning, more broadly, the provision of feedback to the notifier. Indeed, the report stresses that "all platforms respond more frequently to notifications sent from the trusted flagger channels".[182]

However, it is worth noting that a higher removal rate does not necessarily equate to better moderation practices. Indeed, as has been long noted, such data is in many ways ambiguous, especially because of the gaps in information concerning the methodology employed for its collection and because of the absence of qualitative information concerning the rate of correct assessments. In other words, a higher removal rate could entail a higher rate of legal content being misrecognized as illegal hate speech and thus being subjected to over-removal.[183] The issue of the ambiguity inherent to resorting to removal rates as a proxy for effectiveness is even more significant in the light of the almost absolute absence of any reference, within the CoC, to procedural guarantees and safeguards protecting users' freedom of expression, namely that of minority groups and discriminated categories themselves. The CoC, in fact, limits itself to foreseeing a general commitment to further transparency with regard to the practices deployed by signatories. Nor is there, within the Commission's reports, a specific section dedicated to inquiring how IT companies tend to respond to users' complaints concerning possible cases of over-removal.

The limited importance given by the Code to this aspect is especially peculiar in light of the already mentioned goal of promoting freedom, tolerance, and non-discrimination. As argued in Chapter 2, the pursuit of such goals would benefit enormously from taking a substantive equality approach to the countering of hate speech online. Such a perspective, in turn, should nevertheless be supported by granting special attention towards the provision of remedies to individuals, and especially to members of minority or discriminated groups, for addressing cases of incorrect removal of their contents, with a view to ensuring that they are fully able to participate in the public debate.[184]

Overall, evaluating the positive impact of the CoC in countering hate speech online is thus quite a problematic task. According to Bukovská, it appears that, rather than pushing them to implement new and innovative strategies, "the Code of Conduct is primarily publicizing and formalising certain aspects of the internal processes that these IT companies already had in place prior to adoption of the Code to deal with complaints about certain

[182] ibid 4. Namely, Facebook gave feedback to general users in 80.7% of cases and to trusted flaggers or reporters in 97.7% of cases; these rates are respectively of 54.4% and 63.8% for Twitter (today X); 8% and 70.3% for YouTube; 63.1% and 98.1% for Instagram; and 71.99% and 91.3% for TikTok.

[183] "The monitoring reports consist of mere presentation of statistics of removals and statistical information on what grounds was the content removed … with no qualitative assessment whatsoever. There are no 'case studies' and examples of the types of content removed and maintained. This is a significant shortfall, given that such information would provide more insight into the assessment and decision-making and changes within the existing process of the IT companies since the adoption of the Code of Conduct… Overall, the monitoring reports provide very little information on the real effectiveness of the Code of Conduct system and what impact it has in protecting groups at risk of discrimination and hatred and ensuring that the right to freedom of expression is protected". Bukovská (n 163) 9.

[184] See *supra*, §2.5.3; *infra*, §5.4.

types of content".[185] It is arguable that, even though the CoC has certainly contributed to the promotion and sharing of good practices concerning online hate speech moderation and helped in shedding light on the importance of such practices in fostering a democratic society, grounded in the values of pluralism and equality, its practical efficacy in the pursuit of a hate speech- free and fully egalitarian Internet has been rather limited.

Be that as it may, as of October 2022, the EU legal framework addressing hate speech moderation and, more broadly, the moderation of illegal content online, is complemented by the Digital Services Act. Within the framework of the DSA, as will be highlighted below,[186] codes of conduct and codes of practice concerning speech moderation practices are explicitly recognized as parameters for the assessment of a provider's compliance with the set of due diligence obligations set by the new Regulation, thus promoting the adoption of a co-regulatory, rather than self-regulatory, model.[187] Indeed, as has been noted by Vander Maelen, the DSA reflects the general tendency of recent EU legislative acts (including, namely, the GDPR and the AVMSD) to promote a rather novel approach towards codes of conduct, by triggering a "hardening" of these traditionally "soft-law" tools.[188]

The adoption of the DSA, and thus of the new approach towards the implementation of codes of conduct, may thus affect importantly the EU framework on hate speech moderation, thus responding to the observed inefficiencies of the 2016 Code of Conduct.

3.5. The Digital Services Act

3.5.1. *The Digital Services Act package*

In December 2020, the Commission put forward the proposal for a Digital Services Act package, composed of two Acts eventually adopted between September and October

[185] Bukovská (n 163) 6. On the role of the practices of IT companies in defining the practices and sanctions affecting hate speech purveyors see, among others, Wilson and Land (n 79); Giovanni Ziccardi, *Online Political Hate Speech in Europe: The Rise of New Extremisms* (Edward Elgar Publishing 2020) 107–121; Roberto Bortone and Francesca Cerquozzi, 'L'Hate Speech al Tempo Di Internet' (2017) 68 Aggiornamenti Sociali 818, 821.

[186] See *infra*, §3.5.3.5.

[187] According to the Interinstitutional Agreement of 31 December 2003 on Better Law-Making, OJ C 321/1 paras 18, 22, co-regulation includes a mechanism by which a legislative act entrusts the attainment of the objectives defined by the legislative authority to parties which are recognized in the field (e.g., economic operators, social partners, non-governmental organizations, associations), whereas self-regulation consists of the possibility for these parties to adopt amongst themselves and for themselves common guidelines at European level (particularly codes of practice or sectoral agreements). See, in this respect, Senden and others (n 164) 5–6. Besides, the term "co-regulation", as argued by Marsden, "encompasses a range of different regulatory phenomena which have in common the fact that the regulatory regime is made up of a complex interaction of general legislation and a self-regulatory body": see Christopher T Marsden, *Internet Co-Regulation: European Law, Regulatory Governance and Legitimacy in Cyberspace* (Cambridge University Press 2011) 46. On the distinction between legal regulation, co-regulation, and self-regulation, see also Pagallo, Casanovas and Madelin (n 106), arguing nonetheless for the insufficiency of the concept of co-regulation to grasp on the variety of tools available in-between top-down regulation and self-regulation, and suggesting the resort to a "middle-out" approach.

[188] Vander Maelen (n 167).

2022, that is, the Digital Services Act and the Digital Markets Act (DMA).[189] According to the Commission, the package aims, on the one hand, at creating "a safer digital space in which the fundamental rights of all users of digital services are protected" while establishing, on the other hand, "a level playing field to foster innovation, growth, and competitiveness, both in the European Single Market and globally".[190] The purpose of the package, ultimately, is thus that of "taming the giants",[191] by promoting forms of "enhanced responsibility of digital platforms for addressing the different types of risk and harm that can result from their particular business models and market positions".[192] This is done, mainly, through a horizontal, rather than sectoral, perspective and following a strategy that is focused especially upon procedure.[193]

The DMA introduces rules to ensure the contestability and fairness of digital markets through an approach which is "prescriptive" rather than merely "proscriptive",[194] meaning that its goal is that of developing a framework which does not simply react, *ex post*, to the malfunctioning of the market, but sets specific rules to avoid, *ex ante*, such malfunctioning. Most notably, the DMA provides for an innovative competition law framework with respect to "gatekeepers"[195] providing "core platform services".[196]

The DSA complements the DMA's competition law perspective by addressing and regulating the responsibilities of providers of intermediary services to ensure a safer digital environment for the recipients of those services. The DSA seeks, therefore, to promote and guarantee the protection of individuals' fundamental rights in the context of the digital landscape, including the right to freedom of expression and the right to non-discrimination as well as consumer rights,[197] while countering the dissemination of illegal content and the perpetration of illegal conducts across the Internet. The DSA thus represents a seminal step for the purposes of the EU framework on content moderation regulation, representing a significant departure from the previous system emerging from the ECD. Indeed, in the opening to the Explanatory Memorandum accompanying the DSA proposal, the Commission expressly stated:

> Since the adoption of Directive 2000/31/EC (the "e-Commerce Directive"), new and innovative information society (digital) services have emerged, changing the daily lives of

[189] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act), OJ L 265/1.

[190] European Commission, 'The Digital Services Act Package' (*European Commission*, 12 May 2023) <https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package> accessed 2 June 2023.

[191] Martin Eifert and others, 'Taming the Giants: The DMA/DSA Package' (2021) 58 Common Market Law Review 987.

[192] ibid 989.

[193] Pollicino, 'Potere Digitale' (n 114) 439.

[194] Natalia Moreno Belloso and Nicolas Petit, 'The EU Digital Markets Act (DMA): A Competition Hand in a Regulatory Glove' (2023) 48 European Law Review 391, 402.

[195] DMA art 3.

[196] ibid 2(2).

[197] Caroline Cauffman and Catalina Goanta, 'A New Order: The Digital Services Act and Consumer Protection' (2021) 12 European Journal of Risk Regulation 758; Jorge Morais Carvalho, Francisco Arga e Lima and Martim Farinha, 'Introduction to the Digital Services Act, Content Moderation and Consumer Protection' (2021) 3 Revista de Direito e Tecnologia 71.

> Union citizens and shaping and transforming how they communicate, connect, consume and do business. Those services have contributed deeply to societal and economic transformations in the Union and across the world. At the same time, the use of those services has also become the source of new risks and challenges, both for society as a whole and individuals using such services.[198]

3.5.2. *The rules on the liability of providers of intermediary services*

As a matter of fact, the DSA, in its Chapter II dedicated to the "liability of providers of intermediary services", tends to replicate the system established by the ECD: indeed, Articles 4-6 and 8 DSA, concerning the liability of the providers of mere-conduit, caching, and hosting services as well as the prohibition regarding the imposition of general monitoring obligations, are substantially identical to Articles 12-15 ECD, with the exception of excluding from the favourable regime set by Article 6 on hosting service providers those specific cases concerning

> the liability under consumer protection law of online platforms that allow consumers to conclude distance contracts with traders, where such an online platform presents the specific item of information or otherwise enables the specific transaction at issue in a way that would lead an average consumer to believe that the information, or the product or service that is the object of the transaction, is provided either by the online platform itself or by a recipient of the service who is acting under its authority or control.[199]

The "safe harbour" approach towards intermediary liability is thus overall confirmed within the DSA and is, in fact, complemented by a new provision which, mimicking the Good Samaritan clause contained within Section 230 of the US CDA,[200] rules that providers of intermediary services shall not be deemed ineligible for the liability exemptions "solely because they, in good faith and in a diligent manner, carry out voluntary own-initiative investigations into, or take other measures aimed at detecting, identifying and removing, or disabling access to, illegal content", nor solely because they take the "necessary measures" to comply with EU and/or national law, including the DSA itself.[201]

The DSA, therefore, sets *prima facie* a background framework for ISP liability that does not revolutionize the pre-existing system: in fact, the adoption of a good Samaritan clause rather moves in a direction that is quite favourable to ISPs, as it further legitimizes their content moderation and curation practices.

Nevertheless, the DSA introduces a first, rather important, novelty by clarifying that the exemptions from liability, as regulated by Articles 4-6, are not applicable whenever, "instead of confining itself to providing the services neutrally by a merely technical and automatic processing of the information … the provider of intermediary services plays an

---

[198] European Commission, 'Communication of 15 December 2020, 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 1.

[199] DSA art 6, para 3.

[200] See more *infra*, §4.4.2.

[201] DSA art 7. See, with regard to the relationship between arts 6 and 7 of the DSA, Jacob van de Kerkhof, 'Good Faith in Article 6 Digital Services Act (Good Samaritan Exemption)' (*The Digital Constitutionalist*, 15 February 2023) <https://digi-con.org/good-faith-in-article-6-digital-services-act-good-samaritan-exemption/> accessed 24 December 2023.

active role of such a kind as to give it knowledge of, or control over, that information".[202] This way, the DSA confirms and crystallizes that strand of case law initiated by the CJEU in *Google France*[203] that excludes from the safe harbour regime all providers not acting neutrally.[204]

Additionally, the DSA explicitly recognizes that national judicial or administrative authorities have the power to issue orders "to act against one or more specific items of illegal content",[205] as well as orders "to provide specific information about one or more specific individual recipients of the service".[206] These orders, consistently with the prohibition of general monitoring obligations, should be sufficiently well-defined and specific both with respect to their material and territorial scope: however, in the wake of *Glawischnig-Piesczek*,[207] it is rather likely that Member States may enjoy a significant degree of discretion in this respect.

The most significant novelty of the DSA, nevertheless, is the choice of building on top of the rules on liability a whole new set of due diligence obligations, governing and regulating the responsibilities of providers of intermediary services with respect to the fostering of a "transparent and safe online environment".[208] The implementation of these new obligations, moreover, is complemented, as already mentioned above, by the explicit recognition of the role of codes of conduct and of codes of practice, with a view to promoting a shift from a self-regulatory to a co-regulatory approach towards them.[209] The following subsection analyses more in-depth the resulting framework.

3.5.3. *The new due diligence obligations for a transparent and safe online environment*

The new due diligence obligations "for a transparent and safe online environment", which are listed by the DSA within its dedicated Chapter III and are to be complied with under

[202] DSA rec 18.

[203] *Google France* (n 84). See *supra*, §3.4.2.

[204] Folkert Wilman, 'Between Preservation and Clarification: The Evolution of the DSA's Liability Rules in Light of the CJEU's Case Law' in Joris van Hoboken and others (eds), *Putting the DSA into Practice: Enforcement, Access to Justice, and Global Implications* (Verfassungsbooks 2023) 41.

[205] DSA art 9, para 1.

[206] ibid 10, para 1.

[207] *Glawischnig-Piesczek* (n 101). See *supra*, §3.4.2.

[208] See, in this respect, Christoph Busch, 'Regulating the Expanding Content Moderation Universe: A European Perspective on Infrastructure Moderation' (2022) 27 UCLA Journal of Law & Technology 32, 53–56. According to Husovec and Roche Laguna, the major contribution of the DSA is precisely the choice of splitting due diligence obligations from the liability of underlying content. Indeed, "prior to DSA, most of the laws tried to influence the providers' behaviour by threatening with joint liability for what their users do … the courts often faced a binary decision: impose a duty of care or deny it and confirm a liability exemption. The DSA ends this binary. It comes up with its own expectations formulated as due diligence obligations. Providers violating them can be held to account. These legal obligations are separate from those of their users. Violations of DSA have no bearing on the provider's preservation of the liability exemptions. Even providers who are not liable for users remain accountable for their own failings to be diligent". Martin Husovec and Irene Roche Laguna, 'Digital Services Act: A Short Primer' (SSRN, 5 July 2022) 1 <https://papers.ssrn.com/abstract=4153796> accessed 6 June 2023.

[209] DSA art 45. See *infra*, §3.5.3.5.

penalty of "effective, proportionate and dissuasive" pecuniary sanctions,[210] are most notably characterized by an asymmetric approach, meaning that they are not entirely applicable to each and every provider of intermediary services, but are in fact scaled and differentiated based on the services those providers offer and/or upon the number of recipients they reach within the EU.[211] Through such an asymmetric strategy, the DSA aims at implementing "a supervised risk management approach",[212] thus confirming the general trend characterizing recent EU content moderation regulation favouring a risk-based approach.[213]

The DSA identifies, notably, four tiers of risk, each of which calls for an additional layer of due diligence obligations compared to the previous one. The first tier of due diligence obligations is thus represented by a set of rules applicable to all providers of intermediary services, including providers of mere-conduit and caching services, irrespective of their dimensions. The second tier is represented by providers of hosting services, consisting, in the words of the Regulation mimicking the ECD, "of the storage of information provided by a recipient".[214]

The third tier consists of the rules applicable to the providers of so-called "online platforms", unless they qualify as micro or small enterprises as defined by Recommendation 2003/361/EC,[215] that is, unless they employ less than 50 persons and have an annual turnover and/or annual balance sheet total that does not exceed 10 million euros.[216] The DSA, for its purposes, defines an "online platform" as a specific kind of hosting provider that does not simply store information but, at the request of the recipient of its services, also disseminates it to the public.[217] The concept itself of "dissemination to the public" had led, prior to the adoption of the Regulation, to some interpretive challenges: for instance, it was not clear to what extent the providers of interpersonal communications services should be included within the category of the providers of online platforms.[218] For this reason, the final text of the DSA clarifies that the notion of "dissemination to the public

[210] ibid 52.

[211] Joan Barata and others, 'Unravelling the Digital Services Act Package' (European Audiovisual Observatory 2021) 2021–1 35 <https://rm.coe.int/iris-special-2021-01en-dsa-package/1680a43e45> accessed 9 September 2022.

[212] European Commission, 'DSA Proposal' (n 198) 11.

[213] Zohar Efroni, 'The Digital Services Act: Risk-Based Regulation of Online Platforms' (*Internet Policy Review*, 16 November 2021) <https://policyreview.info/articles/news/digital-services-act-risk-based-regulation-online-platforms/1606> accessed 8 June 2023. See also De Gregorio and Dunn (n 116) 483–488, arguing that the DSA tends to conflate elements of both a top-down and a bottom-up perspective on the risk-based approach as it identifies directly the tiers of risk while granting nonetheless a significant leeway for providers of intermediary services in defining, ultimately, the mechanisms, tools, and measures to be implemented.

[214] DSA art 6, para 1.

[215] ibid 19, para 1.

[216] Commission Recommendation 2003/361/EC of 6 May 2003 concerning the definition of micro, small and medium-sized enterprises, OJ L124/36 art 2.

[217] DSA art 3, lett (i), which clarifies, nonetheless, that a service shall not qualify as an online platform if the activity of storing information and disseminating it to the public is a minor and purely ancillary feature of another service or a minor functionality of the principal service and, for objective and technical reasons, that feature or minor functionality cannot be used without the other, main, service.

[218] Busch (n 208) 57–58.

> should entail the making available of information to a potentially unlimited number of persons, meaning making the information easily accessible to recipients of the service in general without further action by the recipient of the service providing the information being required, irrespective of whether those persons actually access the information in question. Accordingly, where access to information requires registration or admittance to a group of recipients of the service, that information should be considered to be disseminated to the public only where recipients of the service seeking to access the information are automatically registered or admitted without a human decision or selection of whom to grant access. Interpersonal communication services … such as emails or private messaging services, fall outside the scope of the definition of online platforms as they are used for interpersonal communication between a finite number of persons determined by the sender of the communication. However, the obligations set out in this Regulation for providers of online platforms may apply to services that allow the making available of information to a potentially unlimited number of recipients, not determined by the sender of the communication, such as through public groups or open channels. Information should be considered disseminated to the public within the meaning of this Regulation only where that dissemination occurs upon the direct request by the recipient of the service that provided the information.[219]

Therefore, while interpersonal communications services are generally considered to be outside the scope of the notion of online platforms, the offering of openly accessible channels such as those provided by Telegram will arguably trigger the responsibility to comply to the obligations set for online platforms.[220]

The fourth tier identified by the DSA is, finally, represented by the providers of so-called "very large online platforms" (VLOPs) and "very large online search engines" (VLOSEs), designated as such by a decision of the Commission when an online platform or an online search engine[221] "have a number of average monthly active recipients of the service in the Union equal or higher than 45 million".[222] For this purpose, the Regulation demanded the publication by 17 February 2023 of their user data:[223] based on such information, on 25 April 2023, the Commission adopted its first designation decision, designating 17 VLOPs and 2 VLOSEs.[224] Four months after, on 25 August 2023, the DSA

[219] DSA rec 14.

[220] This expanded definition of "dissemination to the public" and, consequently, of providers of "online platforms" is indeed particularly significant in the light of the reported frequent use of Telegram channels for the purposes of sharing illegal and harmful content such as non-consensual intimate images, alt-right content, and disinformation. See Silvia Semenzin and Lucia Bainotti, 'The Use of Telegram for Non-Consensual Dissemination of Intimate Images: Gendered Affordances and the Construction of Masculinities' (2020) 6 Social Media + Society 2056305120984453; Valerio Mazzoni, 'Far Right Extremism on Telegram: A Brief Overview' (*European Eye on Radicalization*, 14 March 2019) <https://eeradicalization.com/far-right-extremism-on-telegram-a-brief-overview/> accessed 7 June 2023; Martin Fertmann and Matthias C Kettemann (eds), *Viral Information: How States and Platforms Deal with Covid-19-Related Disinformation; an Exploratory Study of 20 Countries* (Verlag Hans-Bredow-Institut 2021) 11.

[221] Pursuant to the Regulation, the term "online search engine" refers to "an intermediary service that allows users to input queries in order to perform searches of, in principle, all websites, or all websites in a particular language, on the basis of a query on any subject in the form of a keyword, voice request, phrase or other input, and returns results in any format in which information related to the requested content can be found". DSA art 3, lett (j).

[222] ibid 33, para 1.

[223] ibid 24, para 2. Subsequently, online platforms shall publish user data once every six months.

[224] Alibaba AliExpress; Amazon Store; Apple AppStore; Booking.com; Facebook; Google Play; Google Maps; Google Shopping; Instagram; LinkedIn; Pinterest; Snapchat; TikTok; X; Wikipedia; YouTube;

became applicable for providers of VLOPs and VLOSEs in advance pursuant to Article 92 of the Regulation.[225]

Each level of risk entails increased due diligence obligations which move, overall, in at least three directions. First, the DSA aims at promoting and fostering transparency with respect to content moderation and content curation practices. Second, the Regulation sets important and new procedural rules orienting such practices, with a view to ensuring as much as possible the protection of the fundamental rights of the recipients of intermediary services. Third, the DSA introduces a wide array of obligations specifically addressed at actively reducing the levels of risk entailed by the provision of any intermediary service.

3.5.3.1. Provisions applicable to all providers of intermediary services

All providers of intermediary services are required, first of all, to designate a single point of contact to enable direct contact with competent domestic authorities, including most notably the designated national Digital Services Coordinator (DSC),[226] the Commission, and the newly established European Board for Digital Services (EBDS),[227] as well as a single point of contact enabling recipients of the service to communicate directly and rapidly with them.[228] Moreover, all providers that do not have an establishment within the EU but offer their services within it must designate in writing a legal or natural person to act as their representatives: such representatives may be held liable for the providers' failure to comply with the obligations set under the Regulation, without prejudice to the possibility of initiating legal actions against the providers themselves.[229]

Additionally, all providers of intermediary services are required to make available to recipients thorough information concerning any restrictions they may impose in relation to their services within their terms and conditions, indicating namely the policies, procedures, measures, and tools implemented for the purposes of content moderation, including the resort to algorithmic decision-making and human review. In order to ensure full transparency, the terms and conditions must be set out "in clear, plain, intelligible, user-friendly and unambiguous language" and be "publicly available in an easily accessible

---

Zalando. Designated VLOSEs are Bing and Google Search. See European Commission, 'Digital Services Act: Commission Designates First Set of Very Large Online Platforms and Search Engines' (*European Commission*, 25 April 2023) <https://ec.europa.eu/commission/presscorner/detail/en/IP_23_2413> accessed 12 June 2023. On 20 December 2023, the Commission also designated Pornhub, XVideos, and Stripchat as VLOPs: see European Commission, 'Commission Designates Second Set of Very Large Online Platforms under the Digital Services Act' (*European Commission*, 20 December 2023) <https://digital-strategy.ec.europa.eu/en/news/commission-designates-second-set-very-large-online-platforms-under-digital-services-act> accessed 28 December 2023.

[225] "This Regulation shall apply to providers of very large online platforms and of very large online search engines designated … from four months after the notification to the provider concerned … where that date is earlier than 17 February 2024".

[226] DSA arts 49–51. On the nature and role of DSCs, see Julian Jaursch, 'Platform Oversight: Here Is What a Strong Digital Services Coordinator Should Look Like' in Joris van Hoboken and others (eds), *Putting the DSA into Practice: Enforcement, Access to Justice, and Global Implications* (Verfassungsbooks 2023) 94–98.

[227] DSA arts 61–63.

[228] ibid 11–12.

[229] ibid 13.

and machine-readable format".[230] With respect to terms and conditions, however, the DSA does not simply require that providers comply with these transparency requirements. Indeed, the Regulation also addresses from a substantive perspective the matter of their enforcement, stating that providers

> shall 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, including the fundamental rights of the recipients of the service, such as the freedom of expression, freedom and pluralism of the media, and other fundamental rights and freedoms as enshrined in the Charter.[231]

The importance of such a provision emerges clearly if one considers the many challenges raised by the inherently private nature of platform governance which may significantly affect the fundamental rights and freedoms of the recipients of intermediary services.[232] This holds true, especially, when it comes to moderating and reducing the phenomenon of online hate speech which represents a particularly sensitive matter and faces the concrete risk of impacting enormously the right to freedom of expression of minority and discriminated groups.[233]

In this respect, it is worth noting that, while the provision regrettably fails to mention explicitly the need to guarantee the respect of the principle of non-discrimination, a more explicit focus on the promotion of such a principle is contained within the recitals of the Regulation, stating that providers must act "in a non-arbitrary and non-discriminatory manner"[234] when enforcing their terms and conditions. Such wording, especially when applied to providers of hosting services and online platforms, could be interpreted in the sense of requiring them to actively work towards ensuring the equal participation of all users in the public debate online, in line with an approach towards speech governance in general, and hate speech governance in particular, oriented towards (substantive) equality.[235]

However, the efficacy of such an obligation in the long term is yet to be assessed. Literature, in particular, has highlighted on the one hand the challenges related to the enforceability of the new rule,[236] while noting that taking an approach based on

---

[230] ibid 14, para 1.

[231] ibid 14, para 4. The provision, in fact, seemingly aims at promoting, through a legislative act, a framework oriented towards the horizontal effect of fundamental rights as enshrined within the CFREU. This represents a rather significant step taken by the EU lawmaker, as the history of the horizontal effect of the Charter has for long been a prerogative of the CJEU. See, in this respect, Eleni Frantziou, 'The Horizontal Effect of the Charter: Towards an Understanding of Horizontality as a Structural Constitutional Principle' (2020) 22 Cambridge Yearbook of European Legal Studies 208.

[232] See *supra*, §3.2.2.

[233] See *supra*, §2.5.3; *infra* §5.4.

[234] DSA rec 47.

[235] See *supra*, §2.5.

[236] Notably, although art 53 provides that "recipients of the service and any body, organisation or association mandated to exercise the rights conferred by th[e] Regulation on their behalf shall have the right to lodge a complaint against providers of intermediary services … with the Digital Services Coordinator" (and such a right is complemented by a right to the compensation of damages under art 54), it has been argued that "it seems that the DSC has extensive discretionary power to decide whether or not to pick up the complaint, which is particularly relevant since the provision does not create a separate right of action for

individualized human rights prerogatives may not be fully adequate to address collective and systemic harms such as those faced by marginalized communities.[237]

Finally, all providers of intermediary services that do not qualify as small or micro enterprises must publish at least once a year, in a machine-readable format and in an easily accessible manner, "easily comprehensible reports on any content moderation that they engaged in during the relevant period".[238] These reports must include quantitative data, namely: the number of orders received by Member States' authorities to act against illegal content or to provide information and the median time taken to comply with those orders; in the case of hosting providers, the number of notices received concerning the potential presence of illegal content upon their infrastructures, as well as data concerning the actions taken and the timing of response; the number of complaints received by recipients with respect to the enforcement of terms and conditions; in the case of online platforms, information concerning the platform's responsiveness.[239] Moreover, the yearly transparency reports must contain

> meaningful and comprehensible information about the content moderation engaged in at the providers' own initiative, including the use of automated tools, the measures taken to provide training and assistance to persons in charge of content moderation, the number and type of measures taken that affect the availability, visibility and accessibility of information provided by the recipients of the service and the recipients' ability to provide information through the service, and other related restrictions of the service.[240]

Specific information must also be included within the providers' transparency reports whenever they make use of automated means for content moderation, "including a qualitative description, a specification of the precise purpose, indicators of the accuracy and the possible rate of error of the automated means used … and any safeguards applied".[241]

3.5.3.2. Provisions applicable to providers of hosting services

Providers of hosting services are subject to three main obligations pursuant to the DSA: the creation of a notice and action mechanism; the provision of a statement of reasons whenever they take restrictive measures affecting the recipients of their services; and the

---

the recipient of the service" and that "as the DSCs will handle the complaint in accordance with national administrative law, we can expect in this context divergences similar to those observed in for Data Protection Authorities in the GDPR across different Member States". João Pedro Quintais, Naomi Appelman and Ronan Ó Fahy, 'Using Terms and Conditions to Apply Fundamental Rights to Content Moderation' (2023) 24 German Law Journal 881, 905.

[237] ibid 30–31. Griffin underscores the limited significance of a system purely based on the provision of individual remedies rather than focused on confronting the way "tech platforms shape social and cultural norms – for example, by promoting content which reinforces gender stereotypes". This failure to capture "the conceptual framework of rights and discrimination" may indeed lead to an insufficient response against fundamental issues determined by platform governance: for instance, the fact that queer users, people of colour and sex workers are more than often subjected to forms of "shadow banning". See Rachel Griffin, 'Rethinking Rights in Social Media Governance: Why fundamental rights are not enough to remedy the injustices of contemporary social media' (*Verfassungsblog*, 25 February 2022) <https://verfassungsblog.de/rethinking-rights/> accessed 8 June 2023.

[238] DSA art 15, para 1.

[239] ibid 15, para 1, letts (a)-(b), (d).

[240] ibid 15, para 1, lett (c).

[241] ibid 15, para 1, lett (e).

referral to the domestic judicial authorities of "any information giving rise to a suspicion that a criminal offence involving a threat to the life or safety of a person or persons has taken place, is taking place or is likely to take place" they may be aware of.[242] Among these, the first two new obligations are of particular interest.

First, providers of hosting services must put in place easy to access and user-friendly notice and action mechanisms to allow any individual or entity to notify them of the presence on their service of specific items of information they consider to constitute illegal content.[243] The structure of such mechanisms should facilitate the submission of well-substantiated information containing all necessary information such as the reasons why the content flagged is considered to be illegal, the exact electronic location of the information (e.g., the exact URL or URLs), the name and email address of the submitting individual or entity, and a statement confirming the *bona fide* belief that the information and allegations contained in the notice are accurate and complete.[244]

The significance and importance of this provision is represented by the consequences in terms of liability for third-party content that the DSA attaches to the new obligatory mechanism of notice and action. Indeed, once the provider of hosting services has received a notice of the nature just described, it shall be considered to have "actual knowledge or awareness for the purpose of Article 6 in respect to the specific item of information concerned" whenever such notice is substantiated enough to "allow a diligent provider … to identify the illegality of the relevant activity or information without a detailed legal examination".[245] In other words, the submission of a notice under the new notice and action mechanism has the effect of excluding that provider of hosting services from benefitting from the exemption set out in Article 6. Providers are, therefore, required to process all notices in a timely, diligent, non-arbitrary and objective manner, all the time keeping the individuals or entities that submitted them posted on the state of advancement of the verification procedure: in the case of use of automated means for the processing of or decision about the notices submitted, they will have to disclose such information to the notifier.[246]

Second, the DSA also obliges providers of hosting services to provide a clear and specific statement of reasons whenever a recipient is reached by a restrictive measure for having uploaded information that is illegal or in violation of the terms and conditions: such measures include, notably, any "restrictions of the visibility of specific items of information provided by the recipient of the service, including removal of content, disabling access to content, or demoting content";[247] restrictions concerning monetary payments;

[242] ibid 18.
[243] ibid 16, para 1.
[244] ibid 16, para 2.
[245] ibid 16, para 3.
[246] ibid 16, paras 4-6.
[247] The reference to the demotion of content extends the scope of the provision to content curation practices on top of content moderation practices: on the concept of content curation, see *supra*, §2.4.3; *infra*, §5.3.1. Indeed, as underscored by the DSA itself, "restriction of visibility may consist in demotion in ranking or in recommender systems, as well as in limiting accessibility by one or more recipients of the

suspension or termination of the service in whole or in part; and suspension or termination of the recipient's account.[248]

Clearly, the statement of reasons shall contain first and foremost information concerning the restrictive measures adopted; the grounds of the decision (including whether the decision was taken based on a notice received from third parties); the legal or contractual sources based on which the content provided was considered to be illegal or in violation of the provider's terms and conditions; the available possibilities for redress. However, the DSA specifies that the provider should also indicate whether the decision was taken based on the use of automated means (and whether the content had been detected or identified through automated means in the first place):[249] such a specification is particularly welcome in the light of the many challenges raised by the resort to AI for the purposes of content moderation and curation.

3.5.3.3. Provisions applicable to providers of online platforms

Providers of online platforms are the targets of a significant range of additional due diligence obligations, aimed at providing additional procedural guarantees for the recipients of services, at complementing the duties connected to the countering of illegal and harmful content, and at promoting transparency. Moreover, as opposed to the original proposal presented by the Commission in December 2020, the final text of the Regulation contains a specific section dedicated to a restricted group of providers of online platforms, that is, those that allow consumers to conclude distance contracts with traders (so-called "online marketplaces"): the provisions contained therein, clearly, are mainly aimed at protecting and guaranteeing consumer rights.[250]

With respect to the additional procedural guarantees, the DSA sets, notably, the obligation to establish an effective internal complaint-handling system enabling either the recipients of the service who have been reached by restrictive measures or the individuals or entities that have submitted a notice to the platform, concerning the presence of content they deem to be illegal or in violation of the terms and conditions, to lodge a complaint, electronically and free of charge, against the decision taken by the online platform. Such a possibility must be ensured for a period of at least six months from the day on which the recipient is informed about the decision.[251] The Regulation clarifies, notably, that the platform must handle the complaints "in a timely, non-discriminatory, diligent and non-arbitrary manner"[252] and that providers must ensure that the new decision is taken "under the supervision of appropriately qualified staff, and not solely on the basis of automated

service or blocking the user from an online community without the user being aware ('shadow banning')". DSA rec 55. On the DSA's approach to shadow banning, see Paddy Leerssen, 'An End to Shadow Banning? Transparency Rights in the Digital Services Act between Content Moderation and Curation' (2023) 48 Computer Law & Security Review 105790.

[248] DSA art 17, para 1.

[249] ibid 17, para 3.

[250] ibid 29–32.

[251] ibid 20, paras 1-2.

[252] ibid 20, para 4.

means".[253] Arguably, the latter requirement is consistent with the provision of the General Data Protection Regulation (GDPR) on automated individual decision-making.[254]

Nevertheless, the internal complaint-handling system is not the only avenue available to the recipients of the service or to the subjects who have submitted a notice to the online platform. Indeed, the DSA also envisages for them the possibility to refer the decision, free of charge, [255] to out-of-court dispute settlement bodies,[256] certified in accordance with the Regulation itself based on their compliance with a series of conditions proving their expertise, independence, and impartiality.[257] Additionally, as clarified by the Regulation, both the resort to the internal complaint-handling system or to an out-of-court settlement does not prevent the parties from initiating at any stage proceedings before the national judicial authority.[258]

As for the countering of illegal and harmful content, the DSA complements the rules set in general for hosting providers, first, by introducing the category of "trusted flaggers" – entities recognized as such by the domestic DSC based on criteria of particular expertise, competence, and independence, and whose notices submitted under the notice and action mechanism must be given priority and processed and decided upon without undue delay.[259] Second, the DSA mandates that providers of online platforms "suspend, for a reasonable period of time and after having issued a prior warning, the provision of their services to recipients … that frequently provide manifestly illegal content".[260] Third, the Regulation provides for additional protection of minors, ordering to put in place "appropriate and proportionate measures to ensure a high level of privacy, safety, and security of minors, on their services".[261]

Finally, the DSA provides for additional transparency requirements applicable to online platforms, stating that their transparency reports should also include data concerning, on the one hand, the out-of-court disputes involving them and, on the other hand, the suspensions inflicted upon recipients for misuse, as well as information about the number of recipients they reach.[262] Additionally, the Regulation sets rules concerning the online interface design and organization, clarifying that interfaces must not deceive or

[253] ibid 20, para 6.

[254] 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), OJ L/119 art 22.

[255] DSA art 21, para 5.

[256] ibid 21, para 1. The providers of online platforms must inform of such a possibility the recipients of their services as well as the individuals or entities that have submitted a notice to them.

[257] ibid 21, para 3. It is up to the domestic DSC to deal with the certification of out-of-court dispute settlement bodies.

[258] DSA rec 59.

[259] ibid 22.

[260] ibid 23, para 1; specularly, ibid 23, para 2, requires providers of online platforms to also suspend, for a reasonable time and after having issued a prior warning, the processing of notices and complaints submitted through the notice and action mechanisms and internal complaint-handling systems by individuals or entities or by complainants that frequently submit notices or complaints that are manifestly unfounded.

[261] DSA art 28.

[262] ibid 24.

manipulate recipients of the service and must not distort or impair their ability to make free and informed choice,[263] as well as rules governing online advertising.[264]

Most interestingly, the DSA also demands that online platforms set out in their terms and conditions, in plain and intelligible language, whether they use recommender systems,[265] the main parameters used by these systems, and "any options for the recipients of the service to modify or influence those main parameters", including at least "the criteria which are most significant in determining the information suggested" and the "reasons for the relative importance of those parameters". On top of such transparency requirements, the Regulation also establishes that, "where several options are available … for recommender systems that determine the relative order of information presented", providers must "make available a functionality that allows the recipient of the service to select and to modify at any time their preferred option".[266]

With respect to the last point, the DSA adds an additional obligation only applicable to VLOPs and VLOSEs: while providers of online platforms are not actually obliged to provide multiple choices to recipients concerning the recommender systems applicable to them, VLOPs and VLOSEs are explicitly required to "provide at least one option for each of their recommender systems which is not based on profiling"[267] as defined by the GDPR.[268]

The adoption of regulatory measures concerning recommender systems, which, in the first text presented by the Commission, only concerned VLOPs,[269] showcases the EU lawmaker's increasing awareness of the importance of curation, on top of "*stricto sensu*" moderation,[270] for the concrete dissemination of content.[271] Moreover, the Regulation itself reads:

> A core part of the online platform's business is the manner in which information is prioritised and presented on its online interface to facilitate and optimise access to information for the recipients of the service. This is done, for example, by algorithmically suggesting, ranking and prioritising information, distinguishing through text or other visual representations, or otherwise curating information provided by recipients. Such recommender systems can have a significant impact on the ability of recipients to retrieve and interact with information online … They also play an important role in the amplification of certain

[263] ibid 25.

[264] ibid 26.

[265] A recommender system is defined as "a fully or partially automated system used by an online platform to suggest in its online interface specific information to recipients of the service or prioritise that information, including as a result of a search initiated by the recipient of the service or otherwise determining the relative order or prominence of information displayed" in ibid 3, lett (s). On recommender systems, see *infra*, §5.3.1.

[266] ibid 27.

[267] ibid 38.

[268] GDPR art 4, n (4).

[269] Natali Helberger and others, 'Regulation of News Recommenders in the Digital Services Act: Empowering David against the Very Large Online Goliath' (*Internet Policy Review*, 26 February 2021) <https://policyreview.info/articles/news/regulation-news-recommenders-digital-services-act-empowering-david-against-very-large> accessed 13 June 2023.

[270] See *supra*, §2.4.3; *infra*, §5.3.1.

[271] Indeed, the DSA, in defining "content moderation", specifically includes within the notion all "measures taken that affect the availability, visibility, and accessibility … such as demotion". DSA art 3, lett (t).

> messages, the viral dissemination of information and the stimulation of online behaviour.[272]

Such awareness, also confirmed by the additional discipline envisaged for VLOPs and VLOSEs, arguably represents a strong suit of the new legislation, opening the road for a more complete and organic governance of online speech. Particularly, an approach oriented towards the regulation of content curation on top of content moderation could have a positive impact on reducing the spread of hate speech content, while helping promote diversity of opinions and forms of counter-speech.[273]

3.5.3.4. Obligations for providers of very large online platforms and of very large online search engines to manage systemic risks

Apart from the additional rule on recommender systems, the DSA's discipline of VLOPs and VLOSEs is characterized first of all by the introduction of an obligation to put in place a mechanism for the assessment and mitigation of the "systemic risks in the Union stemming from the designing or functioning of their service and its related systems, including algorithmic systems, or from the use made of their service".[274] The risk assessment, to be carried out once every year and whenever a new functionality is deployed that is likely to have a critical impact, must be specific to the services provided and proportionate to the systemic risks, taking into consideration their severity and probability.

The systemic risks to be considered include, clearly, the risk of dissemination of illegal content. However, the DSA extends sensitively the scope of action of the obligation, requiring it to address also other issues that are not directly linked to the commission of unlawful conduct. Most notably, providers of VLOPs and VLOSEs must assess whether the way they provide their services entails the possibility of actual or foreseeable negative effects on the exercise of fundamental rights as protected by the CFREU, including *inter alia* the rights to human dignity, to freedom of expression and information, and to non-discrimination. Additionally, it is necessary to assess the possibility of actual or foreseeable negative effects on civic discourse, electoral processes, and public security, as well as in relation to gender-based violence, the protection of public health and minors and serious negative consequences to the person's physical and mental well-being.[275]

The DSA, therefore, tends to conflate together criminal and harmful materials when it comes to such fourth-tier due diligence obligation. Moreover, the factors to be taken into account when assessing such risks include the design of recommender systems and of any other relevant algorithmic system, their content moderation systems, the applicable terms and conditions and their enforcement, the implementation of systems for selecting and

[272] DSA rec 70.

[273] Recommender systems can, indeed, play an important role in the promotion of pluralism and diversity of content, as stressed, namely, by Natali Helberger, Kari Karppinen and Lucia D'Acunto, 'Exposure Diversity as a Design Principle for Recommender Systems' (2018) 21 Information, Communication & Society 191; Natali Helberger and others, 'A Freedom of Expression Perspective on AI in the Media – with a Special Focus on Editorial Decision Making on Social Media Platforms and in the News Media' (2020) 11 European Journal of Law and Technology 1.

[274] DSA art 34, para 1.

[275] ibid.

presenting advertisements, and the deployment of data-related practices of the provider. The potential impact of intentional manipulation should also be considered, together with specific regional or linguistic aspects.[276]

Once the assessment has been carried out, providers of VLOPs and VLOSEs must "put in place reasonable, proportionate and effective mitigation measures, tailored to the specific systemic risks identified".[277] The Regulation mentions a wide array of potential tools that providers may choose to employ, many of which concern the adaptation of the terms and conditions and of the way these are enforced, as well as the adaptation of content moderation and curation practices, especially when AI-driven.

On top of being highly symbolic of the risk-based approach characterizing the DSA, these obligations, by vesting providers of VLOPs and VLOSEs with the duty, but at the same time with the discretion, to choose, adopt, and implement the necessary mitigating measures, reflect a regulatory strategy that has been defined as "meta-regulation" or "enforced self-regulation":

> "Meta" because one (macro) regulator oversees another (micro) regulator in the management of risk; "enforced" because, in case of inadequacy of the self-regulatory practices, the (macro) regulator has the power to take enforcement measures. To determine whether such measures are warranted, meta-regulation establishes norms of organisation and procedure through which self-regulatory practices can be assessed. By doing so, it assumes a fundamentally "reflexive" character: it focuses on enhancing the self-referential capacities of social systems and institutions outside the legal system to achieve broad social goals, rather than on prescribing particular actions.[278]

The delegation to providers of VLOPs and VLOSEs of such tasks clearly represents a rather delicate choice, the risks of which – namely, the push towards (automated) over-removal of content – were highlighted soon after the presentation of the DSA proposal by the Commission.[279] Accordingly, the final text of the Regulation was complemented with the addition of a new clause with the provision that, when deploying their mitigation strategies, providers must have "particular consideration to the impacts of [the] measures [adopted] on fundamental rights".[280] As a result, providers of VLOPs and VLOSEs are discouraged from resorting to excessively restrictive strategies as a means to comply with the Regulation, as they are, in fact, required to operate a careful balancing of all interests at stake, including those of the recipients of the service, in the light of the principle of proportionality.

To support the correct enforcement of these provisions, the EBDS, in cooperation with the Commission, is vested with the task of publishing annually comprehensive reports

---

[276] ibid 34, para 2.

[277] ibid 35, para 1.

[278] Nicolo Zingales, 'The DSA as a Paradigm Shift for Online Intermediaries' Due Diligence: Hail To Meta-Regulation' in Joris van Hoboken and others (eds), *Putting the DSA into Practice: Enforcement, Access to Justice, and Global Implications* (Verfassungsbooks 2023) 214.

[279] Joan Barata, 'The Digital Services Act and Its Impact on the Right to Freedom of Expression: Special Focus on Risk Mitigation Obligations' (*DSA Observatory*, 27 July 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/> accessed 3 December 2021.

[280] DSA art 35, para 1.

identifying and assessing the most prominent and recurrent systemic risks and sharing best practices, while the Commission, in cooperation with national DSCs, is entitled to issue guidelines in relation to specific risks.[281] Additionally, in times of crisis, that is, "when extraordinary circumstances occur that can lead to a serious threat to public security or public health in the Union or significant parts thereof",[282] the Commission may require one or more providers of VLOPs or VLOSEs to take the necessary additional actions.[283]

On top of the obligations related to risk management, providers of VLOPs and VLOSEs must also comply with additional transparency requirements. First, they must make available to the public information concerning the provision of online advertising.[284] Second, the DSA sets the procedural conditions upon which DSCs may request providers either to grant them access to data that are necessary to monitor and assess compliance with the Regulation or to provide "vetted researchers" with data "for the sole purpose of conducting research that contributes to the detection, identification and understanding of systemic risks in the Union".[285] Third, specific rules apply with respect to the release of transparency reports which, in the case of VLOPs and VLOSEs, must be published every six months and must include additional information, including in particular the number, qualifications, and linguistic expertise of human resources dedicated to content moderation and indicators of accuracy, and related information referring to the use of automated means for such purposes.[286]

Finally, to ensure the full compliance with the obligations set within the DSA, providers of VLOPs and VLOSEs must, on the one hand, establish a dedicated compliance function within their organization, independent from their operational functions and composed of one or more compliance officers,[287] and, on the other hand, undergo independent audits at least once a year.[288]

3.5.3.5. Standards, codes of conduct, and crisis protocols

The DSA complements the rules on providers' new due diligence obligations by recognizing the possibility to develop co-regulatory strategies for complying with such obligations, namely through the development and implementation of voluntary standards set by

[281] ibid 35, paras 2-3.

[282] ibid rec 91, which adds that crises may be the result of armed conflicts, acts of terrorism, natural disasters, pandemics, and other serious cross-border threats to public health.

[283] ibid 36. Besides, the DSA assigns the Commission a central role in the supervision, investigation, enforcement and monitoring activities with respect to the compliance by VLOPs and VLOSEs with their due diligence obligations, as emerges from arts 64–83. Also for this reason, art 43 introduces the additional obligation for VLOPs and VLOSEs to pay supervisory fees to cover the costs incurred by the Commission.

[284] ibid 39.

[285] ibid 40.

[286] ibid 42.

[287] ibid 41.

[288] ibid 37.

European and international standardization bodies[289] and, especially, through the adoption of voluntary codes of conduct at Union level.[290]

In this sense, the DSA requires that where several VLOPs and several VLOSEs are all found to be affected by a certain significant systemic risk, the Commission may invite them, together with any other provider of intermediary services, as appropriate, as well as with relevant competent authorities, civil society organizations and other relevant stakeholders, to participate in the drawing up of codes of conduct to face such systemic risk: these codes shall set out clearly their specific objectives, contain key performance indicators allowing for the measurement of the codes' results, and take into account the interests of any parties affected (including those of citizens and of the Union).[291] The Commission and the EBDS are tasked with the monitoring and evaluations of the codes' objectives and should encourage and facilitate regular review and adaptation of the codes. Additionally, in case of systematic failures to comply with the codes, they may invite signatories to take the necessary actions.[292] The DSA also envisages the possibility to adopt specific codes of conduct for online advertising and for the promotion of accessibility to encourage the full and effective equal participation of persons with disabilities, as well as crisis protocols to address situations of crisis.[293]

The main advantage that the adoption of codes of conduct entails is represented by the possibility of introducing more specific commitments tailored specifically to address certain systemic risks. Indeed, whereas the rules set by the DSA, especially those concerning the obligations to assess and mitigate risk, are of a general nature, and thus suffer from being quite vague and abstract, codes of conduct allow for the introduction of more detailed norms to hold providers of VLOPs and VLOSEs (and, where applicable, also providers of other intermediary services) accountable.[294] At the same time, such additional prescriptions have the advantage, for providers of VLOPs and VLOSEs, of not being identified and defined on a top-down basis but, rather, as a result of negotiations that take into account their own economic and financial interests.

Besides, although the DSA highlights their voluntary nature at the moment of their genesis, such codes of conduct, once they have been agreed upon, tend to acquire under the text of the Regulation a rather compelling force which does not affect only those actors that have participated in their drafting but, rather, all VLOPs and VLOSEs.[295] This emerges, most notably, from the wording of Recital 104, according to which

[289] ibid 44.
[290] ibid 45.
[291] ibid 45, paras 2-3.
[292] ibid 45, para 4.
[293] ibid 46–48.
[294] Rachel Griffin and Carl Vander Maelen, 'Codes of Conduct in the Digital Services Act: Exploring the Opportunities and Challenges' (SSRN, 30 May 2023) 7 <https://papers.ssrn.com/abstract=4463874> accessed 14 June 2023.
[295] As argued by Griffin and Vander Maelen, while the DSA "repeatedly reiterates the 'voluntary' nature of codes", "given their role in risk mitigation and auditing, VLOP/VLOSE participation is all but inescapable as part of DSA compliance". ibid 6.

> adherence to and compliance with a given code of conduct by a very large online platform or a very large online search engine may be considered as an appropriate risk mitigating measure. The refusal without proper explanations by a provider of an online platform or of an online search engine of the Commission's invitation to participate in the application of such a code of conduct could be taken into account, where relevant, when determining whether the online platform or the online search engine has infringed the obligations laid down by this Regulation. The mere fact of participating in and implementing a given code of conduct should not in itself presume compliance with this Regulation.[296]

Therefore, the choice to adhere to and comply with a code of conduct may strengthen the position of a VLOP or VLOSE in proving their compliance with the DSA's due diligence obligations, even though such a choice is not *per se* sufficient to represent an inescapable presumption. Conversely, the refusal to apply a code of conduct, especially upon invitation from the Commission, explicitly represents a strong piece of evidence against VLOPs and VLOSEs. In other words, compliance with a code of conduct ends up representing an almost necessary baseline for the provider to prove its compliance with the DSA. Thus, for instance, X's recent choice to withdraw from the 2022 Strengthened Code of Practice on Disinformation[297] may well expose it to the serious risk of being held responsible for failing to comply with the DSA.[298]

According to Vander Maelen, the DSA's new framework on codes of conduct has important impacts both on their functioning and on their legal nature. On the one hand, indeed, the new discipline affects the three main functional dimensions of codes of conduct, consisting of implementation, accountability, and enforcement. First, the DSA explicitly and clearly recognizes that the codes of conduct have the inherent role of aiding the "further implementation of the hard law instrument itself";[299] second, and as a consequence of the implementation dimension, the codes of conduct under the DSA have the role of contributing to assessing the compliance of providers with their due diligence obligations and thus to the assessment of their accountability in case of failure to do so;[300] third, the DSA envisages a rather strong enforcement system where the Commission and the EBDS play a central role.

On the other hand, as already highlighted above, the DSA framework tends to lead to a process of "hardening" or "juridification" of the codes of conduct themselves, namely

---

[296] DSA rec 104.

[297] Strengthened Code of Practice on Disinformation 2022. On the relation between the Strengthened Code of Practice and the DSA see, among others, Mark R Leiser, 'Analysing the European Union's Digital Services Act Provisions for the Curtailment of Fake News, Disinformation, & Online Manipulation' (SSRN, 24 April 2023) 8–9 <https://papers.ssrn.com/abstract=4427493> accessed 14 June 2023.

[298] Natasha Lomas, 'Elon Musk Takes Twitter out of the EU's Disinformation Code of Practice' (*TechCrunch*, 27 May 2023) <https://techcrunch.com/2023/05/27/elon-musk-twitter-eu-disinformation-code/> accessed 14 June 2023; Carl Vander Maelen and Rachel Griffin, 'Twitter's Retreat from the Code of Practice on Disinformation Raises a Crucial Question: Are DSA Codes of Conduct Really Voluntary?' (*DSA Observatory*, 12 June 2023) <https://dsa-observatory.eu/2023/06/12/twitters-retreat-from-the-code-of-practice-on-disinformation-raises-a-crucial-question-are-dsa-codes-of-conduct-really-voluntary/> accessed 14 June 2023.

[299] Vander Maelen (n 167) 764.

[300] In this respect, Vander Maelen argues that "the DSA's approach is both positive and negative. Positively, it states that adherence to and compliance with codes 'may be considered as an appropriate risk mitigating measure' when dealing with illegal content and systemic risks"; negatively, Vander Maelen refers to the mentioned text of rec 104. See ibid 765.

by affecting three dimensions of "legalization": that is, the dimension of obligation, that of precision, and that of delegation. As regards the dimension of obligation, the DSA causes a shift from a mechanism whereby corporations choose to adhere to and comply with the codes of conduct on a voluntary basis to a system where, conversely, they "are faced with a strong de facto obligation to participate in codes" and are "involved in codes because codes are directly linked to punitive hard law provisions".[301] As for the dimension of precision, the DSA also represents a notable change of direction: indeed, while codes of conduct "are traditionally carriers of 'open' norms', i.e. imprecise broad goals that offer corporations discretion in how to implement them", the DSA "posit[s] broad hard law provisions and determine[s] that codes are meant to specify those provisions by offering prescriptive and specific solutions".[302] Finally, and as a result of the characteristics of the DSA as to the functional dimension of enforcement, the codes of conduct under the new Regulation shall not rely as much as their traditional counterparts on non-judicial monitoring bodies offering advice or making non-binding decisions: rather, the Commission and the EBDS are vested with the duties to monitor the content of codes and ensure they are complied with.[303]

3.5.4. *DSA and hate speech moderation*

3.5.4.1. Applicability of the DSA to hate speech moderation

Clearly, the new regulatory framework introduced by the DSA is highly significant with regard to the governance of online hate speech in the context of the EU, as many of the provisions described above shall be applicable also to hate speech moderation activities.

Notably, the rules concerning the adoption and implementation of the provider's terms and conditions, the provision of a statement of reasons for the adoption, by a provider of hosting services, of a content moderation or content curation measure against the recipient, and the establishment of a complaint-handling system will all affect the way providers address and deal with the issue of hate speech. Therefore, providers should, for instance, state clearly in their terms and conditions how hate speech is treated and sanctioned and should ensure that the enforcement of those terms and conditions does not violate the fundamental rights of recipients as enshrined within the CFREU. In other words, any obligation setting procedural and substantive limits protecting recipients of the service against the private power of providers should be applicable also with respect to the moderation of hate speech. Nevertheless, the same cannot be said about other provisions of the Regulation.

Indeed, while setting a new framework for the countering of "illegal content", the DSA does not intervene to define what such a category includes. In fact, the Regulation generally refers to other sources of EU and domestic law for the definition of what is illegal:

[301] ibid 767.
[302] ibid.
[303] ibid 768.

> 'Illegal content' means any information that, in itself or in relation to an activity, including the sale of products or the provision of services, is not in compliance with Union law or the law of any Member State which is in compliance with Union law, irrespective of the precise subject matter or nature of that law.[304]

Admittedly, the choice to adopt such a generic definition, rather than a closed list, is coherent with the systematization of the DSA as a new horizontal framework for content moderation, applicable broadly and as a general standard. However, this is not without consequences. Namely, the reference to domestic law in the identification of what is to be considered as "illegal" can open the road to a balkanization across the EU as to the applicability of the rules on "illegal content" whenever Member States' laws differ.

This is, precisely, the case of hate speech. Indeed, as discussed above,[305] at the EU level, Framework Decision 2008/913/JHA attempted to harmonize the criminal response against hate speech across all Member States' jurisdictions. However, the Framework Decision only considers the protected grounds of "colour, religion, descent or national or ethnic origin".[306] Conversely, hate speech on other grounds is excluded, so that national legislations vary sensitively in this respect, especially when it comes to the extension of hate speech bans to forms of sexist, ageist, anti-LGBTQIA+, or ableist speech.[307]

As a result, all DSA provisions establishing specific obligations to put in place moderation activities against illegal speech, including notably the rules governing providers' duty to comply with judicial or administrative authorities' orders to act against illegal content and the necessary establishment of notice and action mechanisms, will have different effects depending on the grounds of discrimination concerned. That is, hate speech on grounds of "colour, religion, descent or national or ethnic origin" should in all cases be affected; whereas in other cases, such as is the case of sexist or anti-LGBTQIA+ speech, the applicability of those rules will vary depending on the applicable Member State law.

With respect to VLOPs and VLOSEs' duties of assessment and mitigation of systemic risks, the extent of the applicability of such obligations to hate speech is, moreover, not fully clear. Indeed, the question arises whether VLOPs and VLOSEs should consider as relevant for this purpose the law of the place of establishment or whether, given the extent of their reach, they should break down the assessment and mitigation of risks for each Member State, thus considering the different laws applicable across the European Union. Admittedly, such an interpretive issue could nevertheless be overcome by considering all forms of hate speech, even if not "illegal", as potentially able to lead to actual or foreseeable negative effects "for the exercise of fundamental rights", "on civic discourse and

[304] DSA art 3, lett (h).

[305] See *supra*, §2.2.3.2.

[306] Framework Decision 2008/913/JHA art 1, para 1, lett (a).

[307] In Italy, for instance, as recently as in 2021, the so-called "Zan Draft Law", extending the ban also to hate speech based on sex, gender, gender identity, sexual orientation, and disability, was quashed, so that the current legislation is still rather minimal. Lorenzo Tondo, '"Disgraceful": Italy's Senate Votes down Anti-Homophobic Violence Bill' (*The Guardian*, 27 October 2021) <https://www.theguardian.com/world/2021/oct/27/italy-senate-votes-down-anti-homophobic-violence-bill> accessed 14 June 2023. See more *infra*, §4.2.2.2.

electoral processes, and public security", or "in relation to gender-based violence, the protection of public health and minors and serious negative consequences to the person's physical and mental well-being".

Be that as it may, the system established by the DSA with respect to hate speech has the evident effect of creating two different regimes due to the possibility of qualifying a certain utterance, based on the grounds of discrimination involved, as "illegal hate speech" under national or EU law or as "non-illegal" hate speech. Such a difference of treatment may affect the uniform application of the DSA across the European Union. Additionally, the resulting hierarchization between different grounds of discrimination in the countering of online hate speech is, at the very least, rather objectionable under the lens of the principle of non-discrimination as enshrined in Article 21 CFREU.

In light of this, the Commission's proposal to amend Article 83, paragraph 1, to include hate crimes and hate speech within the group of so-called "EU crimes", by opening up the road to a further harmonization of Member States' criminal legislation in this field, could lead to the indirect effect of clarifying and broadening the scope of application of the DSA to hate speech based on any grounds of discrimination.[308]

3.5.4.2. Hate speech moderation and equality in the DSA

Notwithstanding the complex challenges connected to the precise scope of applicability of the DSA in relation to hate speech, it is nonetheless evident that the new Regulation has the potential to represent an unprecedented step forward in orienting the content moderation and content curation practices of providers of intermediary services, and especially of VLOPs and VLOSEs, when it comes to the governance of such a phenomenon.

Indeed, it should be stressed how, even in the absence of a legal obligation to detect it and remove it, many providers of intermediary services have introduced within their terms and conditions broad prohibitions on hate speech, often encompassing extensive, sometimes explicitly non-exhaustive, lists of grounds of discrimination considered.[309] As a result, the introduction within the DSA of rules setting substantive and, especially, procedural limitations to the exercise of "platform law"[310] could have important results especially in pushing ISPs to adopt strategies of contrast against hate speech internalizing important values such as, potentially, the promotion of substantive equality in the participation in the public debate online.

As a matter of fact, many provisions of the DSA arguably showcase the EU lawmaker's awareness of the threat that an unrestrained private content moderation and curation can pose to individual fundamental rights, including of course freedom of expression and information *vis-à-vis* the possibility of over-removal and/or forms of shadow banning, but also the right to non-discrimination. For instance, as mentioned above, the

[308] See *supra*, §2.2.3.2.

[309] Paolo Cavaliere, 'Digital Platforms and the Rise of Global Regulation of Hate Speech' (2019) 8 Cambridge International Law Journal 282, 290–291; Wilson and Land (n 79) 1047–1053.

[310] David Kaye, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (Human Rights Council 2018) A/HRC/38/35 para 1.

Regulation insists that providers' terms and conditions are enforced in a non-arbitrary and non-discriminatory manner.

The Regulation's focus on the goal of ensuring the protection of recipients against forms of discrimination in the context of content moderation and content curation, however, also emerges starkly from the obligations concerning the establishment of a risk assessment and mitigation system for VLOPs and VLOSEs, with respect to which Article 34 mentions explicitly the need to guarantee the full respect of Article 21 CFREU. Pursuant to Article 35, moreover, risk mitigation measures themselves must be adopted with "particular consideration" to the impacts they might have on "fundamental rights", and a systematic interpretation with the previous provision would suggest including among these fundamental rights also the right to non-discrimination. Additionally, the Regulation complements such rules with obligations concerning the resort to automated means of content moderation and aimed at ensuring correctives against the risk for biased outcomes.

Application of the DSA will likely face the significant challenge of ensuring the full operationalization of such guarantees for the principle of equality, especially with regard to the field of hate speech moderation. Adopting a substantive equality approach towards the right to non-discrimination and, therefore, towards the governance of hate speech, such as that proposed in the previous Chapter,[311] would in this respect contribute significantly to shaping and orienting the moderation strategies of ISPs, with a view to promoting the equal enjoyment of digital rights of minority, discriminated, and marginalized groups.

In this respect, given the general and non-specific scope of the DSA, it may arguably be essential, if not to intervene with *ad hoc* sectoral hard legislation, at least to operate a revision of the 2016 CoC on Illegal Hate Speech Online with a view to updating it and rendering it an effective complementary tool for the promotion of a positive framework towards hate speech governance, invested not only in the simple removal of unwarranted content but also striving to promote counter-speech and, in general, to actively support equal participation of all demographics throughout the Internet.[312] Besides, following the seventh evaluation of the CoC, the European Commission declared in a press release the possibility that a revision of the Code might indeed take place, with a view to addressing the unsatisfactory progresses of signatory IT companies, as well as to adapting it to the new framework established by the DSA.[313]

On such an occasion, it would be desirable that the revision of the CoC were guided by the aim of guaranteeing that moderation and curation measures do not have the controversial effect of interfering with the liberties of groups victimized by hate speech but, rather, that of further empowering them and their role in society.

---

[311] See *supra*, §2.5.2.

[312] See more *infra*, §5.5.2.

[313] European Commission, 'EU Code of Conduct against Online Hate Speech: Latest Evaluation Shows Slowdown in Progress' (*European Commission*, 24 November 2022) <https://ec.europa.eu/commission/presscorner/detail/en/ip_22_7109> accessed 15 June 2023.

3.6. Conclusions

The Chapter has addressed the complex framework on intermediary liability for third-party illegal content within the ECHR and EU frameworks, both of which have undergone important developments from the turn of the millennium and, especially, from the mid-2010s onwards.

On the one hand, the ECtHR has inaugurated with the landmark decision of *Delfi AS v Estonia* a particularly relevant strand of case law, confirming the consistency of the imposition of sanctions for the failure to promptly remove third-party illegal content with the content of Article 10 ECHR on freedom of expression and information. Although subsequent case law, including namely *MTE v Hungary*, has in general mitigated the conclusions adopted in *Delfi*, showcasing a rather more lenient approach towards ISPs, the ECtHR has maintained nevertheless a highly restrictive approach towards the specific case of hate speech. Hate speech is considered by the Strasbourg Court to be of such an egregious nature as to justify the choice of rather limitative measures of intermediaries' Article 10 prerogatives: an approach arguably confirmed in 2023 by the Grand Chamber in *Sanchez v France*.

On the other hand, the EU framework on intermediary liability has undergone a particularly significant evolution in the last few years. Indeed, while in 2000, following the model of the US, the EU had adopted an approach rather favourable towards ISPs, the last two decades have seen the development of a much different strategy, first through the (manipulative) intervention of the CJEU and, subsequently, with the rise of a new legislative season, increasingly oriented towards enhancing the liability, responsibility, and accountability of ISPs for their content moderation and content curation practices. The DSA, in particular, represents a rather revolutionary horizontal reform in this field, having introduced an extensive set of new due diligence obligations.

The practical effects and impact of the DSA are yet to be assessed, as the Regulation, although already applicable to providers of VLOPs and VLOSEs as of 25 August 2023, only became applicable to all other providers of intermediary services on 17 February 2024.[314] Additionally, the first Commission reports on the evaluation of the application and impact of the Regulation are only expected to be presented between 2025 and 2027. Most notably, the extent to which the DSA will be applicable with respect to the moderation of hate speech still raises some doubts, especially in the light of the current lack of harmonization in this field across Member States. Precisely for these reasons, the adoption of complementary strategies, including the revision of the CoC on Illegal Hate Speech Online and the revision of Article 83, paragraph 1, TFEU, are rather desirable at this point.

The next Chapters will complement such reflections addressing two important challenges for the future of the European strategy against online hate speech. Chapter 4 will consider the matter of intermediary liability for third-party hate speech from a

[314] DSA art 93, para 2.

comparative perspective, looking at how, within the EU and within extra-EU and extra-European jurisdictions, such matters are dealt with, namely with a view to inquiring the relationship of the DSA with the multitude of these legal systems.

Chapter 5, conversely, will focus specifically on the issue of private platforms' standards on hate speech moderation and on the functioning and effects of the use of automated content moderation systems. Chapter 5 will thus address the challenges AI raises for hate speech moderation, namely in the light of the principle of substantive equality, and the role the DSA and complementary sources could – and should – play in confronting those challenges. In this respect, ensuring that the DSA, together with any future complementary instrument, has the material effect of promoting discriminated and marginalized demographics' enjoyment of freedom of expression in the digital landscape – and thus of ensuring their full participation in the democratic discourse online – represents a goal of paramount importance.

With respect to the last aspect, an approach oriented towards a substantive equality perspective on hate speech governance, such as that suggested in Chapter 2, would likely be beneficial and offer important insights into how to confront upcoming challenges. Indeed, because the DSA – being a piece of legislation aimed precisely at enhancing the responsibility and accountability of ISPs – carries the inherent risk of pushing providers towards the over-removal of content uploaded by users, it is essential to define strategies capable of avoiding the result that such a risk outruns the advantages brought by the new Regulation. Chapter 5 will thus argue that, in order to maximize the beneficial effects of the DSA while minimizing its collateral effects in the context of the fight against hate speech, the principle of substantive equality may well represent an optimal objective and proxy for future policymaking.

4.

Hate Speech and Intermediary Liability: A Comparative Overview

4.1. Introduction

After having outlined the evolution and main characteristics of the European approach towards online illegal content and, specifically, illegal hate speech, the present Chapter aims to give a broad overview, from a comparative perspective, of how the challenges raised by such content across the Internet have – or have not – been addressed by other jurisdictions. Notably, such a comparative overview will be focused on approaches both

internal and external to the EU, with a view to highlighting common patterns and/or major differences. Indeed, the transnational nature of the Internet and of online hate speech[1] requires an understanding of the way such phenomena are treated elsewhere, in order to investigate what role EU law, and the DSA in particular, may play in shaping the future of moderation policies worldwide.

Section 4.2 considers the relationship between, on the one hand, the DSA and EU strategies against hate speech and, on the other hand, the frameworks of some notable Member States. Among these, specific consideration is given to German law (§4.2.1), as its Network Enforcement Act, enacted in 2017, has served as an internationally relevant blueprint for the governance of intermediary liability for user-generated hate speech (and disinformation). Subsequently, the experiences of France, Italy, and Spain are described, with a view to showcasing alternative – and more or less successful – paths taken in recent years (§4.2.2). Finally, the Section critically discusses the dynamics between the DSA and the speech governance approaches of some Eastern European countries affected by forms of democratic backsliding, namely Poland and Hungary (§4.2.3).

Section 4.3 describes the UK Online Safety Act, with a view to outlining its material, subjective, and territorial scope of application (§4.3.1), the new set of duties imposed upon providers of Internet services (§4.3.2), and the role that the Act will play in the fight against online hate speech in the UK (§4.3.3). The Online Safety Act, indeed, represents a particularly relevant term of comparison, as it aims to reach goals that are in part coinciding with those of the DSA: in this respect, the common aspects and differences between the two instruments will be at the centre of the Section's attention.

Section 4.4 addresses, in turn, the legal framework of the US. As has been noted and stressed more than often by a plurality of commentators, the US takes indeed a radically different approach towards intermediary liability – especially when it comes to hate speech moderation – from that of the EU. This Section, after having briefly underlined the typically tolerant approach of the US towards the "thought we hate" (§4.4.1),[2] thus addresses the rise, at the end of the 1990s, of the notorious Section 230 of the Communications Decency Act, outlining the fundamental role played by the provision in the development of the US case law on intermediary liability (§4.4.2). Section 4.4 also discusses the interplay between Section 230, the state action doctrine (§4.4.3), and the First Amendment, discussing the critiques moved both by conservatives and liberals towards such a system and analysing the attempts made on both sides to amend it (§4.4.4). The success or failure of such attempts may well support or hamper a positive relationship between the DSA and US law, as discussed in subsection 4.4.5.

Section 4.5 aims to give a brief overview of some legislative approaches worldwide. Indeed, although the Section does not aim to give a full and extensive account of how online hate speech governance has been addressed across the various continents, it is nevertheless deemed particularly relevant to highlight the plurality of approaches that can be taken and have been taken with respect to the phenomenon, so as to bear in mind that the

[1] See *supra*, §2.4.2.4.
[2] See also *supra*, §2.3.1.

regulatory strategies of Western democracies do not take place in a void but have to confront themselves with many different jurisdictions. Some selected examples are thus addressed with regard to Asia (§4.5.1), Africa (§4.5.2), and Latin America (§4.5.3); additionally, the case of Australia is also analysed (§4.5.4).

The main conclusions concerning such a comparative inquiry and the relationship of the DSA to the international scenario are discussed, finally, in Section 4.6.

4.2. Domestic legislation of EU Member States

4.2.1. *Germany and the NetzDG: a controversial model?*

4.2.1.1. Content of the NetzDG

In June 2017, the German Bundestag passed an innovative federal law, the Act to Improve Enforcement of the Law in Social Networks (*Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken*), commonly known as "Network Enforcement Act" (*Netzwerkdurchsetzungsgesetz*, NetzDG),[3] which operated a reform of the intermediary liability for third-party illegal content with the declared goal to confront, most notably, the rise of phenomena such as online hate speech and disinformation.[4] Fully operational since January 2018, the NetzDG originally dealt only with "social networks", defined as for-profit tele-media service providers operating Internet platforms designed specifically to enable users to share content with other users or to make such content available to the public.[5] Following the passing of Directive (EU) 2018/1808,[6] however, the text was amended[7] to extend the law's scope of action so as to include video-sharing platforms as well.

The NetzDG is focused on countering the dissemination of unlawful content the publication of which constitutes a criminal offence. In this respect, it is worth mentioning that, in order to define what is to be considered as unlawful content under its scope of intervention, the German Act explicitly refers to a set of provisions contained within the

[3] *Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken* (*Netzwerkdurchsetzungsgesetz* - NetzDG) 2017 (BGBI I S 3352).

[4] For an overview of the history of NetzDG, and notably of the reasons that brought to its adoption, see Thomas Wischmeyer, 'What Is Illegal Offline Is Also Illegal Online: The German Network Enforcement Act 2017' in Bilyana Petkova and Tuomas Ojanen (eds), *Fundamental Rights Protection Online: The Future Regulation of Intermediaries* (Edward Elgar Publishing 2020); Nannerel Fiano, 'Il Linguaggio Dell'Odio in Germania: Tra *Wehrhafte Demokratie* e *Netzwerkdurchsetzungsgesetz*' in Marilisa D'Amico and Cecilia Siccardi (eds), *La Costituzione non odia: Conoscere, prevenire e contrastare l'*hate speech on line (Giappichelli 2021).

[5] NetzDG s 1, para 1. Conversely, platforms offering journalistic or editorial content, the responsibility for which lies with the service provider itself, fall out of the scope of the law, as well as platforms which are designed to enable individual communication or the dissemination of specific content.

[6] See *supra*, §3.4.3.2.

[7] *Gesetz zur Änderung des Netzwerkdurchsetzungsgesetzes* 2021 (BGBI I S 1436). The NetzDG was subsequently amended once again in July 2022, with a view to implementing the TERREG (see *supra*, §3.4.3.2), through *Gesetz zur Durchführung der Verordnung (EU) 2021/784 des Europäischen Parlaments und des Rates vom 29. April 2021 zur Bekämpfung der Verbreitung terroristischer Online-Inhalte und zur Änderung weiterer Gesetze* 2022 (BGBI I S 1182).

German Criminal Code.[8] Most notably, the provisions considered by the NetzDG include, *inter alia*, Section 130 of the Criminal Code, criminalizing the conduct of inciting hatred or of calling for violent or arbitrary measures against individuals or groups based on their nationality, "race", religion or ethnicity, as well as the act of violating the human dignity of such persons or groups by insulting, maliciously maligning or defaming them, provided that such conducts are put in place "in a manner suited to causing a disturbance of the public peace".[9]

Providers of social networks and video-sharing platforms with at least 2 million registered users in Germany are required to comply with several obligations, on penalty of being subjected to hefty fines up to five million euros.[10] First of all, they need to establish a procedure to handle complaints concerning the presence of illegal content upon their infrastructures. Such a procedure must be easily recognizable, directly accessible and permanently available, and must ensure that the provider addresses and responds rapidly to any complaint. Namely, content that is "manifestly unlawful" should be removed or have access to it blocked within 24 hours, whereas content whose unlawfulness is not manifest should be acted upon within 7 days – which may only be exceeded in limited cases.[11] Additionally, the decision concerning the action to be taken against the unlawful content must be notified immediately both to the complainant and to the user subjected to the decision and must be justified and the notification must indicate the possibilities for redress, and inform the complainant of the possibility of filing a criminal suit against the alleged poster of unlawful content.[12] Providers, besides, may decide to refer the decision concerning the unlawfulness of a certain piece of information to a recognized self-regulatory institution, thus agreeing to accept the latter's decision, within 7 days of receiving the complaint. It is up to the administrative authority to recognize such self-regulation institutions, based, namely, on criteria of independence and expertise.[13]

On top of the obligation to establish the described procedure, providers are also required to report to the authorities, including the Federal Criminal Police Office, information concerning the placing of illegal content or the commission of illegal activities

[8] Namely, *Strafgesetzbuch* (StGB) in the version published on 13 November 1998 (BGBI I S 3322) ss 86–86a, 89a, 91, 100a, 111, 126, 129–129b, 130, 131, 140, 166, 184b, 185–187, 189, 201a, 241, 269. See NetzDG s 1, para 3.

[9] "Wer in einer Weise, die geeignet ist, den öffentlichen Frieden zu stören, gegen eine nationale, rassische, religiöse oder durch ihre ethnische Herkunft bestimmte Gruppe, gegen Teile der Bevölkerung oder gegen einen Einzelnen wegen dessen Zugehörigkeit zu einer vorbezeichneten Gruppe oder zu einem Teil der Bevölkerung zum Hass aufstachelt, zu Gewalt- oder Willkürmaßnahmen auffordert oder … wird mit Freiheitsstrafe von drei Monaten bis zu fünf Jahren bestraft". StGB s 130, para 1.

[10] NetzDG s 4, para 2.

[11] ibid 3, para 2, n 1–3. Namely, the period of 7 days for the assessment of cases of non-manifestly unlawful content may be exceeded if "the decision regarding the unlawfulness of the content is dependent on the falsity of a factual allegation or is clearly dependent on other factual circumstances", in which case "the social network can give the user an opportunity to respond to the complaint before the decision is rendered"; or if "the social network refers the decision regarding unlawfulness to a recognized self-regulation institution … within 7 days of receiving the complaint and agrees to accept the decision of that institution".

[12] ibid 3, para 2, n 5.

[13] ibid 3, paras 2, n 3(b), 6–10.

through their infrastructures,[14] and, in case they receive more than 100 complaints per calendar year, publish six-monthly reports – in German – concerning the handling of complaints. This last reporting obligation, clearly, is aimed at further guaranteeing providers' transparency about their content moderation practices. Coherently, subsequent amendments to the NetzDG intervened precisely to extend the range of information to be provided within these six-monthly reports: for instance, the current text of the Act requires providers to give information also regarding the use of automated detection as well as information about the data used.

As a matter of fact, the current text of the NetzDG, if compared with its 2017 original, contains a significant amount of provisions aimed specifically at fostering transparency and promoting the introduction of additional safeguards and guarantees for the individual recipients of the intermediary services concerned: for instance, the NetzDG now foresees the possibility for users to avail themselves of procedures for counter-complaints, allowing both the complainant and the target of a complaint to ask for a revision of any decision on the complaint.[15] Additionally, the law currently envisages the possibility for researchers to request qualified information concerning, *inter alia*, the systems and technologies employed for content moderation.[16]

4.2.1.2. Controversial aspects: NetzDG and freedom of expression

Considered from the beginning to be "arguably the most ambitious attempt by a Western State to hold social media platforms responsible for combating online speech deemed illegal under the domestic law",[17] the NetzDG attracted nevertheless much criticism from a variety of parties, including the tech industry, as well as from activists and academics. Major criticisms concerned, on the one hand, the impact of such a legislation upon the fundamental right to freedom of expression and, on the other hand, its relationship with EU law, namely with the ECD. As regards the first point, two main arguments have been brought against the NetzDG: that of the increased possibility for over-removal and that of the privatization of speech censorship.

First, concerns were raised with respect to the risk of legal content being subjected to over-removal.[18] These concerns were also shared by the then UN Special Rapporteur on

[14] ibid 3a.
[15] ibid 3b.
[16] ibid 5a.
[17] Heidi Tworek and Paddy Leerssen, 'An Analysis of Germany's NetzDG Law' (TWG 2019) 1 <https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf> accessed 12 July 2023.
[18] Article 19, 'Germany: Responding to "Hate Speech"' (2018) 19 <https://www.article19.org/resources/germany-responding-to-hate-speech/> accessed 12 July 2023; Human Rights Watch, 'Germany: Flawed Social Media Law' (*Human Rights Watch*, 14 February 2018) <https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law> accessed 12 July 2023. Indeed, as pointed out by Schulz, being tasked with assessing whether a certain content is illegal within the short time limits established by the NetzDG "puts pressure on a provider and might … push said provider towards the simple but human-rights-adverse solution of taking down the content in almost any case", a risk which is often further amplified by the fact that providers often "lack information about the context, as well as the necessary information-gathering tools, to make a proper assessment": see Wolfgang Schulz,

freedom of expression and opinion, according to whom the German law risked infringing such a freedom as protected by Article 19 ICCPR.[19] According to Hong, the NetzDG does not feature satisfactory provisions protecting speech from over-removal, thus violating the paramount presumption in favour of free speech which should guide any legislation regulating the constitutional freedom of expression, nor is the new procedure for counter-complaints, in force since 28 June 2021, "sufficient to correct this defect, since there is still no explicit duty to put back in content that was erroneously blocked or removed".[20]

Linked to this first argument was also the observation that the adoption of any form of speech regulation tends to lead to chilling effects, that is, to the result of discouraging individuals from expressing their viewpoints and opinions due to fear of being punished. Nevertheless, it has been rightly noted that such an observation is fraught as it does not take into account the fact that the NetzDG does not, in fact, extend the category of conducts punishable under criminal law. It does not, in other words, reduce the extent to which individuals may exercise their freedom of expression prerogatives without incurring criminal sanctions, but simply renders ISPs accountable for the commission of conducts already recognized by the law as offences. Therefore,

> if this effect existed, it would be based on irrational and erratic user behavior because [users] face liability regardless of the NetzDG, which does not change the assessment of the criminal or other liability for statements. The risk of criminal prosecution is much more serious than that of blocking a post. It would, therefore, be incomprehensible if users were to refrain from posting content just because of the NetzDG. At best, it is conceivable that users might fear that their accounts will be blocked or suspended. Yet, this measure is not provided for by the NetzDG but can only be imposed on the basis of self-imposed standards.[21]

Additionally, it has been pointed out that the risk of over-removal is significantly diminished by the fact that the NetzDG does not entail the imposition of fines whenever a social network or video-sharing platform provider fails to remove illegal content but, rather, when such a failure to comply with the law is systemic or persistent. Indeed, data released in the years following the enactment of the new legislation have shown that the NetzDG

'Regulating Intermediaries to Protect Personality Rights Online - The Case of the German NetzDG' in Marion Albers and Ingo Wolfgang Sarlet (eds), *Personality and Data Protection Rights on the Internet: Brazilian and German Approaches* (Springer 2022) 298–299.

[19] "The list of violations is broad, and includes violations that do not demand the same level of protection. Moreover, many of the violations covered by the bill are highly dependent on context, context which platforms are in no position to assess … The short deadlines, coupled with the … severe penalties, could lead social networks to over-regulate expression – in particular, to delete legitimate expression, not susceptible to restriction under human rights law, as a precaution to avoid penalties. Such pre-cautionary censorship, would interfere with the right to seek, receive and impart information of all kinds on the internet". David Kaye, 'Comment on the Social Networks Bill (Netzdurchführungsgesetz)' (Office of the High Commissioner for Human Rights 2017) OL DEU 1/2017 4 <https://www.ohchr.org/en/special-procedures/sr-freedom-of-opinion-and-expression/comments-legislation-and-policy> accessed 12 July 2023.

[20] Mathias Hong, 'Regulating Hate Speech and Disinformation Online While Protecting Freedom of Speech as an Equal and Positive Right – Comparing Germany, Europe and the United States' (2022) 14 Journal of Media Law 76, 86–87.

[21] Patrick Zurth, 'The German NetzDG as Role Model or Cautionary Tale? Implications for the Debate on Social Media Liability' (2021) 31 Fordham Intellectual Property, Media and Entertainment Law Journal 1084, 1131.

had not, in fact, led to an extraordinary increase in the amount of content removed – as initially foreseen by critics.[22]

The second argument brought against the NetzDG concerning its collateral effects on freedom of expression regards the resulting privatization of decision-making with respect to a field, that of speech governance, that necessarily affects sensitive constitutionally relevant matters.[23] With respect to such an argument, it is worth mentioning that German case law has in fact recognized that freedom of expression enjoys a third-party horizontal effect (*Drittwirkung*)[24] *vis-à-vis* providers of intermediary services and, especially, *vis-à-vis* "over-the-top" (OTT) online media service providers.

Namely, in May 2019, the German Federal Constitutional Court, the *Bundesverfassungsgericht* (BVerfG), issued a preliminary injunction ordering Facebook to unblock the account of a far-right wing party, which had posted content considered by the platform to constitute hate speech against asylum seekers, in order to allow it to partake in the electoral campaign for the upcoming European Parliament elections.[25] Arguably, the recognition of such a third-party horizontal effect of the right to freedom of expression, although it does not fully dismiss concerns relating to the privatization of speech governance, at least mitigates them, as it implies the obligation for providers to comply with the main tenets of constitutional law when exercising their moderation prerogatives.

Additionally, from an international perspective, human rights activists and scholars have argued that an inherent issue with the NetzDG is that it could serve as a regulatory model for other jurisdictions, legitimizing the adoption of similar strategies also by authoritarian regimes, with highly negative impacts upon freedom of expression. In other words, some have argued that Germany's choice of adopting this law, by representing itself a form of "authoritarian creep" within the country's democratic regime, could have a worldwide ripple effect leading to a rise of similar measures abroad with a concrete risk of abuse:

> To date, NetzDG's deputization model has not led to the "parade of horribles" often associated with privatization regimes in Germany … However, the law remains problematic. NetzDG sets a punitive, privatized regime as the standard for internet governance. While many authoritarian regimes pursue aggressive action without NetzDG, the law's proliferation has rendered acceptable previously derided policies … NetzDG allows authoritarian regimes to pursue their own antecedent agendas, using social media as a means to further erode civic participation, engagement, and protest.[26]

[22] Wischmeyer (n 4); Zurth (n 21).

[23] On the concerns entailed by the privatization of speech governance, from a constitutional and human rights standpoint, see *supra*, §3.2.2.

[24] Eric Engle, 'Third Party Effect of Fundamental Rights (*Drittwirkung*)' (2009) 5 Hanse Law Review 165.

[25] BVerfG (22 May 2019) 1 BvQ 42/19.

[26] Isabelle Canaan, 'NetzDG and the German Precedent for Authoritarian Creep and Authoritarian Learning' (2022) 28 Columbia Journal of European Law 101, 118.

4.2.1.3. Controversial aspects: NetzDG and EU law

As mentioned above, the NetzDG has been subjected to criticism also from another standpoint: that is, that of its consistency with European Union law and, precisely, with the ECD, for at least three reasons.

First, the NetzDG applies to providers with at least 2 million registered users in Germany, irrespective, therefore, of the Member State where those providers are established. This rule has been considered to be in contrast with the "country-of-origin" principle which characterizes the ECD, based on which it is up to the Member State where the provider is established to assess its compliance with the national provisions applicable to it and "Member States may not, for reasons falling within the coordinated field, restrict the freedom to provide information services from another Member State".[27] As a matter of fact, as of November 2023, the risk of an inconsistency of the NetzDG with the country-of-origin principle set by the ECD appears to be even more likely in the light of the CJEU's decision in *Google Ireland and Others*,[28] where the Court, dealing with the Austrian Federal Law on Measures for the Protection of Users of Communications Platforms,[29] concluded that the rules concerning the system of exceptions to the country-of-origin principle should not be interpreted as allowing Member States to adopt "general and abstract measures aimed at a category of given information society described in general terms and applying without distinction to any provider of that category of services".[30]

Second, whereas the ECD established a safe harbour regime shielding intermediaries from liability whenever they react "expeditiously" upon obtaining knowledge or awareness of illegal activity on their platform, the NetzDG explicitly sets a rather swift time-limit for the removal of "manifestly illegal" content (24 hours). Such a time limit could arguably be considered as being overly rigid and, thus, at odds with the provisions of the ECD. Third, the complaint management system identified by the NetzDG may lead the providers affected to having to monitor actively and constantly the content posted through their infrastructures, in violation of Article 15 ECD.[31]

Clearly, the relationship between the NetzDG and EU law should now be evaluated in the light of the adoption of the DSA. In fact, the recently enacted Regulation seems in many ways to take inspiration from Germany's legislative strategy, namely because of the choice to resort to a due diligence system for providers of intermediary services to enhance their responsibility in the dissemination and spread of illegal and harmful content. As described in Chapter 3, the DSA itself establishes a notice and action mechanism[32] reminiscent of the complaint handling procedure of the NetzDG. Such a procedure,

---

[27] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce'), OJ L 178/1 arts 3, paras 1-2.

[28] Case C-376/22, *Google Ireland Limited and Others v Kommunikationsbehörde Austria* [2023] ECLI:EU:C:2023:835.

[29] *Bundesgesetz über Maßnahmen zum Schutz der Nutzer auf Kommunikationsplattformen* (*Kommunikationsplattformen* - KoPI-G) 2020 (BGBl I, 151/20).

[30] *Google Ireland and Others* (n 28) para 64.

[31] Wischmeyer (n 4).

[32] See *supra*, §3.5.3.2.

therefore, is arguably coherent, at least in principle, with EU law because it is in compliance with the supranational framework regarding transparency duties.

However, issues may well arise with respect to the competences of German authorities, which the DSA, directly applicable also within Germany due to its nature as an EU Regulation, limits significantly. Indeed, as also noted by Advocate General Szpunar,[33] the DSA follows once again the country-of-origin principle as concerns the competences of national DSCs. Therefore, DSCs shall have powers of investigation, enforcement powers, as well as any other exceptional powers envisaged by the Regulation,[34] only with respect to those providers of intermediary services whose place of establishment is located within the territory of their own Member State.[35] In other words, also under the DSA, the German DSC should not be entitled to supervise and enforce compliance with the Regulation of any provider established outside of Germany, even when they have more than 2 million registered users within the country, contrary to the NetzDG. Moreover, the DSA provides that the Commission retains all investigative and enforcement powers concerning the set of obligations exclusive to VLOPs and VLOSEs, whereas DSCs may only exercise such powers with regards to VLOPs and VLOSEs' compliance with any of the other obligations – unless, however, the Commission has initiated proceedings for the same infringements.[36]

Thus, overall, the competences of German authorities shall be much narrower under the framework of the DSA. On top of this, although the DSA does not appear to explicitly prohibit national authorities from setting a time limit for the providers of intermediary services to respond to a complaint, the Regulation rejects the choice of setting such time limits itself, even though it acknowledges that the rules governing the notice and action mechanism should be harmonized at Union level so as to fully ensure the respect of fundamental rights as provided for by the CFREU.[37] As a result, many in Germany have perceived the DSA as a much more feeble regulatory attempt in comparison with the NetzDG.[38] In the light of such a shift in the governance competences from Berlin to

[33] Case C-376/22, *Google Ireland Limited and Others v Kommunikationsbehörde Austria* [2023] ECLI:EU:C:2023:467, Opinion of AG Szpunar [8].

[34] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act), OJ L 277/1 arts 51, paras 1-3.

[35] ibid 56, para 1.

[36] ibid 56, paras 2-4.

[37] DSA rec 52. As a matter of fact, when required to assess the conformity of the French so-called "Avia Law" proposal, which aimed to introduce a similar obligation to take down illegal content, including most notably hate speech, within 24 hours (see *infra*, §4.2.2.1), the European Commission declared that such a time limit, combined with the significant sanctions envisaged by the draft, "could lead to unacceptable outcomes, in particular disproportionate burdens for the online platforms and, in certain circumstances, a risk of over-removal and hence negative effects on freedom of expression". Such a result, in the Commission's opinion, would be in contrast with the ECD and with its goal to promote the free movement of services. European Commission, 'Delivery of Comments Pursuant to Article 5(2) of Directive (EU) 2015/1535 of 9 September 2015. Law Aimed at Combating Hate Content on the Internet' C(2019) 8585 final 7.

[38] Alina Clasen, 'Digital Services Act: Germany Proposes Creation of Advisory Board' (*www.euractiv.com*, 9 May 2023) <https://www.euractiv.com/section/platforms/news/digital-services-act-germany-proposes-creation-of-advisory-board/> accessed 7 August 2023.

Brussels, as well as in the light of the adoption of a less strict approach than that of the NetzDG, it has been argued that "the question may arise as to whether the German Constitution sets constraints on a full European harmonisation of media law, especially on any upper limits for the regulation of intermediaries".[39]

4.2.2. *Beyond the NetzDG: intermediary liability for third-party hate speech across other European experiences*

4.2.2.1. France: the laws against the manipulation of information and the (maimed) Avia Law

In recent years, France has also striven to enact legislation aimed at curtailing the spread of illegal and harmful content including, notably, disinformation and hate speech content.

Thus, in the aftermath of the 2017 presidential elections and amid rising concerns about the meddling of disinformation within democratic processes, France passed at the end of 2018 an organic law[40] and an ordinary law[41] "concerning the fight against the manipulation of information". The new framework introduced transparency obligations and cooperation duties for online platforms, extended the powers of the *Conseil Supérieur de l'Audiovisuel* (CSA, merged at the beginning of 2022 into the *Authorité de Régulation de la communication Audiovisuelle et Numérique*, ARCOM), and promoted media literacy in the educational framework. However, the most crucial part of the law concerns the "electoral period".[42] Indeed, apart from introducing additional transparency requirements regarding sponsored content, the French act establishes a new summary procedure which may be initiated on such occasions to end the dissemination of false information on online communication services.[43] In these cases, the judge is required to evaluate whether the falsehood of the information impugned is obvious, is disseminated massively and through the use of AI means, and leads to the disturbance of public peace or the sincerity of the election, and to act accordingly within 48 hours. Besides, the Senate having referred to it three questions, the *Conseil Constitutionnel* operated a preliminary review of constitutionality of the law, upholding the legislation upon certain conditions.[44]

---

[39] Hong (n 20) 87. Indeed, as recalled by Hong himself, "according to the German Constitutional Court's *Lisbon* judgment, 'the ordering of the freedom of opinion, press and of association' is one of those '[e]ssential areas' in which the principle of democracy … demands that the member states are left with 'sufficient scope' to shape their living conditions … There may therefore be a conflict with the 'variety of opinions, which is so essential for democracy, if Union law no longer leaves sufficient leeway to guarantee this variety of opinions and instead fully harmonises some upper limits for the regulation of intermediaries". ibid 87–88.

[40] *Loi organique n 2018-1201 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information*.

[41] *Loi n 2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information*.

[42] I.e., the three months period prior to the first day of the month when a general election is held. ibid 1.

[43] ibid.

[44] Notably, the *Conseil* held that a judicial order may only be issued against information that can be proven to be false in an objective manner and may not affect simple opinions, parodies, partial inaccuracies, and exaggerations; moreover, both the misleading nature of the information and the risk for consequences impacting the fairness of voting should be "manifest". Thus, the *Conseil Constitutionnel*: "*D'une part, cette*

On top of the laws against the manipulation of information, in June 2020 the French Parliament also approved a law aimed more specifically at "fighting hateful content on the Internet",[45] commonly known as the "Avia Law" from the name of its proponent. The law originally contained a range of provisions establishing new powers for national administrative authorities, including the CSA, and amending the domestic legislation implementing the ECD[46] to introduce new obligations and duties for hosting providers. Namely, the Avia Law obliged hosting providers to remove within 24 hours content such as hate speech, as well as incitement to terrorism, non-consensual pornography, and child pornography.[47] As a matter of fact, the explanatory memorandum to the Avia Law proposal explicitly mentioned Germany's NetzDG as a model for such a provision.[48] However, the *Conseil Constitutionnel*, vested once again by the Senate with the duty to verify the compliance of the law with the French Constitution, struck down many of its provisions, including that foreseeing the mentioned removal obligation, and thus watered down the impact of the new legislation.[49]

---

procédure ne peut viser que des allégations ou imputations inexactes ou trompeuses d'un fait de nature à altérer la sincérité du scrutin à venir. Ces allégations ou imputations ne recouvrent ni les opinions, ni les parodies, ni les inexactitudes partielles ou les simples exagérations. Elles sont celles dont il est possible de démontrer la fausseté de manière objective. D'autre part, seule la diffusion de telles allégations ou imputations répondant à trois conditions cumulatives peut être mise en cause: elle doit être artificielle ou automatisée, massive et délibérée. Cependant, la liberté d'expression revêt une importance particulière dans le débat politique et au cours des campagnes électorales … Dès lors, compte tenu des conséquences d'une procédure pouvant avoir pour effet de faire cesser la diffusion de certains contenus d'information, les allégations ou imputations mises en cause ne sauraient … justifier une telle mesure que si leur caractère inexact ou trompeur est manifeste. Il en est de même pour le risque d'altération de la sincérité du scrutin, qui doit également être manifeste" Cons Const (20 December 2018) 2018-773 DC, *Loi relative à la lutte contre la manipulation de l'information* [21–23]. See, with regard to the *Loi relative à la lutte contre la manipulation de l'information* and to the interpretation of the *Conseil Constitutionnel*, Philippe Mouron, 'Du Sénat Au Conseil Constitutionnel: Adoption Des Lois de Lutte Contre La Manipulation de l'information' (2019) 49 Revue européenne des médias et du numérique 9; Oreste Pollicino, Marco Bassini and Giovanni De Gregorio, *Internet Law and Protection of Fundamental Rights* (Bocconi University Press 2022) 116–118. The first decision applying the new law occurred in May 2019, in a case concerning the request for the removal of a tweet by the incumbent Minister of the Interior: following the indications of the *Conseil Constitutionnel*, however, the Paris TGI dismissed the request. See TGI Paris (10 May 2019) RG 19/53935, *Vieu et Ouzoulias v Twitter France SAS*.

[45] *Loi n 2020-766 du 24 juin 2018 visant à lutter contre les contenus haineux sur internet*. French law, most notably, punishes the act of direct public incitement to hatred, violence, and discrimination against a person or a group of persons "*à raison de leur sexe, de leur orientation sexuelle ou identité de genre ou de leur handicap*" under the *Loi du 29 juillet 1881 sur la liberté de la presse* (French Law on Freedom of the Press) art 24. Additionally, since 2017, non-public incitement to hatred, violence, or discrimination constitutes a minor offence (*contravention*) pursuant to the *Code Pénal* (French Criminal Code) art R625-7.

[46] *Loi n 2004-575 du 21 juin 2004 pour la confiance dans l'économie numérique*.

[47] Avia Law art 1, para II.

[48] "Depuis la loi NetzDG du 1er octobre 2017, l'Allemagne a renforcé la responsabilité des plateformes en exigeant la mise en place de procédures de traitement des signalements efficaces et transparentes, ainsi que le retrait des contenus illicites sous 24 heures sous peine de lourdes sanctions financières". Assemblée Nationale, 'Proposition de Loi n 1785 Visant à Lutter Contre Les Contenus Haineux Sur Internet' (Assemblée Nationale, 20 March 2019) <https://www.assemblee-nationale.fr/dyn/15/textes/l15b1785_proposition-loi#D_non_amendable_0> accessed 10 August 2023.

[49] Cons Const (18 June 2020) 2020-801 DC, *Loi visant à lutter contre les contenus haineux sur internet*. On the judgment of the *Conseil Constitutionnel* see, among others, Évelyne Bonis and Virginie Peltier, 'Chronique de droit pénal et de procédure pénale: (janvier 2020 à juin 2020)' (2020) 5 Titre VII 112, 115–116; Cecilia Siccardi, 'La *Loi Avia*. La Legge Francese Contro l'Odio *On Line* (O Quello Che Ne Rimane)'

With regard to the obligation to remove illegal content within 24 hours, the *Conseil Constitutionnel* observed that such an obligation would arise as soon as the provider were notified about its existence, without the need for any preventive intervention of the judicial authority. Thus, the decision concerning the illegal nature of any information uploaded to its infrastructure would be up to the provider alone, that is, to a private entity, even though such a decision may require a significant degree of legal expertise. Moreover, the *Conseil* considered that a time limit of 24 hours would be extremely short to operate such an assessment.[50] These arguments, coupled with the absence of reasonable causes for the waiving of responsibility and with the entity of the sanctions established,[51] led the *Conseil* to conclude that the provision represented an interference with the right to freedom of expression and communication that did not comply with the criteria of necessity, adequacy, and proportionality that are required under French constitutional law.[52]

As a result, the ultimate text of the Avia Law only contains a set of rather light preventive measures aimed at countering hate speech:[53] notably, the law required to simplify the notification procedure for users to flag unlawful content,[54] introduced a specialized tribunal,[55] amended the French Education Code to promote the education of children in schools with respect to the issue of online hate speech and online violence,[56] and introduced, within the CSA (today ARCOM), a dedicated observatory concerning online hatred (*Observatoire de la haine en ligne*).[57]

4.2.2.2. Italy: of failed legislative attempts and an inconsistent case law

Like Germany and France, Italy has made some attempts to address the issue of the spread of harmful and/or unlawful content across the Internet, mainly with a view to reducing the amount of disinformation and hate speech online: in Italy, however, these attempts have all been short-lived. With respect to disinformation, two legislative proposals were presented in February 2017 and December 2017, aiming respectively at introducing "provisions to prevent the manipulation of online information, to ensure transparency on the

in Marilisa D'Amico and Cecilia Siccardi (eds), *La Costituzione non odia: Conoscere, prevenire e contrastare l'*hate speech on line (Giappichelli 2021) 179–182; Caterina Severino, 'La Democrazia Francese e Le Sfide Del Digitale: Tra Opportunità e Rischi' (2021) 3 Rivista Gruppo di Pisa 33, 37–39.

[50] Cons. Const. (18 June 2020) 2020-801 DC (n 49) paras 14–16.

[51] ibid 17–18.

[52] "Il résulte de ce qui précède que, compte tenu des difficultés d'appréciation du caractère manifestement illicite des contenus signalés dans le délai imparti, de la peine encourue dès le premier manquement et de l'absence de cause spécifique d'exonération de responsabilité, les dispositions contestées ne peuvent qu'inciter les opérateurs de plateforme en ligne à retirer les contenus qui leur sont signalés, qu'ils soient ou non manifestement illicites. Elles portent donc une atteinte à l'exercice de la liberté d'expression et de communication qui n'est pas nécessaire, adaptée et proportionnée. Dès lors … le paragraphe II de l'article 1er est contraire à la Constitution". ibid 19.

[53] Siccardi (n 49) 181.

[54] Avia Law art 2.

[55] ibid 10.

[56] ibid 13–15.

[57] ibid 16.

web and to encourage media literacy" (so-called "Gambaro Draft Law")[58] and at establishing "general rules on social networks and to combat the dissemination of illegal content and fake news on the Internet" (so-called "Zanda-Filippin Draft Law"),[59] but never actually made it to parliamentary debate.[60]

As for online hate speech, at least three proposals were presented during the XVIII Legislature (2018-2022) with the purpose of tackling the phenomenon:[61] AS 634 of 2018, containing "amendments to the Criminal Code and other provisions in the matter of countering incitement to hatred and to discrimination (hate speech)" (Boldrini I Draft Law);[62] AS 1455 of 2019, introducing "measures to counter the phenomenon of incitement to hatred on the web" (Fedeli Draft Law);[63] and AC 2936 of 2021, establishing "measures for the prevention and the countering of the dissemination of manifestations of hate through the Internet network" (Boldrini II Draft Law).[64] On top of these, an attempt was

[58] AS 2688 (XVII), *Disposizioni per prevenire la manipolazione dell'informazione* online, *garantire la trasparenza sul* web *e incentivare l'alfabetizzazione mediatica*. The proposal sought to amend the Italian Criminal Code by introducing two sets of provisions criminalizing: the publication or dissemination of "false, exaggerated or biased news likely to disrupt public order, through computer platforms"; the dissemination of false news capable of raising public alarm or of misleading sectors of the public opinion; and the dissemination of false news "concerning hate campaigns and campaigns aimed at undermining the democratic process". It also provided for communication and rectification requirements, as well as intermediary liability in case of publication or dissemination of "fake news". The Gambaro Draft Law, however, sparked almost unanimous criticism due to its failure to ensure the protection of fundamental constitutional principles: see, among others, Marco Bassini and Giulio Enea Vigevani, 'Primi Appunti Su *Fake News* e Dintorni' (2017) 1 Rivista di Diritto dei Media 11, 13.

[59] AS 3001 (XVII), *Norme generali in materia di* social network *e per il contrasto della diffusione su* internet *di contenuti illeciti e delle* fake news. The Zanda-Filippin Draft Law notably built upon the model of the German NetzDG, which had only recently been adopted, as it did not introduce new types of criminal offences (like the Gambaro Draft Law had aimed to do), but rather relied upon already existing provisions of the Criminal Code. The proposal envisaged however the establishment of new obligations for providers of social networking services, consisting, notably, of the creation of complaint-handling procedures and of the publication of transparency reports. See, with regard to such a proposal, Davide Zecca, 'Tutela Dell'Integrità Dell'Informazione e Della Comunicazione in Rete: Obblighi per Le Piattaforme Digitali Fra Fonti Comunitarie e Disciplina Degli Stati Membri' (2019) 37 DPCE Online 889, 903–904; Matteo Monti, 'La proposta del ddl Zanda-Filippin sul contrasto alle fake news sui social network: profili problematici' (*Diritti Comparati*, 7 December 2017) <https://www.dirittocomparati.it/la-proposta-del-ddl-zanda-filippin-sul-contrasto-alle-fake-news-sui-social-network-profili-problematici/> accessed 11 August 2023.

[60] A subsequent attempt to curtail the phenomenon of online disinformation consisted of the launch at the beginning of 2018 of an operational protocol for the "Fight Against the Diffusion of Fake News through the Web" by the Ministry of the Interior and the Chief of Police, which included the creation of a "red button" alert intended to allow citizens to directly report false content to the Postal Police. Such a tool, however, faced stark opposition and was pulled back almost immediately. In recent years AGCOM, i.e., the Italian Communications Regulatory Authority, has taken a leading role in the fight against disinformation, taking a more careful approach than the legislative and administrative attempts described above. Namely, AGCOM created a technical working group to safeguard "pluralism and fairness of information on digital platforms" to encourage and promote self-regulation of online platforms and the exchange of good practices. See Pollicino, Bassini and De Gregorio (n 44) 114–115.

[61] See, in this respect, Pietro Villaschi, 'I Progetti Di Legge In Discussione In Italia: Analisi Critica' in Marilisa D'Amico and Cecilia Siccardi (eds), *La Costituzione non odia: Conoscere, prevenire e contrastare l'*hate speech on line (Giappichelli 2021).

[62] AS 634 (XVIII), *Modifiche al codice penale e altre disposizioni in materia di contrasto dell'istigazione all'odio e alla discriminazione (*hate speech*)*.

[63] AS 1455 (XVIII), *Misure per il contrasto del fenomeno dell'istigazione all'odio sul* web.

[64] AC 2936 (XVIII), *Misure per la prevenzione e il contrasto della diffusione di manifestazioni d'odio mediante la rete* internet.

made to amend the current criminal law framework on hate speech, with a view to broadening the set of grounds of discrimination considered so as to include also sex, gender, sexual orientation, gender identity, and disability: [65] this proposal, which was commonly known as the "Zan Draft Law" and sparked a considerable (and rather polarized) debate in Italy,[66] did not contain provisions concerning online hate speech specifically, but took a more generic approach towards fighting homo-transphobic, misogynistic, and ableist discrimination and violence; ultimately, it was quashed by the Senate in October 2021.

The Boldrini I Draft Law, which also envisaged an extension of relevant grounds of discrimination, aimed *inter alia* at introducing an obligation for providers of websites, social networks and online platforms to notify the judicial authority of the presence of any content deemed to amount to hate speech. Once the illegal nature of the content had been ascertained, the Postal Police would order the provider to take the necessary technical measures against it.[67] Therefore, as has been noted,[68] the Boldrini I Draft Law would have kept such an assessment a prerogative of the (public) judicial authority, rather than delegating this decision to (private) providers of intermediary services: conversely, both the Fedeli Draft Law and the Boldrini II Draft Law[69] aimed at obliging providers to perform the assessment of the content themselves within a time frame of 24 hours – for this purpose, besides, providers should have created a dedicated independent internal body.[70] Additionally, the draft laws foresaw the duty of providers receiving more than a hundred notifications per year to produce six-monthly transparency reports.[71] The Boldrini II Draft Law also vested the Italian Data Protection Authority with powers to sanction providers not complying with the new requirements.[72] Like the Gambaro and the Zanda-Filippin Draft Laws, however, none of these proposals were ever brought to parliamentary discussion: nor does it seem likely, at the time of writing, that they will be discussed during the current XIX Legislature.[73]

---

[65] AS 2005 (XVIII), *Misure di prevenzione e contrasto della discriminazione e della violenza per motivi fondati sul sesso, sul genere, sull'orientamento sessuale, sull'identità di genere e sulla disabilità*.

[66] Costanza Nardocci, 'Dalla Parola Che Discrimina Alla Parità Nel Linguaggio: La Dimensione Sovranazionale (E Comparata)' in Marina Brambilla and others (eds), *Genere, disabilità, linguaggio. Progetti e prospettive a Milano* (Franco Angeli 2022) 54.

[67] Boldrini I Draft Law art 3.

[68] Villaschi (n 61) 193–194.

[69] The latter also provided for an extension of relevant grounds of discrimination in the same terms as the Zan Draft Law: see Boldrini II Draft Law art 3.

[70] Fedeli Draft Law art 5; Boldrini II Draft Law art 5.

[71] Fedeli Draft Law art 6; Boldrini II Draft Law art 6.

[72] Boldrini II Draft Law art 7. The draft law also foresaw the right (both for adults and minors) to require the shadowing, removal or block of data or images about them from the Internet, as well as the introduction of educational curricula within schools concerning the phenomenon of online hatred and violence and aimed at fostering a more responsible use of the Internet (arts 8-9).

[73] Conversely, AGCOM has passed some important administrative acts to reduce the dissemination of hate speech content across audiovisual media services, as envisaged by *Decreto legislativo 8 novembre 2021, n 208, Testo Unico dei Servizi di Media Audiovisivi* (TUSMA) art 30. The TUSMA, implementing Directive (EU) 2018/1808 (see *supra*, §3.4.3.2), establishes indeed that AGCOM may adopt such instruments to orient providers of AVMS in scheduling programmes that do not contain incitement or justification of criminal offences, including, most notably, incitement to violence or discrimination. The latest of such administrative acts is represented by *Delibera n 37/23/CONS, Regolamento in materia di tutela dei diritti*

In the absence of a clear legislative framework on the responsibilities and duties of providers of intermediary services, Italian courts have had to face some tricky legal questions, often – and quite alarmingly – with rather different outcomes. These questions arose from a number of cases concerning Facebook's decisions to block the accounts of CasaPound and Forza Nuova (as well as of many of their members), these being two alt-right parties whose contents were deemed by Facebook to be in violation of the community standards concerning hate speech. Both parties brought action against the platform, alleging that such a choice represented an infringement of their freedoms of expression and association. Facebook, conversely, argued that its relationship with CasaPound and Forza Nuova was of a contractual nature and that, the parties having failed to comply with the clauses of the community standards, it was Facebook's right to terminate the contract and thus interrupt the provision of its services.[74]

In the case of *CasaPound v Facebook*, the Tribunal of Rome first issued a summary order granting CasaPound's request to have its account reinstated on the social networking platform.[75] According to the judge, the relationship between Facebook and its users, including CasaPound, could not be assimilated to a traditional contractual relationship between private parties, as the platform is set in a "special position" and plays a highly significant role in the enjoyment of users' fundamental rights and prerogatives. Consequently, Facebook has a duty to ensure the full respect of constitutional principles when providing its services to the public. In the present case, the judge considered Facebook's choice to suspend the party's account to be arbitrary and not founded on a full cognizance of the facts of the case, in breach of the principle of pluralism of expression and information.

Facebook impugned the decision, but the appeal was rejected by the collegial court of the Tribunal of Rome.[76] The judges argued that the relationship between Facebook and its users should in fact be recognized as a private contractual one. However, they held nonetheless that the contract between the two parties could not translate into an indiscriminate compression of the fundamental and constitutionally protected rights of the recipients of the service, so that the exercise of such rights, including those to freedom of

fondamentali della persona ai sensi dell'art. 30 del decreto legislativo 8 novembre 2021, n 208, adopted in February 2023. The latter, however, only applies to providers of AVMS that have editorial responsibility upon the programmes they broadcast: providers of video-sharing platforms are thus excluded from its scope of application. See, in this respect, Giulio Enea Vigevani, 'Informazione e Potere' in Marta Cartabia and Marco Ruotolo (eds), *Enciclopedia del Diritto*, vol. *Potere e Costituzione* (Giuffrè 2023) 219 241–242.

[74] See, among others, Angelo Jr Golia, 'L'Antifascismo Della Costituzione Italiana Alla Prova Degli Spazi Giuridici Digitali. Considerazioni Su Partecipazione Politica, Libertà D'Espressione *Online* E Democrazia (Non) Protetta In *CasaPound c. Facebook* E *Forza Nuova c. Facebook*' (2020) 18 Federalismi.it 134; Ottavio Grandinetti, '*Facebook vs. CasaPound e Forza Nuova*, Ovvero La Disattivazione Di Pagine Social E Le Insidie Della Disciplina Multilivello Dei Diritti Fondamentali' (2021) 1 Rivista di Diritto dei Media 173; Marco Bassini, 'Libertà Di Espressione E Social Network, Tra Nuovi "Spazi Pubblici" E "Poteri Privati". Spunti Di Comparazione' (2021) 2 Rivista di Diritto dei Media 67, 93–97; Pietro Villaschi, 'La (Non) Regolamentazione Dei *Social Network* E Del *Web*' in Marilisa D'Amico and Cecilia Siccardi (eds), *La Costituzione non odia: Conoscere, prevenire e contrastare l'*hate speech on line (Giappichelli 2021) 118–123.

[75] Tribunal of Rome, specialized section for business, order of 12 December 2019.

[76] Tribunal of Rome, XVII civil section, order of 29 April 2020.

assembly and to freedom of expression, as protected respectively by Articles 18 and 21 of the Constitution, may not constitute a valid cause for terminating the relationship.[77]

The case of *Forza Nuova v Facebook*, conversely, took the opposite direction. Quite curiously, also this controversy was brought before the Tribunal of Rome. However, this time the judge rejected Forza Nuova's claims, holding not only that Facebook's choice to discontinue its services and to suspend the accounts of the party and of its members was perfectly legitimate but also that Facebook, in blocking the party's account, had in fact complied with a legal obligation to act in such a way. The order, indeed, refers to a wide range of international and supranational sources of law and case law (including the ICCPR, the ICERD, the case law of the ECtHR, Framework Decision 2008/913/JHA, and the ECD),[78] as well as to the domestic Criminal Code provisions on hate speech,[79] concluding that, the contents posted by Forza Nuova being clearly unlawful, Facebook would in fact have been liable had it failed to discontinue its services to the party once it had been made aware that such materials had been uploaded to its infrastructures.[80]

Such a stark difference in the outcomes of the discussed cases reveals a rather concerning uncertainty with respect to the duties and obligations of providers of intermediary services, notably online platforms, when it comes to the moderation of hate speech contents in Italy. On the one hand, the judges in the case concerning CasaPound followed a reasoning reminiscent of the German principle of *Drittwirkung*, that is, they recognized *de facto* the horizontal applicability of fundamental rights *vis-à-vis* the providers of intermediary services. On the other hand, the judge in the Forza Nuova case focused its attention on the responsibility of platforms to moderate content and to reduce the spread of illegal and harmful content. Such an uncertainty clearly represents a significant issue for the Italian framework on intermediary liability and hate speech prevention: hopefully, in

[77] ibid 10.

[78] See *supra*, §2.

[79] *Regio decreto 19 ottobre 1930, n 1398, Codice penale* (Italian Criminal Code) arts 604bis-604ter. Italy first criminalized racist and religious-based hate speech in 1975, when it implemented the ICERD through the *Legge 13 ottobre 1975, n 654, Ratifica ed esecuzione della convenzione internazionale sull'eliminazione di tutte le forme di discriminazione razziale, aperta alla firma a New York il 7 marzo 1966*. Subsequently, a significant revision of the framework was made by the so-called "Mancino Law", i.e., the *Legge 25 giugno 1993, n 205, Conversione in legge, con modificazioni, del decreto-legge 26 aprile 1993, n 122, recante misure urgenti in materia di discriminazione razziale, etnica e religiosa*. Finally, a last reform was operated by *Decreto legislativo 1 marzo 2018, n 21, Disposizioni di attuazione del principio di delega della riserva di codice nella materia penale a norma dell'articolo 1, comma 85, lettera q), della legge 23 giugno 2017, n 103*. For a synthetic overview of the current legislation, as well as of the relevant case law by the Constitutional Court and by the Court of Cassation, see Marilisa D'Amico, 'Odio *On Line*: Limiti Costituzionali e Sovranazionali' in Marilisa D'Amico and Cecilia Siccardi (eds), *La Costituzione non odia: Conoscere, prevenire e contrastare l'*hate speech on line (Giappichelli 2021) 26–29; Costanza Nardocci, 'L'Odio Razziale E Religioso' in Marilisa D'Amico and Cecilia Siccardi (eds), *La Costituzione non odia: Conoscere, prevenire e contrastare l'*hate speech on line (Giappichelli 2021) 44–50; Nannerel Fiano, 'Antisemitismo E Negazionismo. Un Fenomeno Ancora Attuale' in Marilisa D'Amico and Cecilia Siccardi (eds), *La Costituzione non odia: Conoscere, prevenire e contrastare l'*hate speech on line (Giappichelli 2021) 64–66.

[80] Tribunal of Rome, specialized section for the rights of the person and civil immigration, order of 23 February 2020. Besides, a similar conclusion had been reached within Tribunal of Siena, civil section, order of 19 January 2020.

this respect, the DSA may contribute to bringing clarity and to creating a more coherent case law.

4.2.2.3. Spain: the *Protocolo para combatir el discurso de odio en línea*

Whereas Germany, France, and Italy have all attempted – with more or less success – to address the phenomenon of online hate speech through the enactment of hard legislation setting duties and responsibilities for providers of intermediary services, Spain underwent a different path through the adoption of the *Protocolo para combatir el discurso de odio ilegal en línea* (Protocol to Combat Illegal Hate Speech Online).[81] The Protocol was drafted in the light of rising concerns about the spread of illegal hate speech on the Internet, considered to be a threat for the individuals and groups it targets and to have a negative impact upon those who stand up for freedom, as well as representing a challenge for democratic speech and harmonious interactions, especially in the wake of the COVID-19 pandemic. The purpose of the Protocol is, therefore, to define and facilitate forms of cooperation between all signatories, with a view to countering the phenomenon of hate speech in Spain according to Spanish law. In this respect, the Protocol is declared to be mainly inspired by the EU Code of Conduct on Illegal Hate Speech.[82]

The Protocol is structured into six sections. Section I provides an overview of the relevant domestic and EU law, as well as of the ECHR framework, concerning hate speech; it also mentions the liability framework for hosting service providers under the ECD[83] and the *Ley de Servicios de la Sociedad de Información y Comercio Electrónico* (LSSI),[84] implementing the ECD itself. Section II identifies the Computer Crime Unit of the Office of the Prosecutor General (*Unidad contra la Criminalidad Informática de la Fiscalía General del Estado*) as a Point of Contact for competent authorities to communicate with hosting service providers, while proposing at the same time the drafting of a list of those public competent authorities that shall have the actual responsibility to report illegal hate

[81] *Protocolo para combatir el discurso de odio ilegal en línea* 2021. Signatories include the General Council of the Judiciary (*Consejo General del Poder Judicial*), the Office of the Prosecutor General (*Fiscalía General del Estado*), seven Ministries (the *Ministerio de Justicia*, *Ministerio de Interior*, *Ministerio de Educación y Formación Profesional*, *Ministerio de Cultura y Deporte*, *Ministerio de Derechos Sociales y Agenda 2030*, *Ministerio de Igualdad*, and *Ministerio de Inclusión, Seguridad Social y Migraciones*), and the Spanish Digital Economy Association (*Asociación Española de la Economia Digital*), representing the major providers of hosting services.

[82] ibid *Preámbulo*. Pursuant to the Protocol, at I.1, the notion of "illegal hate speech" (*discursos de odio*) includes first and foremost the offences punishable under the *Ley Orgánica 10/1995, de 23 de noviembre, del Código Penal* (Spanish Criminal Code) art 510. These include most notably the public promotion and incitement to hatred, discrimination, or violence against persons or groups "*por motivos racistas, antisemitas, antigitanos u otros referentes a la ideología, religión o creencias, situación familiar, la pertenencia de sus miembros a una etnia, raza o nación, su origen nacional, su sexo, orientación o identidad sexual, por razones de género, aporofobia, enfermedad o discapacidad*". Additionally, the notion applies to any other hate crime consisting of acts of expression or communication based on discriminatory bias pursuant to art 22, para 4, of the Criminal Code, as well as to the offences referred to in the *Ley 19/2007, de 11 de julio, contra la violencia, el racismo, la xenofobia y la intolerancia en el deporte* art 23, para 1, letts b)-c) (criminalizing a range of acts of expression or communication in the context of sports meetings and competitions).

[83] See *supra*, §3.4.1.

[84] *Ley 34/2002, de 11 de julio, de Servicios de la Sociedad de Información y Comercio Electrónico*.

speech online to the Point of Contact; the Section also provides for some operational rules to guarantee the cooperation between hosting service providers and the Point of Contact. Section III introduces the category of trusted flaggers, requiring hosting service providers to grant notifications received from them preferential processing, whereas Section IV concerns the accreditation procedure and training programmes for trusted flaggers. Section V, following Commission Recommendation (EU) 2018/334 on Illegal Content Online,[85] concerns the implementation of redress mechanisms to inform citizens about alternative dispute settlement mechanisms concerning hate speech, without the need to refer to Spanish criminal courts to seek justice. Finally, Section VI addresses the monitoring and evaluation of the Protocol, which is attributed to the Monitoring Committee established by the Interinstitutional Agreement to cooperate in combating racism, xenophobia, LGBTI-phobia, and other forms of intolerance.[86]

The Protocol clearly features a rather different approach from that adopted by Germany's NetzDG and those attempted by France and Italy, as it is characterized by a co-regulatory, instead of regulatory, strategy. Indeed, the Protocol contains some important guidelines for providers of hosting services with a view to orient them in the implementation of their moderation practices against hate speech and, contextually, in complying with the existing rules on intermediary liability under the ECD and the LSSI. In this respect, the Spanish strategy appears to anticipate, albeit at a merely domestic level, the DSA's choice to vest with increased importance the adoption of codes of conduct as co-regulatory means to promote the fight against unlawful and harmful content across the Internet.[87] Arguably, the Spanish Protocol will thus represent an instrument coherent and consistent with the framework envisaged by Regulation (EU) 2022/2065.

4.2.3. *Democratic backsliding and speech governance in Eastern Europe: the case of "memory laws" in Poland and Hungary*

The previous subparagraphs have outlined some of the attempts to regulate online hate speech and, notably, the responsibilities and duties of providers of intermediary services, with a view to highlighting the similarities, differences, and potential sources of conflict between the EU framework and Member States' domestic strategies. For instance, as noted above, the approval of the DSA raises constitutional concerns in the German context, as it may well translate into an increased shift of speech governance powers from Berlin to Brussels. It will thus be paramount for the EU to address in a careful manner

[85] Commission Recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online, OJ L63/50.

[86] *Acuerdo suscrito entre el Consejo General del Poder Judicial, la Fiscalía General del Estado, el Ministerio de Justicia, el Ministerio de Interior, el Ministerio de Educación y Formación Profesional, el Ministerio de Trabajo, Migraciones y Seguridad Social, el Ministerio de la Presidencia, Relaciones con las Cortes e Igualdad, el Ministerio de Cultura y Deporte y el Centro de Estudios Jurídicos para cooperar institucionalmente en la lucha contra el racismo, la xenofobia, la LGBTIfobia y otras formas de intolerancia* 2018.

[87] See *supra*, §3.5.3.5.

these sensitive matters, so as to avoid dangerous crises and the risk of a balkanization of national practices in the application of the new Regulation.

At the same time, the application and enforcement of the DSA will likely have to face the delicate issues resulting from the open concept of "illegal content" which, pursuant to Article 3, letter (h), refers to any information that "is not in compliance with Union law or the law of any Member State which is in compliance with Union law, irrespective of the precise subject matter or nature of that law". As argued in Chapter 3, such a definition could lead to asymmetrical responses across Member States with respect to hate speech governance, especially in light of the different grounds of discrimination treated as relevant depending on the jurisdiction considered:[88] thus, for instance, Italy is still missing a law proscribing misogynistic, LGBTQIA+-phobic, or ableist hate speech.[89] However, a different set of issues may arise with respect to domestic legislations setting limitations to freedom of expression that are arguably at risk of promoting the silencing of minorities and pose, therefore, significant concerns with respect to ECHR and CFREU principles themselves. An example of this is represented by the rise of "memory laws" across Eastern European countries, and especially in Poland and Hungary, which have been associated with those countries' democratic backsliding.[90]

The expression "memory laws" broadly refers, in general, to any legislative measure aimed at governing the relationship with history in a given country, including provisions recognizing and commemorating historical events and figures, but also punitive measures against the denial of historical atrocities or bans on the use of symbols that are connected to totalitarian regimes.[91] A clear example is, therefore, the prohibition of Holocaust denialism, which is, as described in Chapter 2, consistent with the ECHR and CFREU frameworks.[92] Similarly, many former Soviet Union countries introduced, after 1989, memory laws concerning the violences, crimes and episodes of oppression linked to the communist experience of the Soviet Union. Nevertheless, more recent memory laws in countries such as Poland and Hungary (as well as, outside of the EU, Ukraine and Russia) have become tools to promote deeply nationalistic propaganda, to such an extent as to give rise to what have been defined "memory wars".[93]

The resort to such mnemonical strategies to promote nationalistic instances has been referred to by Mälksoo as "militant memocracy":

---

[88] See *supra*, §3.5.4.1

[89] See *supra*, §4.2.2.2.

[90] See, e.g., Marta Bucholc, 'Commemorative Lawmaking: Memory Frames of the Democratic Backsliding in Poland After 2015' (2019) 11 Hague Journal on the Rule of Law 85; Marina Bán and Uladzislau Belavusau, 'Memory Laws' (SSRN, 9 May 2022) <https://papers.ssrn.com/abstract=4104552> accessed 14 August 2023.

[91] Uladzislau Belavusau, 'Law and the Politics of Memory' in Maria Mälksoo (ed), *Handbook on the Politics of Memory* (Edward Elgar Publishing 2023).

[92] See *supra*, §2.2.3.

[93] Uladzislau Belavusau, Aleksandra Gliszczyńska-Grabias and Maria Mälksoo, 'Memory Laws and Memory Wars in Poland, Russia and Ukraine' (2021) 69 Jahrbuch des öffentlichen Rechts 95. Quite evidently, the governance of the country's historical memory has not been uninfluential in the context of the 2022 Russian attack on Ukraine.

> Similar to militant democracy, militant memocracy is ready to compromise certain democratic standards for the sake of thus defending the system's feasibility – only that its prevailing political concern is defending a state-endorsed version of the past to sustain a national/state identity in the present rather than the protection of core democratic values as the foremost normative criteria … Unlike its militant democracy counterpart, militant memocracy is definitively not about defending the liberal core of democracy (such as fundamental rights, the rule of law, pluralism, and the protection of minorities) in the first place. Quite the opposite: seeking to protect a national historical memory/mnemonic narrative from alternative accounts contesting it, militant memocracy can go to great lengths about restricting liberal rights … militant memocracy has the status of state's official narrative, its national honour, good name, and standing in contemporary international relations at its core.[94]

According to Belavusau, although Western memory laws condemning the denial, justification or gross trivialization of genocides and other crimes against humanity (including, notably, the Holocaust) may clearly be themselves questioned and subjected to criticisms, a significant difference lies precisely in that they are "strongly embedded into the paradigm of *militant democracy*" and are, as a result, the reflection of a "dignity-based paradigm … leading to the adoption of … so-called *self-inculpatory* memory laws" where the dignity of victims of such genocides and crimes is central. Conversely, the recent wave of Central and Eastern European memory laws aims to "fortif[y] a victimhood of national states and majority nations", so that such legislation rather takes a "*self-exculpatory*" turn.[95]

In Poland, the most striking example of such memory laws driven by purposes of "militant memocracy" is represented by the much-contested amendments to the 1998 Institute of National Remembrance Act (INRA).[96] At the beginning of 2018, under the guidance of the Polish government, led by the right-wing party Law and Justice (*Prawo i Sprawiedliwość*), a first Act Amending the INRA[97] was passed. The reform, in particular, criminalized the act of publicly and "contrary to the facts" attributing to the Polish nation or to the Polish state a responsibility or co-responsibility for partaking in Nazi crimes. Notably, the provision would apply *inter alia* to any form of recognition of the responsibility of Poles in participating in the Jedwabne Pogrom of 1941, when over 300 Jewish people were burnt alive in a barn.[98]

---

[94] Maria Mälksoo, 'Militant Memocracy in International Relations: Mnemonical Status Anxiety and Memory Laws in Eastern Europe' (2021) 47 Review of International Studies 489, 504–505.

[95] Uladzislau Belavusau, 'Mnemonic Constitutionalism and Rule of Law in Hungary and Russia' (2020) 1 The Interdisciplinary Journal of Populism 16, 17.

[96] *Ustawa z dnia 18 grudnia 1998 r. o Instytucie Pamięci Narodowej - Komisji Ścigania Zbrodni przeciwko Narodowi Polskiemu* (Act of 18 December 1998 on the Institute of National Remembrance - Commission for the Prosecution of Crimes against the Polish Nation), Dz.U. 1998 Nr 155 poz. 1016.

[97] *Ustawa z dnia 26 stycznia 2018 r. o zmianie ustawy o Instytucie Pamięci Narodowej - Komisji Ścigania Zbrodni przeciwko Narodowi Polskiemu, ustawy o grobach i cmentarzach wojennych, ustawy o muzeach oraz ustawy o odpowiedzialności podmiotów zbiorowych za czyny zabronione pod groźbą kary* (Act of 26 January 2018 Amending the Act on the Institute of National Remembrance – Commission for the Prosecution of Crimes against the Polish Nation, the Act on Military Graves and Graveyards, the Museums Act as well as the Act on the Criminal Liability of Collective Entities for Punishable Offences), Dz.U. 2018 poz. 369.

[98] Belavusau, Gliszczyńska-Grabias and Mälksoo (n 93).

Following widespread domestic and international criticisms, as well as a souring of Poland's relations with numerous countries including Israel,[99] a second Act Amending the INRA was passed in June 2018,[100] repealing the most controversial provisions of the January reform. Most notably, the attribution to the Polish nation or state of responsibilities in Nazi-related crimes was de-criminalized. Nevertheless, Polish civil courts may still have a margin of action to levy civil law sanctions,[101] even though recent case law has been rather ambiguous on the point:[102] thus, the posting of contents accusing Poles of partaking in the Holocaust may well still represent an illegal activity under Polish civil legislation.

Hungary, throughout the 2010s, also adopted legal measures to govern the collective historical memory of the country with a view to promoting nationalistic instances. In the case of Hungary, in fact, the adoption of such a strategy is even embedded within the state's constitutional framework: thus, the Hungarian approach has been considered to constitute a form of "mnemonic constitutionalism".[103] Notably, the 2011 Hungarian Fundamental Law[104] features a preamble containing a National Avowal which has been

[99] Jon Henley, 'Poland Provokes Israeli Anger with Holocaust Speech Law' *The Guardian* (1 February 2018) <https://www.theguardian.com/world/2018/feb/01/poland-holocaust-speech-law-senate-israel-us> accessed 16 August 2023.

[100] *Ustawa z dnia 27 czerwca 2018 r. o zmianie ustawy o Instytucie Pamięci Narodowej - Komisji Ścigania Zbrodni przeciwko Narodowi Polskiemu oraz ustawy o odpowiedzialności podmiotów zbiorowych za czyny zabronione pod groźbą kary* (Act of 27 June 2018 Amending the Act on the Institute of National Remembrance – Commission for the Prosecution of Crimes against the Polish Nation and the Act on the Criminal Liability of Collective Entities for Punishable Offences), Dz.U. 2018 poz. 1277.

[101] Jörg Hackmann, 'Defending the "Good Name" of the Polish Nation: Politics of History as a Battlefield in Poland, 2015-18' (2018) 20 Journal of Genocide Research 587, 603.

[102] As highlighted by Aleksandra Gliszczyńska-Grabias and Michał Jabłoński, 'Is One Offended Pole Enough to Take Critics of Official Historical Narratives to Court?' (*Verfassungsblog*, 12 October 2019) <https://verfassungsblog.de/is-one-offended-pole-enough-to-take-critics-of-official-historical-narratives-to-court/> accessed 16 August 2023, civil courts have generally referred to the Civil Code's provisions on defamation to levy civil law sanctions against persons associating Poles with Nazi crimes and antisemitic violences. Most notably, a number of lawsuits have been brought by the Polish League Against Defamation's president, Mr Świrski, against such articles and declarations: following the interpretation of the Supreme Court, a variety of decisions have recognized Mr Świrski's legal stance and awarded him the right to receive an apology. More recently, in the case of *Leszczyńska v Engelking and Grabowski*, a Polish woman brough action against Prof. Barbara Engelking and Prof. Jan Grabowski, who had co-edited a book touching on the complicity of Poles in the genocide of Jews under the Nazi occupation: the book referred, specifically, to the possible implications of the plaintiff's uncle, Edward Malinowski, wartime mayor of a village in North-Eastern Poland, in a massacre of Jewish people. The first-instance court held the book's claims historically inaccurate and thus ordered the co-editors to issue an apology to Leszczyńska; nevertheless, in *Leszczyńska v Engelking and Grabowski* [2021] Warsaw Court of Appeals, I Civil Division I ACa 300/21, the Warsaw Court of Appeals overturned the decision claiming that sanctioning the respondents would represent an unacceptable violation of the freedom of scientific research and of the freedom of expression. See Jon Henley, 'Fears for Polish Holocaust Research as Historians Ordered to Apologise' *The Guardian* (9 February 2021) <https://www.theguardian.com/world/2021/feb/09/fears-polish-holocaust-research-historians-ordered-apologise> accessed 16 August 2023; The Guardian, 'Polish Appeals Court Overturns Ruling against Holocaust Historians' *The Guardian* (16 August 2021) <https://www.theguardian.com/world/2021/aug/16/polish-appeals-court-overturns-ruling-against-holocaust-historians> accessed 16 August 2023.

[103] Belavusau (n 95).

[104] *Magyarország Alaptörvénye* (Fundamental Law of Hungary) 2011. English translation available at <https://njt.hu/jogszabaly/en/2011-4301-02-00>, accessed 16 August 2023.

increasingly charged, in the following years, with nationalistic references to the history of Hungary as well as to its linkages with the Christian religion.

The 2018 Seventh Amendment to the Fundamental Law, for example, introduced within the Avowal the statement: "We hold that the protection of our identity rooted in our historic constitution is a fundamental obligation of the State".[105] Such an addition, in particular, was adopted in the context of a reform which expressly and declaredly addressed the "mass immigration affecting Europe" and intended to "protect the national sovereignty and to prohibit the settlement of alien populations in Hungary",[106] thus vesting the Government with "a fundamental obligation to protect the constitutional self-identity" of the country.[107] The Fundamental Law also contains a condemnation of the inhuman rights violations committed against Hungary and its citizens by national socialist and communist dictatorships, although it "fails to acknowledge that war crimes against humanity were committed … also between 1920 and 1944 by extreme right-wing 'free troops' and the security forces of the independent Hungarian state … also against other peoples".[108]

Furthermore, and rather consistently with its preamble, the Fundamental Law, in protecting the right to freedom of expression, establishes that such a right may not be exercised with the aim of violating the dignity "of any national, ethnic, racial or religious community", including, first and foremost, the "dignity of the Hungarian nation".[109] This is reflected also within the Hungarian Criminal Code which, while criminalizing incitement to violence and hatred against persons defined by a rather ample range of protected grounds (including disability, gender identity, and sexual orientation), specifically considers as punishable any act of incitement to violence or hatred against the "Hungarian nation".[110] According to the NGO Article 19, this represents a rather controversial aspect of the legislation:

> Although it is possible that one could incite hatred and violence against a majority population, some argue that belonging to the majority Hungarian ethnic group is hardly an element of one's identity which would put one in a vulnerable, threatened position, and which would permit the individual to benefit from increased protection under the criminal law in Hungary. This is problematic also in light of existing trends, by which law enforcement agencies and the courts are more likely to find perpetrators guilty of violent offences if they are committed by members of minority groups against members of the majority

[105] *T/332 számú javaslat. Magyarország Alaptörvényének hetedik módosítása* (Seventh Amendment to the Fundamental Law of Hungary) 2018 art 1. English translation available at <https://helsinki.hu/wp-content/uploads/T332-Constitution-Amendment-29-May-2018-ENG.pdf>, accessed 16 August 2023.

[106] ibid General Reasoning.

[107] ibid Detailed Reasoning art 1.

[108] Gábor Halmai, 'Memory Politics in Hungary: Political Justice without Rule of Law' (*Verfassungsblog*, 10 January 2018) <https://verfassungsblog.de/memory-politics-in-hungary-political-justice-without-rule-of-law/> accessed 16 August 2023. According to Halmai, "the current Hungarian government's attitude towards public discussion of history, similar to that of the Polish one, reflects the position of these illiberal populist regimes towards the rights of their citizens".

[109] Fundamental Law of Hungary art IX, para 5.

[110] See *2012. évi C törvény a büntetőtörvénykönyvről* (Hungarian Criminal Code) 2012 s 332.

> population, rather than the other way round. Guilty verdicts in cases where minorities, such as the Roma or LGBTQI people, are targeted are extremely rare.[111]

The described frameworks on historical memory legislation in Poland and Hungary arguably represent a highly relevant matter for the application and enforcement of common content moderation strategies across the EU and, especially, of the DSA. Indeed, the speech policies described above have been harshly criticized by academic scholarship as well as by activist groups, as they seemingly clash with the fundamental rights and prerogatives of individuals. Clearly, the focal points of such criticisms concern the general safeguard of the right to freedom of expression and information protected under Article 10 ECHR and Article 11 CFREU.

Moreover, the approach followed by Poland and Hungary represents a clear hurdle for a substantive equality-driven fight against the phenomenon of hate speech. Their memory laws being focused on countering the expression of positions that are critical of the past of the nation, and thus of the nation itself, Poland and Hungary tend through such legislation to promote an ethical and historical hierarchy where the majority population is set at the apex. The risk of such a dynamic is, in turn, that of producing a degradation of all other groups, that is, those groups that represent minority demographics. In this respect, the quoted criticism from Article 19 of the Hungarian provision on hate speech clearly confirms the likelihood of such a result.

If, as argued in Chapter 2,[112] the European approach to the governance of hate speech should be driven by the goal of overcoming situations of disparity and domination between certain categories of people over others that have traditionally been subjected to discrimination and marginalization, then the adoption of deeply nationalistic memory laws, potentially quashing and silencing the critical voices of those groups, represents a problematic hurdle. Besides, such risks are overall confirmed and enhanced by the policies adopted in recent years by Poland and Hungary *vis-à-vis* a number of minority demographics. Think, for instance, of the 2021 Hungarian Act "on taking more severe actions against pedophile offenders and amending certain Acts for the protection of children"[113] which, surreptitiously conflating queer people with pedophilia, has limited significantly the representation and portrayal of LGBTQIA+ people across the media[114] and was consequently referred by the Commission to the CJEU the following year.[115]

In the light of such concerns, the EU will have to face a rather crucial challenge when applying the DSA's definition of "illegal content", as allowing Member States wide

---

[111] Article 19, 'Hungary: Responding to "Hate Speech"' (2018) 20 <https://www.article19.org/resources/hungary-responding-hate-speech/> accessed 16 August 2023.

[112] See *supra*, §2.5.

[113] *2021. évi LXXIX. törvény a pedofil bűnelkövetőkkel szembeni szigorúbb fellépésről, valamint a gyermekek védelme érdekében egyes törvények módosításáról* 2021.

[114] Zoltán Kovács, 'Portrayal and Promotion – Hungary's LGBTQI+ Law Explained' *Euractiv* (24 June 2021) <https://www.euractiv.com/section/non-discrimination/news/portrayal-and-promotion-hungarys-latest-anti-lgbt-law-explained/> accessed 16 August 2023.

[115] Krisztina Than, 'Hungary Vows to Fight in EU Court to Defend Anti-LGBT Law' *Reuters* (9 March 2023) <https://www.reuters.com/world/europe/hungary-vows-fight-eu-court-defend-anti-lgbt-law-2023-03-09/> accessed 16 August 2023.

discretion in identifying what is illegal under national law could open the doors to unwarranted consequences and censorial practices. Thus, the clause of Article 3, letter (h), specifying that the notion of "illegal content" includes any content that is not in compliance with the law of a Member State inasmuch as that law is itself "in compliance with Union law", shall likely play a paramount role in this respect, and will require the Commission and EU institutions to carefully evaluate the coherence of domestic legislation with the core principles and goals of the Union, while ensuring, where possible, the respect of the different sensitivities characterizing the various Member States: a task which, however, may well be rather tricky.

4.3. The United Kingdom's Online Safety Act

4.3.1. *Scope of the Act*

4.3.1.1. Material scope of the Act: the debate over the "legal but harmful" provisions and the new "triple shield"

Following the publication in April 2019 of the Online Harms White Paper (OHWP),[116] which promoted, in particular, the establishment by the Government of a statutory duty of care for companies to ensure the safety of Internet users and tackle online harm,[117] the Online Safety Act (OSA) was introduced in the UK House of Commons in March 2022 and eventually enacted in October 2023.[118]

The declared purpose of the OSA is to make the use of regulated Internet services safer for individuals in the UK. To achieve such a purpose, the Act seeks most notably to impose upon providers a range of duties aimed at identifying, mitigating, and managing the risk of harm originating from illegal content and activity, as well as from content and activity that is harmful to children. These duties are focused on ensuring that regulated Internet services are safe by design and are designed and operated in such a way as to guarantee higher standards of protection for adults and children, protect users' fundamental rights to freedom of expression and privacy, and promote transparency and accountability in their provision. On top of these duties, new functions and powers are attributed to the UK's Office of Communications (Ofcom). [119]

In fact, precisely with respect to its scope and objectives, the text of the Act underwent one of its major transformations between December 2022 and January 2023, when the

[116] Department for Digital, Culture, Media and Sport, and the Home Office, 'Online Harms White Paper' (HM Government 2019) CP 57. On the Online Harms White Paper see the dedicated Symposium on vol 11 of the Journal of Media Law (2019) and, in particular, Damian Tambini, 'The Differentiated Duty of Care: A Response to the Online Harms White Paper' (2019) 11 Journal of Media Law 28.

[117] See Lorna Woods, 'The Duty of Care in the Online Harms White Paper' (2019) 11 Journal of Media Law 6.

[118] Online Safety Act 2023. For an overview of the history, as well as of the content, of the Act, see Victoria Nash and Lisa Felton, 'Treating the Symptoms or the Disease? Analysing the UK Online Safety Bill's Approach to Digital Regulation' (SSRN, 2 June 2023) <https://papers.ssrn.com/abstract=4467382> accessed 23 August 2023.

[119] OSA s 1.

House of Commons decided to give heed to wide-ranging criticisms concerning the intention of original draft to include, among other duties, also an obligation for providers to deal with content that is "legal but harmful" to adult users.[120] Indeed, whereas the final text of the OSA only targets content that is illegal, or sets obligations fostering protection of children from content that is specifically harmful for them (e.g., pornography), the original Act proposal, as the OHWP before it,[121] envisaged a framework which required to address also material considered to be *per se* legal under British law but, at the same time, harmful for individual adults and society at large, such as disinformation, cyberbullying, self-harm imagery, or content related to eating disorders.

The "legal but harmful" provisions were nevertheless accused of being a serious hazard for freedom of expression, particularly because of the lack of a clear explanation about the relevant criteria to evaluate the harmfulness of a specific content: according to commentators, not only would this lead to a heavy reliance upon the platforms' own judgments,[122] but it would also open the door to serious governmental interference with freedom of expression. Indeed, in this respect, Trengove and others argued:

> In designing the structure of its regulation, the Bill assigns Ofcom the power to issue Codes of Practice that will determine how … to fulfil the more abstract duties set out in the legislation … the Bill also grants the Minister [of State] the power to interfere with the Codes of Practice by exercising a veto power or by directing Ofcom to align the Codes of Practice with government policy … Our concern is that this creates the possibility of a democratic deficit in the Bill: the Minister retains sweeping powers to interfere with the limits and regulation of speech on the internet's key platforms, with Parliament playing only a minimal negative oversight role. This power is sweeping since the remit of the Bill is wide (particularly in defining the "harmful but legal" content for which services are responsible). This means that the Minister has significant power to interfere with an important set of rights (including free speech and free press) without the particulars of their interventions being vetted by Parliament or subject to public scrutiny.[123]

Due to such criticisms, the House of Commons ultimately chose to amend the text of the OSA and to scrap the "legal but harmful" provisions, while replacing them, in response

---

[120] Jon Porter, 'The UK's Tortured Attempt to Remake the Internet, Explained' *The Verge* (4 May 2023) <https://www.theverge.com/23708180/united-kingdom-online-safety-bill-explainer-legal-pornography-age-checks> accessed 20 August 2023.

[121] Victoria Nash, 'Revise and Resubmit? Reviewing the 2019 Online Harms White Paper' (2019) 11 Journal of Media Law 18, 21–23.

[122] "The meaning of 'content that is harmful to' children and adults is prescribed by clauses 45 and 46 respectively, pursuant to which content is harmful if 'there is a material risk of the content having, or indirectly having, a significant adverse physical or psychological impact on a child [or adult] of ordinary sensibilities'. Clauses 45(7) and 46(6) stipulate that where the platform has knowledge about a particular child or adult at whom relevant content is directed, or who is the subject of it, then the child's or adult's 'characteristics' must be taken into account. Unfortunately, this is the limit of the Bill's explanation of what amounts to legal yet 'harmful' content. It does not account for the fact that how we determine what is harmful will depend on the individual concerned, nor does it define a child or adult of 'ordinary sensibilities' or prescribe the 'characteristics' that that would make them susceptible to harm. As the Bill currently stands, evaluating user content will be entrusted to the subjective judgment of the platform". Peter Coe, 'The Draft Online Safety Bill and the Regulation of Hate Speech: Have We Opened Pandora's Box?' (2022) 14 Journal of Media Law 50, 68–69.

[123] Markus Trengove and others, 'A Critical Review of the Online Safety Bill' (2022) 3 Patterns 100544, 7.

to the opposition's argument that this entailed a major weakening of the Act,[124] with a "triple shield" to protect users. Substantially, the "triple shield" consists of the creation of new criminal offences (and, therefore, of a broadening of the notion of "illegal content" itself), of a strengthened obligation for providers to enforce their own terms and services (especially those concerning children's access), and of the introduction of duties to empower adult users by allowing them more control over the filtering of harmful content they may want to avoid.[125] These three layers will be discussed in the following subsections.

4.3.1.2. Subjective scope of the Act: regulated services

Regulated services under the OSA include, mainly, user-to-user and search services, subject to the majority of the obligations set by the new legislation.[126] "User-to-user services" are defined as Internet services "by means of which content that is generated directly on the service by a user of the service, or uploaded to or shared on the service by a user of the service, may be encountered by another user, or other users, of the service".[127] "Search services", instead, are defined as Internet services that are, or include, a search engine,[128] that is, "a service or functionality which enables a person to search some websites or databases" or "to search (in principle) all websites or databases" and which "does not include a service which enables a person to search just one website or database".[129] Additionally, if a user-to-user service includes a public search engine, it is referred to as a "combined service".[130]

Like the DSA, the OSA also establishes different sets of obligations for providers to comply with depending on their characteristics and, therefore, on the risks they pose to individuals and society. With regard to this aspect, not only are providers of user-to-user services subjected to broader duties than providers of search services, but these two classes are themselves divided into additional categories.

First, the OSA vests the Secretary of State with the task of making regulations setting the thresholds to determine whether a user-to-user service amounts or not to a "Category

---

[124] Toby Helm, 'Labour Pledges to Toughen "Weakened and Gutted" Online Safety Bill' *The Observer* (1 January 2023) <https://www.theguardian.com/technology/2023/jan/01/labour-pledges-toughen-online-safety-bill> accessed 20 August 2023.

[125] Peter Coe, 'Hate Speech, Free Speech and Draft Online Safety Bill' (*Birmingham Law School Research Blog*, 12 December 2022) <https://blog.bham.ac.uk/lawresearch/2022/12/hate-speech-free-speech-and-draft-online-safety-bill/> accessed 20 August 2023. See more *infra*, §4.3.2.

[126] Additionally, specific rules apply to providers of pornographic content. OSA ss 79–82.

[127] ibid 3, subsection 1. Subsection 2 clarifies that, for the purposes of defining a service as a "user-to-user service", it does not matter if a content is actually shared between users as long as a functionality exists allowing for such sharing, nor does it matter what is the proportion of UGC that is present on the service.

[128] ibid 3, subsection 4. Pursuant to subsections 5-7, when an Internet service allowing for the sharing between users of user-generated content features also a search engine, it is nevertheless a user-to-user service unless that user-generated content consists of content of one of the kinds mentioned within subsection 6 (e.g., e-mails, SMS and MMS messages, one-to-one aural communications; comments on provider content; internal business service conditions).

[129] ibid 229, subsection 1.

[130] ibid 4, subsection 7. Subsection 7 also clarifies that a "public search engine" is "a search engine other than one in relation to which the conditions in paragraph 7(2) of Schedule 1 (internal business service conditions) are met".

2B" service, subject to additional transparency requirements,[131] or pertains to the even higher risk tier of "Category 1" services, subjected to the highest degree of duties.[132] The criteria to determine such a categorization are based on the "number of users of the user-to-user part of the service", on the "functionalities of that part of the service" and on "any other characteristics of that part of the service or factors relating to that part of the service that the Secretary of State considers relevant". Similarly, the Secretary of State must also set the thresholds for "Category 2A" services, which include those search services whose search engines fulfil criteria based on the number of their users and on any other characteristic or factor deemed relevant. [133] Once such thresholds are identified, it will be up to Ofcom to keep a register of the providers of Category 1, Category 2A, and Category 2B services.[134]

Additionally, certain duties apply specifically to services, be they user-to-user or search services, that are "likely to be accessed by children",[135] as a way of preventing harm to minors.

4.3.1.3. Territorial scope of the Act

With regard to its territorial scope, the OSA clarifies that it only regulates Internet services that have "links with the United Kingdom",[136] that is, when those services have a "significant number" of UK users or when UK users "form one of the target markets of the service (or the only target market)".[137] Besides, it is not fully clear what is to be considered a "significant number" of UK users, nor how it shall be determined whether the UK represents a "target market" of the service considered.

Additionally, services are considered to have links with the UK if they are capable of being used in the country by individuals and there are "reasonable grounds to believe that there is a material risk of significant harm to individuals in the United Kingdom" presented by user-generated content.[138] Nevertheless, also in this case, the OSA leaves the notion of "material risk of significant harm" open to interpretation. Overall, the current genericness of the provisions concerning the territorial scope of the Act arguably leaves quite an ample leeway for the interpretation of enforcing authorities.

4.3.2. *The new duties for Internet service providers*

The OSA envisages a wide range of duties for providers of regulated Internet services, described analytically within Parts 3-6. Namely, Part 3 is dedicated to the duties of care of providers of user-to-user services and of search services.[139] Part 4 addresses the other

[131] ibid 77–78, Schedule 8.
[132] See more *infra*, §4.3.2.
[133] OSA s 95, Schedule 11, para 1, subparagraphs 1-3.
[134] ibid 95–96.
[135] ibid 37.
[136] ibid 4, subsection 2.
[137] ibid 4, subsection 5.
[138] ibid 4, subsection 6.
[139] ibid 6–63.

duties of the providers of such services, including obligations concerning the verification of users' identities;[140] the reporting of child sexual exploitation and abuse content;[141] and the enforcement of the services' terms of service in a such a way as to ensure the principles of transparency and accountability and the protection of freedom of expression;[142] as well as rules applicable to the case of deceased child users[143] and provisions about transparency reporting.[144] Part 5 contains the duties of providers of pornographic content.[145] Finally, Part 6 concerns the providers' duty to pay fees.[146]

4.3.2.1. Main duties of care

The new duties of care provided for within Part 3 of the Act represent the central content of the legislation, encompassing a wide-ranging set of obligations aimed at countering the presence and spread of content that is illegal or harmful to children. As mentioned above, the original bill also required providers to take measures to remove content that is harmful to adults. Although these provisions were removed, the Act was complemented with new obligations and with the introduction of new criminal offences punishing, for example, false communications[147] and threatening communications,[148] as well as the sending or showing of flashing images (which could cause seizures in individuals with epilepsy),[149] the encouraging of or assisting in the commission of self-harm,[150] and additional online sexual misconducts.[151] Thus, thanks to the adoption of this first tier of the "triple shield",[152] the scope of applicability of the new duties of care also encompasses the prevention and combatting of these novel offences.

All providers of user-to-user services must, first and foremost, undertake suitable and sufficient illegal content risk assessments, taking into account a wide range of elements, including their user base and the possible role of the use of algorithms in making the dissemination of content easier, quicker, and wide-reaching.[153] Second, user-to-user service providers must comply with a number of duties to mitigate and reduce the

[140] ibid 64–65.
[141] ibid 66–70.
[142] ibid 71–74.
[143] ibid 75–76.
[144] ibid 77–78.
[145] ibid 79–82.
[146] ibid 83–90.
[147] ibid 179. The new false communications offence shall be committed, namely, when a person sends a message conveying information they know to be false if, at the time of sending it, they intended the message, or the information in it, to cause "non-trivial psychological or physical harm to a likely audience", and they did not have a reasonable excuse to send that message. Pursuant to s 180, however, the new offence cannot be committed by a recognized news publisher, by the holder of a licence under the Broadcasting Act 1990 or 1996, by the holder of a multiplex licence, by an on-demand programme service, or in connection with the act of showing a film made for cinema to members of the public.
[148] ibid 181.
[149] ibid 183.
[150] ibid 184.
[151] ibid 187–188.
[152] See *supra*, §4.3.1.1.
[153] OSA s 9.

dissemination of that illegal content and to minimize the exposure of individuals to it, also based, clearly, on the last risk assessment made.

Most notably, to this aim, proportionate measures in terms of the design or operation of services must be put in place; the provider must swiftly take down illegal content as soon as it is alerted of its presence; the terms of service must specify the measures adopted to counter the risk of encountering illegal content (including the resort to proactive technologies to ensure compliance) and must be applied consistently; additionally, providers of Category 1 services must also summarize in the terms of service the findings of their latest risk assessments.[154] Specific rules apply for user-to-user service providers that are likely to be accessed by children with respect to content that is harmful to children.[155] Moreover, user-to-user services are required to allow users and affected persons to easily report content that is illegal or harmful to children, and to operate complaints procedures against the provider's lack of compliance with its duties.[156]

As anticipated above, another layer of the "triple shield" – introduced in the Act to compensate the scraping of the "legal but harmful" provisions – is represented by the new rules on "user empowerment".[157] The OSA, indeed, provides that Category 1 services must carry out an assessment of the incidence of and of the likelihood for adults, and notably for adults "with a certain characteristic or who are members of a certain group", to encounter, also due to the functioning of the service's algorithms, the following types of user-generated content: content that encourages, promotes, or provides instructions for suicide or an act of deliberate self-injury, or an eating disorder or behaviours associated with an eating disorder; content that is abusive against persons based on their race, religion, sex, sexual orientation, disability, or gender reassignment; content that incites hatred against people of a particular race, religion, sex or sexual orientation, people who have a disability, or people who have the characteristic of gender reassignment.[158]

Vis-à-vis such materials, providers of Category 1 services must include, to the extent that it is proportionate to do so, features allowing adult users to increase their control over their own exposure to it. These features, in particular, should allow users to reduce the likelihood of encounter or alert of the presence of one of the mentioned types of content. Additionally, similar features should be developed to allow adult users to filter out non-verified users, either by preventing interaction with content generated, uploaded, shared by the latter or by reducing encounters with it.

Providers of search services have to comply with a range of duties of care as well, including, most notably, obligations to assess the risk of and counter illegal content comparable to those established for user-to-user service providers.[159] Also in this case,

[154] ibid 10.
[155] ibid 11–13.
[156] ibid 20–21.
[157] ibid 14–16.
[158] "A person has the characteristic of gender reassignment if the person is proposing to undergo, is undergoing or has undergone a process (or part of a process) for the purpose of reassigning the person's sex by changing physiological or other attributes of sex". ibid 16, subsection 9.
[159] ibid 26–27.

additional duties apply to Category 2A services and to services that are likely to be accessed by children.

Given the possible implications (and collateral effects) connected to the enforcement of the described duties of care, the OSA also provides that due regard should be had notably to the protection and guarantee of the rights to freedom of expression and to privacy. Providers of Category 1 services are subject to more stringent provisions also in this case, as they are required to carry out an assessment of the impact on such fundamental rights of the safety measures or policies they wish to adopt and of those they have already adopted; additionally, they are required to specify in a publicly available statement the positive steps they have taken accordingly.[160]

4.3.2.2. Codes of practice for duties of care

Like the DSA, the OSA foresees the drafting of codes of practice and states that, where a provider of a user-to-user or search service takes or uses the measures described therein, it shall be treated as complying with the duties of care under Part 3 of the Act, including those duties concerning the due regard of freedom of expression and privacy rights. Conversely, when providers seek to comply with the OSA through alternative means other than those set out by the codes of practice, they must have particular regard to the protection of individuals' freedom of expression and to their right to privacy, while Ofcom shall have the power to assess the adequacy of those alternative measures.[161] Additionally, although codes of practice may be brought to court as evidence of the provider's compliance or non-compliance with the law, the OSA clarifies that a failure to act in accordance with a provision of a code is not sufficient of itself to make the provider liable to legal proceedings.[162]

Besides, a rather significant difference between the DSA and the OSA is that, whereas the former states that the Commission and the EBDS shall simply encourage and facilitate the drawing up of voluntary codes of conduct to be elaborated together with all relevant stakeholders,[163] the OSA provides that Ofcom shall have a duty to prepare and issue itself those codes of practice, under the direction of the Secretary of State and the supervision of Parliament, while all other relevant stakeholders will have to be consulted by Ofcom but shall not play an active role in the material drafting of the codes.[164] The OSA thus

[160] ibid 22, 33. These provisions have nevertheless been deemed insufficient by some commentators. Thus, Peter Coe argues that they ultimately give providers of Internet services "a statutory footing to produce boilerplate policies that say they 'have had regard' to free speech or privacy, or 'taken into account' the protection of democratic or journalistic content" since, as long as they "can point to a small number of decisions where moderators have had regard to, or taken these duties into account", they will be able to prove compliance with those rules. Peter Coe, 'Is the New Online Safety Bill Built to Fail?' (*University of Birmingham*, 18 January 2023) <https://www.birmingham.ac.uk/news/2023/is-the-new-online-safety-bill-built-to-fail> accessed 23 August 2023. On top of these provisions, ss 17-19 establish duties for providers of Category 1 services concerning the protection of content of "democratic importance", the protection of news publisher content, and the protection of journalistic content.

[161] OSA s 49.

[162] ibid 50.

[163] DSA art 45. See *supra*, §3.5.3.5.

[164] OSA arts 41–48.

takes a different strategy from the DSA's co-regulatory one and maintains, rather, a top-down approach.

4.3.2.3. Enforcement of Category 1 providers' terms of service

Another highly significant set of provisions is represented by those contained within Chapter 3 of Part 4 of the OSA, concerning rules on the enforcement of providers' terms of service. Above all, the Act requires providers of Category 1 services to ensure, on the one hand, proportionate systems and processes to guarantee that UGC is not taken down, and that users are not restricted in their access to UGC, suspended, or banned, unless this is done in accordance with their terms of service or in compliance with the duties set out in the OSA itself or to avoid civil or criminal liability.[165] On the other hand, those same providers must also ensure that, whenever their terms of service – which must be clear, accessible and detailed – indicate that they will take measures against a particular kind of content (e.g., take down; access restriction; suspension or ban of the user), those terms of service will have to be applied consistently and the measures enforced.[166] This represents, as already mentioned, another layer of the so-called "triple shield".

4.3.3. *Online Safety Act and hate speech*

The framework envisaged by the OSA touches on the governance of hate speech in the UK on at least two levels.

4.3.3.1. Hate speech constituting a criminal offence

First, the duties of care concerning the countering of illegal speech clearly apply to those instances of hate speech that are punishable under the existing criminal legislation. With respect to this level, it has nevertheless been observed that the definition of what "hate speech" is and of what it includes – also with regard to the protected categories considered – is still rather "murky" under British legislation and subject to frequent modifications.[167] Rather than a systemic framework on hate speech proscription, the UK features a plethora of legislative acts differently addressing the phenomenon.[168] As a matter of fact, a 2020

[165] ibid 71. Such a duty, nevertheless, does not apply in relation to consumer content and to terms of service which deal with the treatment of consumer content.

[166] ibid 72, subsections 3-4.

[167] Coe, 'The Draft Online Safety Bill and the Regulation of Hate Speech' (n 122) 68.

[168] As mentioned in Jeremy Waldron, *The Harm in Hate Speech* (Harvard University Press 2012) 204–207, the UK might in fact be among the European countries with the longest-standing tradition of punishing forms of hate speech. Indeed, in *Regina v Osborne* [1732] W Kel 230, 25 Eng Rep 584, a man was convicted for antisemitic libel and for fomenting antisemitic disorders with a view to protecting the public order. Currently, the UK framework on hate speech includes first of all the Public Order Act 1986. Part III of the Act, dedicated to "Racial Hatred", includes a number of provisions criminalizing several acts intended or likely to stir up "hatred against a group of persons … defined by reference to colour, race, nationality (including citizenship) or ethnic or national origins" (s 17). The Racial and Religious Hatred Act 2006 amended the Public Order Act by introducing a Part 3A, containing provisions against hate speech on grounds of religious discrimination. Besides, whereas the Public Order Act makes it a criminal offence to use "threatening, abusive or insulting words or behaviour" (s 18) on grounds of racial discrimination, s 29B

Law Commission consultation on hate crimes and hate speech laws was published investigating their functioning and suggesting possibilities for reform.[169] In light of this, the OSA has been subjected to criticisms with regard to its consistency with the fundamental right to freedom of expression as protected under the ECHR:

> By making online intermediaries responsible for the content on their platforms, the Bill requires them to act as our online social conscience, thereby making them *defacto* gatekeepers to the online world. Although 'privatised censorship' has taken place on platforms such as Facebook and Twitter since their creation, the Bill gives platforms a statutory basis for subjectively evaluating and censoring content. This … could lead to platforms adopting an over-cautious approach to monitoring content by removing anything that may be illegal (including content that they think could be hate speech) or may be harmful, and that would therefore bring them within the scope of the duty and regulatory sanctions. This risk is amplified by the lack of clear definitions of what is hate speech. Such an approach could lead to legitimate content being removed because it is incorrectly thought to be illegal. Cynically, it may also provide platforms with an opportunity, or an excuse, to remove content that does not conform with their ideological values on the basis that it could be illegal.[170]

With respect to this first level, the OSA moves in a direction which is rather similar to the solutions adopted by the DSA, as it aims to broaden providers' due diligence responsibilities against the dissemination of illegal speech, including hate speech. In both cases, this is done through the imposition of asymmetric obligations, that is, through the establishment of different duties for providers of different services based on risk factors. The OSA thus takes a risk-based approach that is rather similar to that of the DSA, an aspect which is confirmed by the choice of the UK Act to impose on providers of all user-to-user and all search services a duty to carry out an assessment of the risks concerning the presence of illegal content upon their infrastructures and to act accordingly. At the same time, however, the OSA appears to take a more top-down perspective with regard to the mitigation measures to be adopted once that risk assessment has been carried out, as the drafting of relevant codes of practice has been retained by public authorities themselves.

Be that as it may, the OSA faces concerns and challenges that are rather similar to those expressed with respect to the DSA:[171] that is, the risk of an unclear scope of application when it comes to hate speech moderation due to the generic reference to "illegal content" calling for the resort to and interpretation of third, often "murky", legislation,

---

only criminalizes the use of "threatening words or behaviour" (s 29B) thus reducing the scope of the proscription: this is confirmed by the Football (Offences) Act 1999, which punishes the repeated uttering of any words or sounds abusive or insulting to a person on the sole grounds listed within s 17 of the Public Order Act 1986. The subsequent Criminal Justice and Immigration Act 2008 broadened Part 3A of the Public Order Act so as to expand the provisions against the stirring of hatred on religious grounds also to the case of stirring hatred on grounds of sexual orientation. Besides, this further addition was accompanied by the clarification that nor "the discussion or criticism of sexual conduct or practices or the urging of persons to refrain from or modify such conduct or practices" nor "any discussion or criticism of marriage which concerns the sex of the parties to marriage" shall "be taken of itself to be threatening or intended to stir up hatred" (s 29JA). In addition to the described framework, legislation affecting the governance of hate speech in the UK includes the following: Crime and Disorder Act 1998; Criminal Justice Act 2003; Malicious Communications Act 1988; Communications Act 2003.

[169] Law Commission, 'Hate Crime Laws: A Consultation Paper' (2020) Law Com CP 250.

[170] Coe, 'Hate Speech, Free Speech and Draft Online Safety Bill' (n 125).

[171] See *supra*, §3.5.4.

and the delegation of content moderation duties, and thus of speech governance, to private profit-driven business actors, with all the (constitutional) issues this entails.[172] As noted by Coe, the mere provision of "due regard" to the fundamental rights of freedom of expression and privacy may not be sufficient to overcome the human rights risks entailed by the Act.

Furthermore, the OSA arguably fails to fully capture the need to embed the principle of equality within the content moderation practices of providers, so as to ensure that no excessive disparity of treatment affects users that are members of minority, discriminated, marginalized, or victimized communities. In fact, the role of equality in the enjoyment of users' fundamental rights is only mentioned by the OSA with regard to the drafting by Ofcom of codes of practice and guidance for transparency reports, as the Act mandates the prior consultation of persons that are considered to have "relevant expertise in equality issues and human rights, in particular (i) the right to freedom of expression set out in Article 10 of the Convention, and (ii) the right to respect for a person's private and family life, home and correspondence set out in Article 8 of the Convention".[173] It is thus desirable that Ofcom will take into due account the opinions of these experts, so as to promote, across regulated providers, good practices that are capable of encouraging the utterance and spread of precious counter-narratives and minority or marginalized voices.

4.3.3.2. "Legal but harmful" hate speech

The second level concerns those utterances that do not amount to illegal content under UK law but are nevertheless to be considered as "legal but harmful" material. These hate speech contents are covered by at least two layers of the triple shield enforceable *vis-à-vis* providers of Category 1 services. First, pursuant to Chapter 3 of Part 4, these providers will have to ensure that they enforce consistently their own terms of service, meaning that any failure to sanction users that violate the providers' private standards concerning hate speech will constitute a failure to comply with the OSA itself. Second, providers of Category 1 services must enable adult users to exert control over their exposure to content that is abusive or incites to hatred against persons based on their race, religion, sex, sexual orientation, disability, or gender reassignment.

The requirement of deploying features to allow adult users to reduce their exposure to "legal but harmful" forms of hate speech represents a particularly innovative strategy to address and counter the negative effects of such a phenomenon. To a certain extent, the provision reminiscent of those rules, enshrined within the DSA, aimed at guaranteeing a higher degree of transparency with respect to the use of recommender systems for content curation purposes and at ensuring a wider margin of choice for users as to the content those automated systems present to them.[174] In the case of the OSA, however, such a strategy appears to be much more targeted, as the Act focuses specifically on allowing users to avoid certain specific types of content.

---

[172] See *supra*, §3.2.2.
[173] OSA ss 41, subsection 6, 78, subsection 2.
[174] See *supra*, §3.5.3.3.

The new OSA duties on user empowerment, nevertheless, arguably represent a double-edged sword in the fight against harmful hate speech. On the one hand, they allow persons – especially members of targeted groups – to avoid encountering content that may be hurtful and trigger significant physical harm and/or psychological distress. On the other hand, such a strategy does not *per se* help reduce the spread of hate content across those demographics that are more susceptible to the incitement to hatred, violence, or discrimination. The user empowerment tools, the practical functioning of which is additionally yet to be defined, may in other words guarantee that individuals that already reject the premises, ideas, and arguments embedding hate speech can avoid entering into contact with it; however, they are seemingly not at all helpful in addressing the mounting extremization of hate groups. Thus, the result could well be that of enhancing, rather than reducing, the polarization of public discourse.

4.4. The United States

4.4.1. *United States' tolerance towards the "thought we hate"*

The approach of the US towards hate speech governance and its relationship with free speech have already been discussed in Chapter 2.[175] In the US mindset and SCOTUS case law, the adoption of measures against hate speech generally represents an impermissible breach of the Constitution, as it translates into a form of viewpoint discrimination that is at odds with the First Amendment. In discussing the status of free speech in the constitutional mindset and jurisprudence of the US, Alexander Meiklejohn expressed over seventy years ago the following thoughts:

> We Americans, in choosing our form of government, have made, at this point, a momentous decision. We have decided to be self-governed. We have measured the dangers and the values of the suppression of the freedom of public inquiry and debate. And, on the basis of that measurement, having regard for the public safety, we have decided that the destruction of freedom is always unwise, that freedom is always expedient … It is a reasoned and sober judgment as to the best available method of guarding the public safety. We, the People, as we plan for the general welfare, do not choose to be "protected" from the "search for truth". On the contrary, we have adopted it as our "way of life", our method of doing the work of governing for which, as citizens, we are responsible. Shall we, then, as practitioners of freedom, listen to ideas which, being opposed to our own, might destroy confidence in our form of government? Shall we give a hearing to those who hate and despise freedom, to those who, if they had the power, would destroy our institutions? Certainly, yes! Our action must be guided, not by their principles, but by ours. We listen, not because they desire to speak, but because we need to hear.[176]

The model of the "tolerant society" and the idea that the "search for truth" should not be constricted by governmental interference represent the *fil rouge* of SCOTUS case law on hate speech. Therefore, as opposed to the European framework, US constitutional law tends to grant hate speech the status of a legitimate form of expression. Quite notoriously,

[175] See *supra*, §§2.2.1, 2.3.1.

[176] Alexander Meiklejohn, *Free Speech And Its Relation to Self-Government* (1st edn, Harper & Brothers 1948) 65–66.

in *Matal v Tam*, the SCOTUS argued that "speech that demeans on the basis of race, ethnicity, gender, religion, age disability, or any other similar ground is hateful; but the proudest boast of our free speech jurisprudence is that we protect the freedom to express 'the thought that we hate'".[177] Consequently, limitations and restrictions upon these utterances tend to be considered admissible only in those rather rare cases when they pertain to categories of "low-value speech" – notably, "true threats" or "fighting words".[178]

The highly tolerant and liberal approach to speech characterizing the US and its First Amendment jurisprudence is also reflected within its legislation and case law on intermediary liability. The following subsections will give an overview of such a framework focusing, first of all, on the notorious Section 230 of the Communications Decency Act, which inspired, after all, the EU's ECD at the turn of the millennium.[179]

4.4.2. *Intermediary liability in the US and the rise of Section 230*

When Congress passed Section 230 CDA in 1996, it sought primarily to promote the development of the Internet and of interactive computer services and media, to preserve the "vibrant and competitive free market" characterizing those emerging technologies, and to encourage the development of new technologies to maximize user control over the information received through the Internet.[180] In fact, Section 230 was in good part a response to some judgments that had brought about the issue of the legal regime applicable to ISPs,[181] namely, the decision rendered by the US District Court for the Southern

[177] *Matal v Tam* 582 US __ (2017) 25. Another emblematic case is, in this respect, that of *Snyder v Phelps*, where the plaintiff, Mr Snyder, claimed compensation for the damages inflicted upon him by the respondent, leader of a Baptist congregation, who had organized demonstrations during the funeral of Snyder's son, a twenty-year-old marine who had died in Iraq. Phelps' congregation had, as a matter of fact, begun demonstrating on occasion of a number of funerals of members of the US army, arguing most notably that such deaths were caused by the army's "moral turpitude" and for its supposed tolerance towards gay people. The demonstrators' placards exhibited slogans such as: "They turned America over to f***; they're coming home in body bags"; "Thank God for dead soldiers"; "God hates f***"; "You're going to hell"; "God hates you". The SCOTUS ultimately rejected Snyder's claims for compensation, affirming: "While these messages may fall short of refined social or political commentary, the issues they highlight – the political and moral conduct of the United States and its citizens, the fate of our nation, homosexuality and the military, and scandals involving the Catholic clergy – are matters of public import … Speech is powerful. It can stir people to action, move them to tears of both joy and sorrow and – as it did here – inflict great pain. On the facts before us, we cannot react to that pain by punishing the speaker. As a Nation we have chosen a different course – to protect even hurtful speech on public issues to ensure that we do not stifle public debate". *Snyder v Phelps* 562 US 443 (2011) 454, 460–461.

[178] See, *ex multis*, *Brandenburg v Ohio* 395 US 444 (1969); *Village of Skokie v Nat'l Socialist Party of America* 373 NE2d 21 (Ill 1978); *RAV v City of St Paul* 505 US 377 (1992); *Virginia v Black* 538 US 343 (2003). See *supra*, §2.2.1. In an *obiter dictum* of its judgment concerning the legitimacy of sanctioning of a flower shop owner who had refused to arrange flowers for a same-sex wedding, the Supreme Court of the state of Washington corroborated this interpretation of federal case law by stating that "the First Amendment protects even hate speech, provided it is not 'fighting words' or a 'true threat'". *State v Arlene's Flowers, Inc* 441 P3d 1203 (Wash 2019) 1225.

[179] See *supra*, §3.4.1.

[180] Communications Decency Act 1996 ss 230, subsection b, nn 1–3.

[181] Danielle Keats Citron and Benjamin Wittes, 'The Internet Will Not Break: Denying Bad Samaritans Sec. 230 Immunity' (2017) 86 Fordham Law Review 401, 404–406; Jeff Kosseff, *The Twenty-Six Words That Created the Internet* (Cornell University Press 2019) 36–56; Mary Anne Franks, 'How the Internet Unmakes Law Distinguished Lecture Series on the State of Internet Law' (2020) 16 Ohio State Technology Law Journal 10, 17–18.

District of New York in the case of *Cubby v CompuServe*[182] and that rendered by the New York Supreme Court in the case of *Stratton Oakmont v Prodigy Services*.[183]

Cubby v CompuServe dealt with an episode of online defamation. CompuServe, together with Prodigy and America Online, was one of the major commercial online service providers in the US, offering its clients access to the Internet and, notably, to the content posted and shared by other subscribers to the service. Cubby, a company that developed products and services for the computer industry, launched in 1990 a newsletter on the broadcast industry, called "Skuttlebut", which was distributed via fax machine. However, a similar newsletter, "Rumorville", was also being distributed through CompuServe's infrastructure, to which one of Cubby's founders, Mr Blanchard, was subscribed. Soon, Rumorville began posting content accusing Skuttlebut of stealing news items, as well as articles against Blanchard. Blanchard ultimately brought action for defamation against both Rumorville and CompuServe, the latter being allegedly in a position comparable to that of a publisher, and thus subject to editorial responsibility. Judge Leisure, vested with the duty to decide for the first time on a case concerning the liability of an ISP for third-party content, eventually ruled in favour of CompuServe, arguing that CompuServe did not act as an editor but, rather, as a mere distributor, so that, following the SCOTUS' precedent of *Smith v California*,[184] it could not be held liable for Rumorville's defamatory content. However, such a decision inherently implied that, according to Judge Leisure, a court should in similar circumstances always assess whether the ISP had in fact acted as a distributor or, rather, as an editor. Only in the latter case could an ISP could be considered to be liable for third-party content without infringing the First Amendment.

Indeed, this conclusion would later be confirmed by *Stratton Oakmont v Prodigy*. Prodigy was, like CompuServe, an early online service provider that allowed for the exchange of information between users. Prodigy featured, namely, an online bulletin board, Money Talk, upon which serious allegations of fraud had been posted by a pseudonymous user against a securities firm, Stratton Oakmont, and its president Mr Porush. The greatest (and widely marketed) difference between the service offered by CompuServe and that offered by Prodigy, however, was that the latter had set standards for third-party content and scanned through the notes that went on bulletin boards to remove material that violated those standards. Notably, Prodigy employed an automated software to filter out profanity. Although the service provider argued that it was not possible for it to edit out (manually) all posts and messages uploaded by its users, Justice Ain concluded that the ISP was in fact acting as an editor and, therefore, should be accountable for editorial responsibility. In other words, the choice of Prodigy to remove questionable content had cost it the defamation lawsuit.

[182] *Cubby, Inc v CompuServe, Inc* 776 FSupp 135 (SDNY 1991).

[183] *Stratton Oakmont, Inc v Prodigy Services Co* 23 Media L Rep 1794 (NY Sup Ct 1995).

[184] *Smith v California* 361 US 147 (1959). The case concerned the conviction, under a local statute, of the owner of a bookstore, Mr Smith, for the mere owning of a book judicially determined to be obscene, even though Smith had been unaware of the book's content. The SCOTUS determined that such a conviction was in breach of the Fourteenth Amendment.

The outcomes of *Cubby v CompuServe* and *Stratton Oakmont v Prodigy* soon triggered the reaction of the Congress. Proponents of the CDA, indeed, were unsatisfied with the courts' findings, which could impair the development of the Internet and, at the same time, discourage ISPs from taking action against "indecent" materials. As a result, the CDA established that "no provider or user of an interactive computer service shall be treated as the publisher or speaker of any information content provider",[185] thus shielding ISPs from any form of liability that may arise from third-party content. At the same time, through its "Good Samaritan clause", Section 230 establishes that ISPs shall not incur in civil liability on account of "any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected".[186] In other words, pursuant to Section 230, providers are not liable for any content published online by users but may nevertheless choose to intervene to remove content they deem inappropriate without this implicating that they are acting as editors and thus without risking incurring editorial liability.

In fact, the precise extent to which ISPs would be shielded from liability under Section 230 was not initially fully clear. In this respect, a seminal decision was the one rendered in 1997 by the Court of Appeals for the Fourth Circuit in the case of *Zeran v America Online*,[187] defined as "probably the most important ruling in Internet Law".[188]

A false announcement had been published upon a bulletin of America Online concerning the sale of t-shirts that were offensive towards the victims of a recent domestic terrorist attack in Oklahoma City. The announcement reported the name and telephone number of Mr Zeran who, in the following weeks, was flooded with telephone calls and death threats. Although Zeran had repeatedly informed America Online (AOL) of such defamatory content, the ISP had failed to effectively remove it, so that Zeran ultimately filed a lawsuit against it, arguing that AOL was liable for negligence. Zeran's attorney acknowledged that Section 230 prohibited judges from treating providers of interactive computer services as publishers, thus excluding the possibility to hold them liable for the presence of illegal content they were unaware of, however, it did not prohibit treating them as distributors, as had been done in the case of *Cubby v CompuServe*. That being the case, consistently with *Cubby*, an ISP could still be considered liable for tort for third-party content as soon as it was made aware of the presence of such content.

The Court of Appeals for the Fourth Circuit, nevertheless, confirmed the first-instance decision of the District Court for the Eastern District of Virginia to dismiss the lawsuit, arguing as follows:

> Zeran simply attaches too much importance to the presence of the distinct notice element in distributor liability. The simple fact of notice surely cannot transform one from an

[185] CDA ss 230, subsection c, n 1.

[186] ibid 230, subsection c, n 2, lett A.

[187] *Zeran v America Online, Inc* 129 F3d 327 (4th Cir 1997). See, for an overview of the history of the case, Kosseff (n 181) 79–97.

[188] Eric Goldman, 'The Ten Most Important Section 230 Rulings' (2017) 20 Tulane Journal of Technology and Intellectual Property 1, 3.

> original publisher to a distributor in the eyes of the law. To the contrary, once a computer service provider receives notice of a potentially defamatory posting, it is thrust into the role of a traditional publisher. The computer service provider must decide whether to publish, edit, or withdraw the posting. In this respect, Zeran seeks to impose liability on AOL for assuming the role for which § 230 specifically proscribes liability – the publisher role … If the original party is considered a publisher of the offensive messages, Zeran certainly cannot attach liability to AOL under the same theory without conceding that AOL too must be treated as a publisher of the statements … [I]nterpreting § 230 to leave distributor liability in effect would defeat the two primary purposes of the statute and would certainly "lessen the scope plainly intended" by Congress' use of the term "publisher".[189]

As the first appellate federal judgement applying the new provision, *Zeran* represented a landmark precedent in the following years. Thus interpreted, Section 230 introduced a legal discipline extremely favourable to ISPs, opening the doors to the evolution of the digital environment. In this respect, Eric Goldman has declared that, when it comes to the Internet, Section 230 is even "better than the First Amendment".[190]

4.4.3. *Private moderation and the state action doctrine*

The regime established by Section 230 CDA in favour of providers of interactive computer services is not counterbalanced in the US by proper duties to guarantee the protection of users' free speech. Indeed, a principle such as that of the German *Drittwirkung*, imposing horizontal duties upon private actors to respect others' constitutional rights, is barred in the US by the state action doctrine, according to which those rights may only be invoked against state actors.[191]

Besides, the definition of the exact boundaries between what is state action and what is private action – and therefore the precise definition of what and who is, in fact, a "state actor" – can represent a tricky question especially in those cases "where private parties are at least arguably imbued with governmental power, and have allegedly abused that power".[192] In its landmark judgment of *Marsh v Alabama*, for instance, the SCOTUS concluded that the First Amendment barred the private owners of a company town from prohibiting the appellant, a Jehova's Witness, to distribute religious literature on the town's sidewalk, in application of an Alabaman statute. Indeed, the SCOTUS affirmed that "whether a corporation or a municipality owns or possesses the town, the public in either case has an identical interest in the functioning of the community in such a manner that the channels of communication remain free".[193]

---

[189] *Zeran v America Online, Inc* (n 187) 332–334.

[190] Eric Goldman, 'Why Section 230 Is Better than the First Amendment' (2019) 95 Notre Dame Law Review Reflection 33.

[191] See, among others, Oreste Pollicino, 'The Quadrangular Shape of the Geometry of Digital Power(s) and the Move towards a Procedural Digital Constitutionalism' (2023) 29 European Law Journal 10, 22–24.

[192] Wilson R Huhn, 'The State Action Doctrine and the Principle of Democratic Choice' (2006) 34 Hofstra Law Review 1379, 1389.

[193] *Marsh v State of Alabama* 326 US 501 (1946) 507.

As noted by Gardbaum,[194] the SCOTUS has identified some tests to draw a line between private and state actors. For instance, the "public function" test, which has its roots precisely in *Marsh v Alabama*, stipulates that a private actor's actions may be deemed state action for constitutional purposes when it exercises functions that are traditionally reserved to the state. In other cases, the existence of a "nexus" between the state and the private actor may be investigated, with a view to understanding whether the former is significantly entangled with, or is participating jointly in, the actions of the latter. An additional test refers to the possibility of inquiring whether the state has provided the private actor with "such significant encouragement, either overt or covert, that the choice must in law be deemed to be that of the state".[195] Finally, the enforcement, through court orders, of certain voluntary private actions has in some cases (but less consistently) been deemed to be subject to constitutional scrutiny (e.g., court orders enforcing racially restrictive covenants between homeowners).[196]

The paramount role of the Internet and of social media with respect to the full enjoyment of expression, communication, and information liberties in the digital age could serve as a reasonable basis for the argument that providers of such intermediary services today do, in fact, exert a "public function" and may therefore be considered state actors responsible of guaranteeing the protection of those rights. Such an argument was brought up in the case of *Prager University v Google*.[197] Prager University (PragerU) is a non-profit advocacy group and media organization promoting conservative propaganda who contested that YouTube – and, therefore, its parent company Google – had infringed PragerU's First Amendment rights by reducing the visibility of and demonetizing several of its videos.

PragerU, in particular, invoked as a relevant precedent the case of *Packingham v North Carolina*, where the SCOTUS had struck down a local statute banning registered sex offenders from creating any social media accounts as a way to prevent children from entering into contact with them. On that occasion, the majority of the Court, in an opinion authored by Justice Kennedy, had held that the prohibition, sanctioned under criminal law, was not narrowly tailored to accomplish its goals and thus violated the First Amendment by prohibiting *in toto* access to websites which represent, today, the "principal sources for knowing current events, checking ads for employment, speaking and listening in the modern public square, and otherwise exploring the vast realms of human thought and knowledge".[198] The Court of Appeals for the Ninth Circuit rejected PragerU's

---

[194] Stephen Gardbaum, 'The "Horizontal Effect" of Constitutional Rights' (2003) 102 Michigan Law Review 387, 412–414.

[195] *Blum v Yaretsky* 457 US 991 (1982) 1004.

[196] See, similarly, Cheong: "To determine whether a party is a state actor, courts have developed four discernible tests: (1) the existence of a *symbiotic relationship* between the private actor and the state, (2) the state *commanding or encouraging* private discriminatory action, (3) the private party performing a *traditionally public function*, and (4) *the involvement of a governmental authority* in the unlawful conduct". Inyoung Cheong, 'Freedom of Algorithmic Expression' (2023) 91 University of Cincinnati Law Review 680, 684.

[197] *Prager University v Google LLC* 951 F3d 991 (9th Cir 2020).

[198] *Packingham v North Carolina* 582 US __ (2017) 8.

argument that *Packingham* represented a relevant precedent, as in that case the actor abridging Packingham's First Amendment rights was, in fact, a state actor – that is, the state of North Carolina.

Instead, the Court relied on the precedent of *Manhattan v Halleck*,[199] where the SCOTUS had dealt with the case of a non-profit corporation (operating public access channels in the state of New York) that had chosen to suspend some cable television show producers from the provision of its services and facilities due to the content of a programme submitted for airing. Having excluded that the respondent was a state actor,[200] the SCOTUS had concluded that no infringement of the First Amendment had taken place. Following this precedent, the Court of Appeals for the Ninth Circuit held in the case of *PragerU* that Google could not be considered as exercising a public function and, therefore, did not have any First Amendment obligations towards PragerU. In particular, the Court argued:

> It is true that a private entity may be deemed a state actor when it conducts a public function, but the relevant function "must be both traditionally and exclusively governmental" … This test is difficult to meet. It is "not enough" that the relevant function is something that a government has "exercised … in the past, or still does" or "that the function serves the public good or the public interest in some way". Rather, the relevant function must have been "traditionally the *exclusive* prerogative of the [s]tate" … The relevant function performed by YouTube – hosting speech on a private platform – is hardly "an activity that only governmental entities have traditionally performed" … Private parties like "[g]rocery stores" and "[c]omedy clubs" have "open[ed] their property for speech" … YouTube does not perform a public function by inviting public discourse on its property.[201]

4.4.4. *The Untouchables? Critics and recent developments on the interplay between Section 230, state action doctrine, and the First Amendment*

The choice to reject a categorization of online private platforms as state actors, coupled with the immunity established under Section 230, implies that providers of intermediary services are, under US law, protected by a layer of untouchability as regards their content moderation (and curation) practices. On the one hand, as a general rule,[202] providers are not required to put in place measures to counter the presence of illegal (or harmful) content on their infrastructures. On the other hand, they are not subject to any duties to ensure the protection of the First Amendment rights of the recipients of their services.[203] In other

---

[199] *Manhattan Community Access Corp v Halleck* 587 US __ (2019).

[200] "Merely hosting speech by others is not a traditional, exclusive public function and does not alone transform private entities into state actors subject to First Amendment constraints". ibid 10.

[201] *Prager U v Google* (n 197) 997–998.

[202] In fact, sectoral legislation may provide otherwise. For example, this is the case, as is well known, of the US DMCA which establishes a notice and take down regime with respect to copyright infringement.

[203] A rather different situation concerns the case where a public figure, e.g., the President of the United States, blocks accounts and/or removes content posted by users on their own personal wall. In these cases, because those measures have been adopted directly by a state figure, a violation of the First Amendment has been found to take place. The leading case is, in this respect, *Knight First Amendment Institute at Columbia v Trump* 928 F3d 226 (2nd Cir 2019), which held that the President's account on social media

words, providers do not risk facing liability or accountability for the content they do not remove, nor for the content they do remove.

This two-fold immunity is particularly evident and, to a certain extent, striking when it comes to hate speech governance. Clearly, ISPs are not compelled to sanction the upload of hate speech content, not only because of Section 230 but also because that content is *per se* considered to be free speech protected under the First Amendment. However, providers are nevertheless allowed to ban hate speech and take actions against it under their own standards and terms of service and have, in fact, proven rather often to be willing to do so.[204] As a result, the choice to counter or not the phenomenon of hate speech on the Internet is left to the complete discretion of private platforms, with no scrutiny whatsoever from the state as regards the respect of constitutional principles, such as the right to freedom of expression or the protection of equality. While hate speech is, as such, protected by the First Amendment and its removal is not mandated to ISPs by the law, there is no countervailing right for users whose content has been erroneously categorized as hate speech to have that content reinstated or any other measures be removed. This could, clearly, have rather significant effects in terms of the promotion of the right to equality of victimized groups, especially if the automated systems for hate speech detection are vitiated by unwarranted biases.[205]

---

represents a public forum for free speech; *Biden v Knight First Amendment Institute at Columbia* 593 US __ (2021) subsequently vacated the judgment of the Court of Appeals for the Second Circuit due to the shift from Donald Trump's to Joe Biden's presidency. See also *Davison v Randall* 912 F3d 666 (4th Cir 2019). On this point, see, among others, Marco Bassini, 'Social Networks as New Public Forums? Enforcing the Rule of Law in the Digital Environment' (2022) 1 The Italian Review of International and Comparative Law 311, 323–325. On 31 October 2023, the question whether officials can block critics on their social media accounts was heard by the SCOTUS: see Adam Liptak, 'Supreme Court to Decide Whether Officials Can Block Critics on Social Media' *The New York Times* (24 April 2023) <https://www.nytimes.com/2023/04/24/us/elected-officials-social-media-supreme-court.html> accessed 18 September 2023; Adam Liptak, 'Biden Asks Supreme Court to Lift Limits on Contacts With Social Media Sites' *The New York Times* (14 September 2023) <https://www.nytimes.com/2023/09/14/us/politics/supreme-court-social-media-misinformation.html> accessed 18 September 2023; John Kruzel and Andrew Chung, 'US Supreme Court Weighs If Public Officials Can Block Critics on Social Media' *Reuters* (31 October 2023) <https://www.reuters.com/legal/us-supreme-court-decide-if-public-officials-can-block-critics-social-media-2023-10-31/> accessed 27 December 2023. Eventually, on 15 March 2024, the SCOTUS concluded that a public official, posting content about topics concerning their work, may be held liable under the First Amendment for having blocked comments from their critics only inasmuch as they have the power to speak on behalf of the state and are, in that instance, exercising that power. Indeed, the SCOTUS underscored that when "the public employee does not use his speech in furtherance of his official responsibilities, he is speaking in his own voice", meaning that that official retains their own First Amendment rights, including that of blocking other users' comments: see *Lindke v Freed* 601 US 187 (2024) 201.

[204] Richard Wilson and Molly Land, 'Hate Speech on Social Media: Content Moderation in Context' (2021) 52 Connecticut Law Review 1029. See *infra*, §5.2.

[205] As highlighted by Citron and Franks, "some of the most serious threats to free speech come not from the government, but from non-state actors. Marginalized groups in particular, including women and racial, minorities, have long battled with private censorial forces as well as governmental ones. But the unregulated internet – or rather, the selectively regulated internet – is exacerbating, not ameliorating, this problem. The current state of Section 230 may ensure free speech for the privileged few; protecting free speech for all requires reform … the internet has rolled back many gains made for racial and gender equality. The anonymity, amplification, and aggregation possibilities offered by the internet have allowed private actors to discriminate, harass, and threaten vulnerable groups on a massive scale … the internet has been used to further chill the intimate, artistic, and professional expression of individuals whose rights were already

As a matter of fact, over twenty years after the adoption of Section 230 and the landmark judgment of *Zeran*, the "power without responsibility"[206] enjoyed by providers has often been called into question both by the left and the right: namely, conservatives "claim that Section 230 gives tech companies a license to silence speech based on viewpoint", whereas liberals "criticize Section 230 for giving platforms the freedom to profit from harmful speech and conduct".[207]

4.4.4.1. The strange case of Texas' HB 20 and Florida's SB 7072

With respect to the first line of criticisms, that is, the argument that Section 230 gives online platforms the right to censor speech on the basis of viewpoint discrimination, the question has arisen whether the introduction of state laws restricting such a liberty of platforms to freely moderate content as they wish would be barred by a third-layer shield, on top of Section 230 and of the state action doctrine, that is, the First Amendment. In other words, the question concerns the possibility of comparing any content moderation or content curation choice of a platform to free speech and, therefore, the recognition of a First Amendment "editorial" right to remove unwarranted content and the consistency of such a right with Section 230.

As a matter of fact, Florida and Texas, both led by conservative governments, have enacted legislation aimed at reducing the extensive liberty of platforms in moderating content.[208] Namely, their goal is to prevent providers from applying measures that are considered to be viewpoint discriminatory, following conservatives' general belief that content moderation tends to be biased against their own views and ideas – a belief which was further intensified following the deplatforming of former US President Donald Trump from Twitter (today X) and other social media in the aftermath of the assault on Capitol Hill of 6 January 2021. Texas' HB 20,[209] on top of introducing transparency and due process obligations, including a complaint-and-appeal system for users,[210] prohibits social media platforms from

---

under assault offline". Danielle Keats Citron and Mary Anne Franks, 'The Internet as a Speech Machine and Other Myths Confounding Section 230 Reform What's the Harm? The Future of the First Amendment' (2020) 2020 University of Chicago Legal Forum 45, 67–68.

[206] Rebecca Tushnet, 'Power Without Responsibility: Intermediaries and the First Amendment' (2008) 76 The George Washington Law Review 986.

[207] Citron and Franks (n 205) 46–47.

[208] See, among others, Ioanna Tourkochoriti, 'The Digital Services Act and the EU as the Global Regulator of the Internet' (2023) 24 Chicago Journal of International Law 129, 144–146; Cheong (n 196) 691–693; Enrico Andreoli, 'Continuities and Discontinuities. First Amendment and Digital Free Speech in US Constitutionalism' (2023) 56 DPCE Online 261, 275–283. At the federal level, some proposals have been presented for the adoption of a Disincentivizing Internet Service Censorship of Online Users and Restrictions on Speech and Expression Act (DISCOURSE Act) and of a 21st Century FREE Speech Act by members of the 117th Congress, although such attempts appear to have been unsuccessful: see DISCOURSE Act, S. 2228, 117th Cong. (2021); 21st Century FREE Speech Act, S. 1384, 117th Cong. (2021). See, with respect to the federal proposals, Dawn Carla Nunziato, 'The Digital Services Act and the Brussels Effect on Platform Content Moderation' (2023) 24 Chicago Journal of International Law 115, 124–125.

[209] Tex HB 20 relating to censorship of or certain other interference with digital expression, including expression on social media platforms or through electronic mail messages 2021.

[210] ibid 2.

> censor[ing] a user, a user's expression, or a user's ability to receive the expression of another person based on:
> (1) the viewpoint of the user or another person;
> (2) the viewpoint represented in the user's expression or another person's expression; or
> (3) a user's geographic location in [Texas] or any part of [Texas].[211]

Florida's SB 7072[212] also introduces some content moderation restrictions, prohibiting for instance social media platforms from deplatforming any known candidate for office,[213] as well as from using post-prioritization or shadowbanning algorithms affecting content by or about that candidate.[214] Additionally, the law includes significant disclosure obligations, notably the duty to "publish the standards, including detailed definitions, it uses or has used for determining how to censor, deplatform, and shadow ban",[215] and user-data requirements for deplatformed users, namely the right "to access or retrieve" all of their "information, content, material, and data for at least 60 days after the user receives the notice".[216] Both legislations have nevertheless been subjected to constitutional scrutiny with respect to their actual consistency with the First Amendment, with inconsistent results.

At the beginning of December 2021, in *NetChoice v Paxton*,[217] the District Court for the Western District of Texas first issued a preliminary injunction against HB 20 that enjoined the enforcement of the mentioned new duties upon social media platforms, arguing that social media platforms "curate both users and content to convey a message about the type of community the platform seeks to foster and, as such, exercise editorial discretion over their platform's content",[218] so that HB 20's prohibition of viewpoint-based moderation – entailing, for instance, even the prohibition to remove content such as speech promoting Nazism – ultimately "restricts social media platforms' First Amendment right to engage in expression when they disagree with or object to content" and the threat of lawsuits under that law "chills the social media platforms' speech rights".[219]

On 11 May 2022, the Court of Appeals for the Fifth Circuit granted the state of Texas' motion to stay the preliminary injunction pending appeal,[220] whereas on 31 May of the same year the SCOTUS vacated the Fifth Circuit's order, thus reinstating the District Court's preliminary injunction, with the contrary vote of Justices Kagan, Alito, Thomas, and Gorsuch. Notably, in his dissenting opinion joined by Justices Thomas and Gorsuch, Justice Alito argued that the case concerned a "ground-breaking Texas law" addressing "the power of dominant social media corporations to shape public discussion of the important issues of the day"[221] and that, because of the novel legal questions the law arose

[211] ibid 7.
[212] Fla SB 7072 on Social Media Platforms 2021.
[213] ibid 2.
[214] ibid 4.
[215] ibid.
[216] ibid. See Andreoli (n 208) 280.
[217] *NetChoice, LLC v Paxton* 573 FSupp3d 1092 (WDTex 2021).
[218] ibid 1108.
[219] ibid 1109–1110.
[220] *NetChoice, LLC v Paxton* 2022 WL 1537249 (5th Cir 2022).
[221] *NetChoice, LLC v Paxton* 596 US __ (2022) 1.

and the uncertain applicability of precedent case law, he did not feel "comfortable intervening at th[at] point in the proceedings".[222]

On 16 September 2022, however, the Court of Appeals for the Fifth Circuit finally issued its ruling, reversing the District Court's injunction for abuse of discretion and remanding for further proceedings, as it concluded that Texas' HB 20 did not, in fact, violate of the First Amendment. Quite solemnly, Circuit Judge Oldham declared in his opinion: "Today we reject the idea that corporations have a freewheeling First Amendment right to censor what people say".[223] *Inter alia*, the Court of Appeals for the Fifth Circuit argued that HB 20 does not chill speech but, rather, censorship, so that the prohibitions it introduced would "cultivate rather than stifle the marketplace of ideas".[224] Additionally, the Court thus investigated the relationship between HB 20 and Section 230:

> Section 230 provides that the Platforms "shall [not] be treated as the publisher or speaker" of content developed by other users … Section 230 reflects Congress's judgment that the Platforms do not operate like traditional publishers and are not "speak[ing]" when they host user-submitted content … Section 230 undercuts both of the Platforms' arguments for holding that their censorship of users is protected speech. Recall that they rely on two key arguments: first, they suggest the user-submitted content they host is *their speech*; and second, they argue they are *publishers* akin to a newspaper. Section 230, however, instructs courts *not* to treat the Platforms as "the publisher or speaker" of the user-submitted content they host … In sum, § 230 reflects Congress's judgment that the Platforms are not acting as speakers or publishers when they host user-submitted content. While a statute may not abrogate constitutional rights, Congress's factual judgment about the role of online platforms counsels against finding that the Platforms "publish" (and hence speak) the content that other users post. And that's particularly true here, because the Platforms have long relied on and vigorously defended that judgment – only to make a stark about-face for this litigation. Section 230 thus reinforces our conclusion that the Platforms' censorship is not protected speech under the First Amendment.[225]

In the meantime, on 30 June 2021, in *NetChoice v Moody*,[226] the District Court for the Northern District of Florida had issued a preliminary injunction enjoining enforcement of Florida's SB 7072, similar to that issued by the District Court for the Western District of Texas in *NetChoice v Paxton*. This time, however, the Court of Appeals for the Eleventh Circuit affirmed on 23 May 2022 most of the District Court's decision, vacating it only partially with respect to a number of provisions (e.g., that on user-data access rights) that it held, in fact, to be constitutionally legitimate.[227] In this respect, the reasoning of the Court of Appeals for the Eleventh Circuit with regard to the possibility of recognizing online platforms' content moderation and curation activities as a form of free speech protected by the First Amendment is seemingly strikingly different from that rendered by the Court of Appeals for the Fifth Circuit in *NetChoice v Paxton*:

> Social-media platforms like Facebook, Twitter, YouTube, and TikTok are private companies with First Amendment rights … and when they (like other entities) "disclos[e],"

[222] ibid 5–6.
[223] *NetChoice, LLC v Paxton* 49 F4th 439 (5th Cir 2022) 445.
[224] ibid 450.
[225] ibid 465–466, 468.
[226] *NetChoice, LLC v Moody* 546 FSupp3d 1082 (NDFla 2021).
[227] *NetChoice, LLC v Moody* 34 F4th 1196 (11th Cir 2022) 1231–1232.

> "publish[ ]," or "disseminat[e]" information, they engage in "speech within the meaning of the First Amendment," … More particularly, when a platform removes or deprioritizes a user or post, it makes a judgment about whether and to what extent it will publish information to its users – a judgment rooted in the platform's own views about the sorts of content and viewpoints that are valuable and appropriate for dissemination on its site. As the officials who sponsored and signed S.B. 7072 recognized when alleging that "Big Tech" companies harbor a "leftist" bias against "conservative" perspectives, the companies that operate social-media platforms express themselves (for better or worse) through their content-moderation decisions. When a platform selectively removes what it perceives to be incendiary political rhetoric, pornographic content, or public-health misinformation, it conveys a message and thereby engages in "speech" within the meaning of the First Amendment.[228]

The judgments rendered respectively by the Court of Appeals for the Fifth Circuit and by the Court of Appeals for the Eleventh Circuit in the cases of *NetChoice v Paxton* and *NetChoice v Moody* are thus clearly split in their rulings, offering conflicting interpretations, at the federal level, of the relationship of content moderation and content curation practices of online platforms with the First Amendment and Section 230. Such a split outcome reflects, in this respect, an internal struggle within the US constitutional framework. On 29 September 2023, the SCOTUS agreed to hear appeals on the two decisions' outcomes.[229]

The arguments were eventually heard in the hearing of 26 February 2024: although, at the time of writing, the Court has not issued its decision yet, several (but not all) of the judges – including Justices Kagan, Kavanaugh, and Barrett – appeared to be rather skeptical on the constitutional validity of the two state legislations.[230] Be that as it may, it is clear that the judgment of the SCOTUS, expected to be delivered by summer 2024, will play a paramount role in shaping the US jurisprudence and case law on the relationship between the First Amendment and the Internet.

4.4.4.2. Questioning platforms' immunity for harmful content: *Gonzalez v Google*, *Twitter v Taamneh*, and *Volokh v James*

With respect to the second line of criticisms, that is, the argument that online platforms would profit from harmful speech and harmful conduct, some attempts have also been made to reduce the degree of immunity of online platforms.

In this regard, most notably, the SCOTUS found itself at a crossroads when it heard the related cases of *Gonzalez v Google* and *Twitter v Taamneh* in February 2023. In both cases, plaintiffs claimed compensation of damages from online platforms following the killing of their relatives in terrorist attacks in Paris and Istanbul, based on the argument that those platforms' moderation systems had contributed to the dissemination of the

[228] ibid 1210.

[229] Adam Liptak, 'Supreme Court to Hear Challenges to State Laws on Social Media' *The New York Times* (29 September 2023) <https://www.nytimes.com/2023/09/29/us/supreme-court-social-media-first-amendment.html> accessed 1 October 2023.

[230] Amy Howe, 'Supreme Court Skeptical of Texas, Florida Regulation of Social Media Moderation' (*SCOTUSblog*, 26 February 2024) <https://www.scotusblog.com/2024/02/supreme-court-skeptical-of-texas-florida-regulation-of-social-media-moderation/> accessed 26 April 2024.

ideologies promoted by the Islamic State of Iraq and Syria (ISIS). In fact, the two cases had been dealt jointly, together with a third case (*Clayborn v Twitter*), by the Court of Appeals of the Ninth Circuit.[231] On that occasion, the Court of Appeals had noted how the cases of *Taamneh* and *Clayborn* concerned the possibility of interpreting platforms' conducts as "aiding and abetting" the promotion of ISIS-related propaganda, whereas *Gonzalez* touched more directly on the consistency of the plaintiffs' claims with Section 230.

The plaintiffs in *Gonzalez*, most notably, submitted that Section 230 should not be applicable in their case and should therefore not bar them from the possibility of claiming compensation for the damages suffered. Three main arguments were brought in this respect: first, they argued against the extra-territorial applicability of the Section 230 immunity; second, they held that Congress' 2016 Justice Against Sponsors of International Terrorism Act,[232] amending the Anti-Terrorism Act,[233] had implicitly repealed Section 230 at least with respect to terrorist content online; third, they submitted that Section 230 immunity does not apply to Anti-Terrorism Act claims based on criminal statutes.[234] All these arguments, however, including in particular that concerning the suggested implied repeal of Section 230, were rejected by the Court of Appeals for the Ninth Circuit.

When hearing the case of *Gonzalez*, the Justices of the SCOTUS clearly acknowledged that their decision would likely have an extraordinary impact on the future of the Internet. Such an acknowledgment was expressed, namely, by Justice Kagan who, during the hearing, admitted on the one hand that it is not at all clear why the tech industry, as opposed to any other industry, should not internalize the costs deriving from its own conduct, while declaring on the other hand that accepting the views of the plaintiffs would lead to rather uncertain consequences. Such a choice, according to Justice Kagan, should therefore be taken by the Congress rather than by the SCOTUS – which, in her own words, is clearly not composed of the "nine greatest experts on the Internet".[235]

Besides, the SCOTUS heard contextually the related case of *Twitter v Taamneh*, which was eventually decided on 18 May 2023.[236] The judgment of the Court, unanimous and authored by Justice Thomas, rejected the plaintiffs' claims, arguing that the content moderation and content curation practices put in place by the social network could not be considered to amount to aiding and abetting of ISIS-related propaganda, especially because the notion itself of "aiding and abetting" implies a willful and active conduct which, according to the Court, was not the case of the social media platform:

> To be sure, plaintiffs assert that defendants' "recommendation" algorithms go beyond passive aid and constitute active, substantial assistance. We disagree … Viewed properly,

[231] *Gonzalez v Google LLC* 2 F4th 871 (9th Cir 2021).
[232] Justice Against Sponsors of Terrorism Act 2016.
[233] Anti-Terrorism Act 1990.
[234] *Gonzalez v Google* (n 231) 886.
[235] Recording available on Robert Barnes and others, 'Supreme Court Considers If Google Is Liable for Recommending ISIS Videos' *Washington Post* (21 February 2023) <https://www.washingtonpost.com/technology/2023/02/21/gonzalez-v-google-section-230-supreme-court/> accessed 19 September 2023.
[236] *Twitter, Inc v Taamneh* 598 US 471 (2023).

> defendants' "recommendation" algorithms are merely part of that infrastructure. All the content on their platforms is filtered through these algorithms, which allegedly sort the content by information and inputs provided by users and found in the content itself. As presented here, the algorithms appear agnostic as to the nature of the content, matching any content (including ISIS' content) with any user who is more likely to view that content. The fact that these algorithms matched some ISIS content with some users thus does not convert defendants' passive assistance into active abetting. Once the platform and sorting-tool algorithms were up and running, defendants at most allegedly stood back and watched; they are not alleged to have taken any further action with respect to ISIS.[237]

Twitter v Taamneh's focus on the issue of the possibility of recognizing the platform's actions as a conduct of aiding and abetting played an essential role in helping the SCOTUS to overcome the *impasse* it had found itself in during the hearing for the case of *Gonzalez*. Indeed, the Court rendered a rather short decision on the same date, 18 May 2023, where it rejected the plaintiffs' arguments by simply referring to that "precedent" and thus avoiding the need to deal with the role of Section 230:

> We need not resolve either the viability of plaintiffs' claims as a whole or whether plaintiffs should receive further leave to amend. Rather, we think it sufficient to acknowledge that much (if not all) of plaintiffs' complaint seems to fail under either our decision in *Twitter* or the Ninth Circuit's unchallenged holdings below. We therefore decline to address the application of § 230 to a complaint that appears to state little, if any, plausible claim for relief. Instead, we vacate the judgment below and remand the case for the Ninth Circuit to consider plaintiffs' complaint in light of our decision in *Twitter*.[238]

Nevertheless, in her concurring opinion to *Twitter v Taamneh*, Justice Brown Jackson argued that while she joined the Court's opinion with the understanding that both decisions were closely related (both had also been filed by the same counsel) and rested on the specific allegations brought up by the plaintiffs' counsel, "other cases presenting different allegations and different records may lead to different conclusions".[239]

Apart from such attempts to overcome, judicially, the wide-ranging immunity of online platforms, another relevant case concerned the imposition by the state of New York of limited due diligence duties upon social media networks with specific respect to countering "hateful conduct" online. The New York Hateful Conduct Law, adopted following the livestream of a mass shooting by a white supremacist in Buffalo, introduced, first of all, the duty to "provide and maintain a clear and easily accessible mechanism for individual users to report incidents of hateful conduct", where "hateful conduct" is defined as "the use of a social media network to vilify, humiliate, or incite violence against a group or a class of persons on the basis of race, color, religion, ethnicity, national origin, disability, sex, sexual orientation, gender identity or gender expression"; additionally, the Law introduced an obligation to "have a clear and concise policy readily available and accessible on their website and application which includes how such social media networks will respond and address the reports of incidents of hateful conduct on their platform".[240]

---

[237] ibid 499.
[238] *Gonzalez v Google LLC* 598 US 617 (2023) 622.
[239] *Twitter v Taamneh* (n 236) 507.
[240] NY State Assembly Bill 2021-A7865A 2021 s 1.

The Law also clarified that the new duties should not be interpreted as imposing upon social media networks an obligation adversely affecting the rights and freedoms of any person, notably their right to free speech, or as "add[ing] or increas[ing] liability of a social media network for anything other than the failure to provide a mechanism for a user to report to the social media network any incidents of hateful conduct on their platform and to receive a response on such report".[241] In other words, the Law did not require social media networks to take down content amounting to "hateful conduct" but, rather, to simply make available a system for users to report such cases and to have a response thereupon.

The Hateful Conduct Law was, nevertheless, brought before the District Court for the Southern District of New York which, in *Volokh v James*,[242] issued on 14 February 2023 a preliminary injunction enjoining the enforcement of the statute. The Court held, in this respect, that the Law was not, in fact, in violation of Section 230, precisely because it did not "even require that social media networks remove instances of 'hateful conduct' from their websites"[243] – thus, the Law did not treat providers as publishers. However, the Law was considered to be in breach of the First Amendment because it entailed a content-based regulation of speech which, on the one hand, could have had a negative chilling effect on constitutionally protected speech[244] and, on the other hand, would have compelled social media networks to "speak", through their policies, against hate speech itself:

> Here, the Hateful Conduct Law requires social media networks to disseminate a message about the definition of "hateful conduct" or hate speech – a fraught and heavily debated topic today … [T]he dissemination of a policy about "hateful conduct" forces Plaintiffs to publish a message with which they disagree. Thus, the Hateful Conduct Law places Plaintiffs in the incongruous position of stating that they promote an explicit "pro-free speech" ethos, but also requires them to enact a policy allowing users to complain about "hateful conduct" as defined by the state.[245]

The case of *Volokh v James* thus represents a highly significant episode, as the Hateful Conduct Law giving rise to it constitutes a rather rare and unusual attempt within the US to counter, albeit in a limited fashion, a phenomenon, that of online hate speech, that has been repeatedly considered by constitutional case law as being protected by the First Amendment. The Law, however, although constructed in such a way as to strive to reduce

[241] ibid.

[242] *Volokh v James* 2023 WL 1991435 (SDNY 2023).

[243] ibid 10.

[244] "Even though the law does not require social media networks to remove 'hateful conduct' from their websites and does not impose liability on users for engaging in 'hateful conduct', the state's targeting and singling out of this type of speech for special measures certainly could make social media users wary about the types of speech they feel free to engage in without facing consequences from the state. This potential wariness is bolstered by the actual title of the law – 'Social media networks; hateful conduct prohibited' – which strongly suggests that the law is really aimed at reducing, or perhaps even penalizing people who engage in, hate speech online". ibid.

[245] ibid 7. Additionally, the Court denies that such a compelled speech would represent a form of commercial speech, i.e., a form of "low-value" speech not subjected to strict scrutiny, as "the policy requirement compels a social media network to speak about the range of protected speech it will allow its users to engage (or not engage) in". ibid.

as much as possible the impact on users' free speech, was not able to pass the strict scrutiny test operated by the District Court.

The judgment, although rendered by a lower court, thus confirms once more the general hostility of US case law towards the imposition of any duties upon providers of intermediary services and, especially, towards the adoption of governmental measures against the dissemination of hate speech on the Internet. Such a perspective may, clearly, enter into contrast with foreign policies and, in particular, with the EU's DSA.

4.4.5. *Digital Services Act and the United States*

The composite legal framework of the US concerning platform governance and content moderation, both in general terms and with specific regard to hate speech, is indeed rather different from the approach currently under development within the EU and, particularly, from the due diligence system introduced by the DSA.[246] In this respect, the opposition between the two value-frameworks characterizing the legislative approaches followed on the two sides of the Atlantic appears to be particularly striking.[247] Although it is true that Section 230 has been subjected to increasing bipartisan criticisms which may, possibly, lead to a reform of the intermediary liability framework of the US, it is yet to be assessed which will be the direction taken by such a hypothetical reform. However, it emerges from the above-described case law that, currently, the prevailing orientation seems to move further away from the DSA system rather than closer.

The choices made by the SCOTUS in the cases of *Twitter v Taamneh* and *Gonzalez v Google* reflect a strong perplexity on the part of the Justices with regard to the prospect of undercutting the scope of action of Section 230 so as to recognize forms of intermediary liability for third-party content. Admittedly, Justice Brown Jackson argued that she might have voted otherwise had the specific circumstances of the case been different, while Justice Kagan expressed her doubts about the appropriateness of a free-for-all intermediary liability framework and suggested that Congress should probably take matters into its own hands. However, it is undeniable that the SCOTUS has demonstrated a very cold attitude towards the plaintiffs' suggestion of holding platforms directly accountable for the harm caused by the content they host.

Besides, if such a conclusion is valid, in general terms, for any type of third-party illegal or harmful content, it is even more valid when it comes to hate speech moderation. The case of *Volokh v James*, albeit at the level of a District Court, represents in this respect a clear confirmation of the aversion towards the introduction of any form of accountability of ISPs for the dissemination of hate speech content. Indeed, in that case, the New York Hateful Conduct Law had not even introduced an obligation to remove hate speech content, but was still considered to be in violation of the First Amendment simply because it envisaged a duty to include anti-hate speech provisions within the intermediaries' terms and conditions.

[246] See *supra*, §3.5.3.
[247] Pollicino (n 191) 6–9.

At the same time, state legislative attempts like Texas' HB 20 and Florida's SB 7072 would pave the way for overcoming platforms' immunity against users' free speech expectations, thus greatly reducing the scope of effectiveness of the Good Samaritan clause. Also in this respect, the trends emerging in the US are likely to force further apart the two frameworks on the opposite sides of the Atlantic, especially when it comes to hate speech governance. It is true that the DSA, too, aims to limit the extensive moderating power of platforms, with a view to protecting the rights and liberties of users, including freedom of expression and the right to non-discrimination. However, the spirit and intentions behind these statutes, passed by deeply conservative state governments, is much different from that of the DSA, namely because their paramount goal is precisely that of guaranteeing service recipients' right to express views that are protected by the First Amendment but in violation of ISPs' own rules, such as, precisely, hate speech.

Therefore, as has been noted, the introduction of rules affecting the liberty of platforms to freely implement their own terms and conditions and imposing a prohibition of viewpoint discrimination on their part would not only represent a choice to move in a direction different from that of the EU but would lead, in fact, to an actual conflict between the two legal regimes:

> Under the Texas law, a platform's removal of content that, for example, denies or questions the extent of the Holocaust, or that is critical of immigration policies or immigrants or COVID-19 vaccines, would likely be considered illegal viewpoint discrimination in content moderation. Yet, a platform's *refusal* to remove such content upon notice would likely violate the terms of the DSA.[248]

In deciding the cases of *NetChoice v Paxton* and *NetChoice v Moody*, the SCOTUS will likely find itself once again at a fundamental crossroads for the future of platform governance and content moderation – not only at the national level but, rather, on a global scale. The choice of recognizing the constitutional legitimacy of statutes such as Texas' HB 20 and Florida's SB 7072 would, indeed, lead to the creation of an additional rift between the framework of the EU and that of the US.

Besides, what the outcome of *NetChoice v Paxton* and *NetChoice v Moody* might be is not fully clear, especially on account of how split the Court was when it first decided to vacate the stay ordered by the Fifth Circuit with respect to the District Court's preliminary injunction. On that occasion, as already mentioned, four Justices – Kagan, Alito, Thomas, and Gorsuch – voted against the majority's decision and, therefore, in favour of allowing the enforcement of HB 20. Such a small margin between minority and majority, so different from the unanimous judgments of *Twitter v Taamneh* and *Gonzalez v Google*, leaves quite open the possibility that the SCOTUS Justices will decide for the constitutional legitimacy of the contested laws.[249]

---

[248] Nunziato (n 208) 123. See, on the same point, Tourkochoriti (n 208) 144–146.

[249] In fact, some of the SCOTUS Justices have for a long time stressed a need to rein in the extraordinary power of platforms. In his concurring opinion for the case of *Biden v Knight First Amendment Institute*, Justice Thomas argued: "Today's digital platforms provide avenues for historically unprecedented amounts of speech, including speech by government actors. Also unprecedented, however, is the concentrated

4.5. A global overview on hate speech and intermediary liability

4.5.1. *Asia*

The intermediary liability framework, notably when referred to hate speech, is rather varied across Asia. The present subsection shall focus, most notably, on three major jurisdictions of the continent, which showcase different legal approaches to the phenomenon: Japan, India, and the People's Republic of China.

In Japan, the 2001 Provider Liability Law provides that intermediaries shall not be held liable for the infringement of others' rights due to the flow of information upon their infrastructures, unless it is technically feasible to take the necessary measures and unless the intermediary either knew about the infringement or knew the existence of relevant information and there were reasonable grounds for it to have knowledge of the infringement.[250] To this extent, the liability framework established by Japanese law is quite similar to the ECD safe harbour system, although the exemption is slightly narrower in that liability arises also in the case of "constructive knowledge", that is, also in the cases where there are simply "reasonable grounds" for knowledge of illegality of user-generated content.[251]

Be that as it may, the applicability of such a liability framework to the case of hate speech is rather debatable in the light of the limited scope of the 2016 Discriminatory Speech Law[252] which, apart from adopting a rather vague notion of hate speech (and only considering hate speech uttered against "persons from outside of Japan"), does not actually contain provisions proscribing and punishing hate speech but simply promotes institutional actions to promote equality.[253] In fact, some case law has opened the doors to the possibility of resorting to other tort-related lawsuits (e.g., defamation). Nevertheless, it seems that the Japanese legal response to hate speech as such is, in general, altogether rather limited.[254]

Intermediary liability in India is first and foremost regulated by Section 79 of the Information Technology Act, pursuant to which "an intermediary shall not be liable for any third party information, data, or communication link hosted by him" unless it fails to "expeditiously remove or disable access" to a material "upon receiving actual knowledge, or on being notified by the appropriate Government or its agency that any information,

---

control of so much speech in the hands of a few private parties. We will soon have no choice but to address how our legal doctrines apply to highly concentrated, privately owned information infrastructure such as digital platforms". *Biden v Knight First Amendment Institute* (n 203) 2.

[250] Act on the Limitation of Liability for Damages of Specified Telecommunications Service Providers and the Right to Demand Disclosure of Identification Information of the Sender 2001 s 3, para 1.

[251] Kyung-Sin Park, 'From Liability Trap to the World's Safest Harbour: Lessons from China, India, Japan, South Korea, Indonesia, and Malaysia' in Giancarlo Frosio (ed), *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020) 265.

[252] Act on the Promotion of Efforts to Eliminate Unfair Discriminatory Speech and Behavior against Persons with Countries of Origin other than Japan 2016.

[253] Craig Martin, 'Striking the Right Balance: Hate Speech Laws in Japan, the United States, and Canada' (2018) 45 Hastings Constitutional Law Quarterly 455, 466–469.

[254] ibid 470.

data or communication link residing in or connected to a computer resource controlled by the intermediary is being used to commit [an] unlawful act".[255] Interpretation of the scope of this provision has undergone significant developments during the last twenty years. In its 2015 judgment of *Shreya Singhal*,[256] the Indian Supreme Court reduced significantly the range of action-triggering notifications, limiting such a power to court decisions only and making India "one of the 'safest' harbours in the world where intermediaries do not have to take down anything unless the courts find the content to be unlawful".[257]

However, more recently, Indian institutions have increasingly begun turning towards the promotion of proactive monitoring by ISPs, as showcased in particular by the adoption in 2021 of new Intermediary Guidelines[258] which, complementing the Information Technology Act, provide for a range of due diligence responsibilities.[259] On top of this, with specific regard to hate speech – the countering of which has often been justified by Indian case law as essential to promote equal participation in society[260] – the government has in the past often resorted to the rather radical strategy of imposing Internet shutdowns, thus cutting down citizens' access to the digital environment itself.[261]

In the People's Republic of China, tort law provides that the victim of a tort (including defamation) committed through the use of a network shall have the possibility to notify the provider of that network service so as to require it to adopt the necessary measures such as deletion, block, or disconnection. Once the provider of the network has been notified, it shall be jointly and severally liable for any additional harm if it fails to act. Similarly, the provider shall be jointly and severally liable where it knows that a user is infringing upon the right or interest of another person through the use of its services and fails to take the necessary measures.[262]

Although some Chinese commentators consider these rules reminiscent of the US framework, and especially of the safe harbour and notice-and-action framework set by the DMCA,[263] it has been observed that the Chinese approach is in fact rather different from that of a safe harbour. Namely, whereas frameworks such as the US one (as well as the ECD and DSA) were "enacted to specify when intermediaries would *not* be held liable", Chinese law rather specifies when intermediaries will be held liable but fails to include a "clause that states that intermediaries 'shall be exempt' from liability in certain

[255] Information Technology Act 2000 s 79.

[256] *Shreya Singhal v Union of India* AIR 2015 SC 1523.

[257] Park (n 251) 263.

[258] Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules 2021.

[259] Indranath Gupta and Lakshmi Srinivasan, 'Evolving Scope of Intermediary Liability in India' (2023) 37 International Review of Law, Computers & Technology 294.

[260] Smarika Lulz and Michael Riegner, 'Freedom of Expression and Hate Speech' in Philipp Dann and Arun K Thiruvengadam (eds), *Democratic Constitutionalism in India and the European Union: Comparing the Law of Democracy in Continental Polities* (Edward Elgar Publishing 2021) 204–208.

[261] Chinmayi Arun and Nakul Nayak, 'Preliminary Findings on Online Hate Speech and the Law in India' (2016) Berkman Klein Center Research Publication No. 2016-19 <https://cyber.harvard.edu/publications/2016/HateSpeechIndia> accessed 25 September 2023.

[262] Tort Law of the People's Republic of China 2010 art 36.

[263] Huaiwe He, 'Online Intermediary Liability for Defamation under Chinese Laws' (2013) 7–8 <https://www.law.uw.edu/media/1403/china-intermediary-liability-of-isps-defamation.pdf> accessed 25 September 2023.

circumstances": thus, whereas the legal texts adopted by the US and EU systems tend to have a "liability-exempting" nature, the text adopted by China follows, rather, a "liability-imposing" orientation.[264] Such a distinction, moreover, has become increasingly evident in the light of interpretive approaches taken by courts "gravitating towards broad conceptions of knowledge and effective notification, unfairly holding intermediaries liable, and naturally incentivizing them into proactive censorship".[265]

The implications with regard to the protection and guarantee of fundamental rights are, thus, cause for concern and this is even more so in the light of the Chinese approach towards the phenomenon of hate speech: it has been observed, in fact, that the country's hate speech legislation has often been used, in contrast with the European perspective,[266] to foster political propaganda and even to repress minority views (including, for instance, the voices of the LGBTQIA+ community).[267]

Besides, like the People's Republic of China, a range of other jurisdictions have introduced legislation concerning the countering of online harmful content, including hate speech and disinformation, that have been subjected to widespread criticisms due to their potentially dangerous impact on the fundamental rights and liberties of individuals and citizens. The Russian Federation, for instance, adopted in the wake of the NetzDG an approach rather similar to that of Germany, however, this translated into a significant legal encroachment on speech expressing "extremist" perspectives criticizing the government.[268] Similarly, legislation against disinformation and "fake news" have been adopted in countries such as Malaysia and Singapore that have proven to be quite harsh on freedom of expression.[269]

Overall, the framework on hate speech and intermediary liability across Asian countries thus moves from jurisdictions adopting a rather liberal approach favourable to freedom of expression (e.g., Japan) to jurisdictions which, conversely, tend to enforce much more stringent rules typical of illiberal legal systems (e.g., People's Republic of China).

---

[264] Park (n 251) 254–255.

[265] ibid 258.

[266] See *supra*, §2.5.2.

[267] "In post-socialist China, the regulation of online hate speech under the aegis of protecting equality rights is largely characterised by a guideline of nationalist campaign and demoralised pragmatism that informs a patchwork of administrative law, criminal law and civil law. Administrative law serves as an online valve of prior restraint guided by a vague concept of national security and cyber sovereignty. Criminal law acts as subsequent punishment to deter any anti-state and anti-social online speech. Civil law provides merely marginal protection of speech and equality rights by imposing stringent liabilities on ISPs and content providers … Whereas the law imposes restrictions on online speech by virtue of mandates of equal protection, the statist campaign of hate(ful) speech often trumps the latter and serves to fend off open criticism of government policies and public discussion of topics that potentially contravene the mainstream political ideologies". Ge Chen, 'How Equalitarian Regulation of Online Hate Speech Turns Authoritarian: A Chinese Perspective' (2022) 14 Journal of Media Law 159, 178.

[268] Canaan (n 26) 123–125.

[269] Pollicino, Bassini and De Gregorio (n 44) 119–122.

4.5.2. *Africa*

Intermediary liability legislation generally began emerging later in the African context if compared to countries of the Global North, also as a result of economic disparities leading to a wider gap in the access to the Internet. However, in the early 2000s, South Africa adopted its Electronic Communications and Transactions Act,[270] which was in good part inspired by the US CDA and DMCA and by the EU's ECD and which subsequently served, towards the end of the 2000s and the beginning of the 2010s, as a blueprint for the development of a "first generation" of liability limitation laws for many countries across the continent, namely Ghana,[271] Zambia,[272] and Uganda.[273]

In the second half of the 2010s, a second generation of laws on intermediary liability began spreading across Africa, following the African Union Convention on Cyber Security and Personal Data Protection.[274] Although the Convention was only ratified by three countries – as opposed to the fifteen which were required for entry into force – it nevertheless contributed to pushing for a new wave of regulations. These, however, focused more on making intermediaries accountable for unlawful conducts rather than on exempting them from liability. Thus, this "second generation" of African intermediary liability laws, which affected countries such as Malawi,[275] Ethiopia,[276] and Kenya, [277] was more inspired by the goal of removing and acting against illegal content. South Africa, too, enacted a Cybercrimes Act[278] inspired by the African Union Convention.[279]

In fact, the fight against illegal content in African countries has been significantly hampered by the gap between the way online platforms and social media deploy their content moderation practices in countries of the Global North and in countries of the Global South. Indeed, limited interest by IT companies with regard to the African market has often led to a lack of adequate resources and machine-learning training for automated detection systems when it comes to non-Western languages spoken in the continent.[280] As a reaction, many African governments, accusing platforms of bolstering hate speech and disinformation content, have begun adopting more and more frequently measures

[270] Electronic Communications and Transactions Act 2002.
[271] Electronic Transactions Act 2008.
[272] Electronic Communications and Transactions Act 2009. The Law was subsequently substituted by the Electronic Communications and Transactions Act 2021.
[273] Electronic Transactions Act 2011.
[274] African Union Convention on Cyber Security and Personal Data Protection 2014.
[275] Electronic Transactions and Cyber Security Act 2016.
[276] Computer Crime Proclamation 2016.
[277] Computer Misuse and Cybercrimes Act 2018.
[278] Cybercrimes Act 2020.
[279] On the historical reconstruction of the two generations of intermediary liability law in Africa, see Nicolo Zingales, 'Intermediary Liability in Africa: Looking Back, Moving Forward?' in Giancarlo Frosio (ed), *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020).
[280] Giovanni De Gregorio and Pietro Dunn, 'Artificial Intelligence and Freedom of Expression' in Alberto Quintavalla and Jeroen Temperman (eds), *Artificial Intelligence and Human Rights* (Oxford University Press 2023) 88–89.

pertaining to direct censorship (such as Internet shutdowns) rather than cooperative strategies involving providers of those platforms:[281]

> Some countries … chose to rely on tactics other than reporting hate speech. Computational propaganda is an increasingly common tool employed by governments, including Ethiopia, Rwanda and Sudan … The proliferation of hate speech on social media has become a primary justification for the increasing governmental use of internet shutdowns. These measures can range from throttling internet speed to the point [o]f making it practically unusable, to completely switching it off … increasingly [such forms of censorship] are understood as one of the few mechanisms available for addressing online speech and offline harms in a moment of crisis … The escalation of internet shutdowns also reflects the frustration on the part of some governments due to their inability to intervene in the governance of online platforms that are often in another jurisdiction, on another continent. In the absence of concerted cooperation with companies, shutting down the entire network or specific digital spaces has become increasingly popular.[282]

Clearly, similar reactions represent a serious encroachment on freedom of expression, showcasing the many challenges still existent with reference to a transnational (and global) fight against the phenomenon of online hate speech.

4.5.3. *Latin America*

Article 13 of the American Convention on Human Rights (ACHR), while recognizing the right to freedom of thought and expression, encompassing the "freedom to seek, receive, and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing, in print, in the form of art, or through any other medium of one's choice",[283] explicitly allows for the imposition of liability when expressly established by the law and when necessary to ensure either the "respect for the rights or reputations of others" or "the protection of national security, public order, or public health or morals".[284] Additionally, the ACHR provides, in a manner that resembles Article 20 ICCPR,[285] that

> any propaganda for war and any advocacy of national, racial, or religious hatred that constitute incitements to lawless violence or to any other similar action against any person or group of persons on any grounds including those of race, color, religion, language, or national origin shall be considered as offenses punishable by law.[286]

Accordingly, many Latin American countries have variously attempted to enact anti-hate speech bans, with a specific focus, in some cases, on online hate speech and on the imposition of *ad hoc* forms of liability upon providers of intermediary services. This has been

---

[281] Giovanni De Gregorio and Nicole Stremlau, 'Platform Governance at the Periphery: Moderation, Shutdowns and Intervention' in Judit Bayer and others (eds), *Perspectives on Platform Regulation. Concepts and Models of Social Media Governance Across the Globe* (Nomos 2021).

[282] Giovanni De Gregorio and Nicole Stremlau, 'Inequalities and Content Moderation' (2023) 14 Global Policy 870, 875.

[283] American Convention on Human Rights (Pact of San José) 1969 art 13, para 1.

[284] ibid 13, para 2.

[285] See *supra*, §2.2.2.1.

[286] ACHR art 13, para 5. With respect to the relationship between the ACHR and the banning of hate speech, see Martina Brun Pereira and others, 'Nuevas Posibilidades de Comunicación, Nuevos Peligros, Nuevos Desafíos: La Libertad de Expresión y El Discurso de Odio En Internet' (2022) 75 Revista IIDH 101.

the case, for instance, of Brazil, Ecuador, Guatemala, Honduras, and Peru, although most of these attempts, to date, are either still ongoing projects or have been unsuccessful. Argentina, conversely, has had in force a law penalizing "discriminatory acts" since 1988.[287]

The case of Venezuela is, nevertheless, one of the most controversial in the context of Latin America. The 2017 Constitutional Law against Hatred[288] introduced criminal sanctions, punishable with incarceration, with respect to the utterance of hate speech, as well as other measures including the possibility for the authorities to order the removal of content and to revoke the concessions granted to communications media and IT companies. However, the way the law has been enforced has been widely criticized from many angles, as its rules have more than often been exploited to censor positions critical of the government.[289]

The case of Brazil is, in turn, rather different, as the federal Civil Rights Framework for the Internet (*Marco Civil da Internet*) provides for a wide intermediary liability exemption for third-party content, stating that an application provider can be liable for civil damages only inasmuch as it has not taken steps to make the content unavailable after a specific court order has been issued.[290] The presence of a court order is not mandatory only in limited cases, such as those concerning copyright infringement, dissemination of non-consensual intimate content, and child sexual abuse material: in these cases, a simple notice and take down mechanism is envisaged. However, hate speech is not included amongst these categories of illegal content.[291]

4.5.4. *Australia*

The matter of intermediary liability in Australia has been dealt with by Australian courts in different ways depending on the type of third-party illegal content being shared through digital infrastructures. A variety of authorities, sometimes in apparent conflict with each other, have emerged with regard to such a legal question in the attempt to find a

---

[287] Marianne Díaz Hernández, *Discurso de Odio En América Latina: Tendencias de Regulación, Rol de Los Intermediarios y Riesgos Para La Libertad de Expresión* (Derechos Digitales América Latina 2020) 10–15 <https://www.derechosdigitales.org/wp-content/uploads/discurso-de-odio-latam.pdf> accessed 27 September 2023; Juan Carlos Lara Gálvez, 'La Defensa de La Libertad de Expresión, La Ciberseguridad, y El Derecho a Una Información Veraz Frente a Las Fake News y La Neutralidad de Internet' in Renata Ávila and others, *Derechos digitales en Iberoamérica: situación y perspectivas* (Fundación Carolina 2023) 97.

[288] *Ley Constitucional contra el Odio, por la Convivencia Pacífica y la Tolerancia* 2017.

[289] "Esta Ley ha sido denunciada por la sociedad civil dado que falla la prueba tripartita de proporcionalidad, legalidad y necesidad de la norma jurídica, e impone penas privativas de libertad hasta por 20 años por difusión de mensajes considerados de odio. En este sentido … la Ley otorga poderes al Estado para aplicar una serie de medidas que suprimen el ejercicio de la libertad de expresión y profundizan la censura, entre las cuales se encuentran la eliminación de contenidos de internet, el bloqueo de sitios web, la revocatoria de licencias a medios de comunicación y la imposición de penas de cárcel hasta por veinte años". Díaz Hernández (n 287) 14.

[290] *Marco Civil da Internet* 2014 art 19.

[291] Luiz Fernando Marrey Moncau and Diego Werneck Arguelhes, 'The Marco Civil Da Internet and Digital Constitutionalism' in Giancarlo Frosio (ed), *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020) 199–201.

reasonable balance between some general principles of common law that may enter into contrast in cases involving intermediary liability.

On the one hand, Australian common law rejects, in principle, the idea that obligations may be imposed upon institutions or individuals to protect the rights of another against harm caused by third parties. this implies, in practice, the idea that no general "duty to rescue" should exist and that no affirmative actions are compelled by the law for the active protection of others' rights. On the other hand, however, common law provides that, for every wrong, the law should always provide a remedy: nevertheless, in cases concerning online harm, such an effective remedy has sometimes proven to be non-existent.[292]

In the case of defamation, liability of ISPs has been found to exist, in general terms, when hosts exercise some degree of control over the content disseminated. In other cases, where the ISP takes a less active role, secondary liability is recognized once the intermediary is actually informed about the likely possibility of carrying defamatory content.[293] Similarly, with regard to the phenomenon of racial vilification, punishable under the Racial Discrimination Act 1975 – which declares it unlawful to do, unless it is in private, an act which is "reasonably likely, in all the circumstances, to offend, insult, humiliate or intimidate another person or a group of people … because of the race, colour or national or ethnic origin of the other person or of some or all of the people in the group" –,[294] courts have considered that the provision of facilities enabling the spread of such content and the failure to remove it can in fact constitute itself an act of publication thereof, at least once the operator has knowledge of it. Therefore, the material act proscribed by the statute is considered to have been put in place in case of failure to remove hate speech content after that content has been notified to the provider. However, the main question that is still under debate concerns the intentional element, that is, the question whether it can be considered that the failure to remove unlawful hate speech contents fulfils the statute's requirement that the publication "act" has, indeed, been committed precisely "because of the race, colour or national or ethnic origin of the other person or of some or all of the people in the group".[295]

On top of such judiciary developments in Australian common law, a composite set of regulatory and self-regulatory interventions have been passed in recent years with a view to promoting a duty of care on social media platforms to counter the presence of unlawful or harmful content online:[296] in 2019, for instance, the government passed an amendment

[292] Kylie Pappalardo and Nicolas P Suzor, 'The Liability of Australian Online Intermediaries' in Giancarlo Frosio (ed), *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020) 236–239.

[293] ibid 240–241.

[294] Racial Discrimination Act 1975 s 18C.

[295] Pappalardo and Suzor (n 292) 243. In this respect, see namely *Silberberg v Builders Collective of Australia Inc* (2007) 164 FCR 475, which excluded the possibility to draw the conclusion of the presence of the intentional element, arguing instead that the failure to remove the contested comments could be just as easily be attributed to inattention or lack of diligence; *Clarke v Nationwide News Pty Ltd* (2012) 289 ALR 345, where it was held that in the case of an active solicitation and moderation of readers' contributions, the ISP can well be found to have committed the act of publication because of racial bias.

[296] Rachel Tan, 'Social Media Platforms Duty of Care – Regulating Online Hate Speech' (2022) 37 Australasian Parliamentary Review 143, 156–158.

of the Criminal Code targeting ISPs failing to notify or delete live or streaming violent content;[297] in February 2021, an Australian Code of Practice on Disinformation was adopted (subsequently updated in December 2022);[298] on 21 January 2022, the Australian Online Safety Act entered into force, determining the creation of an eSafety Commissioner with the power to disable access to harmful and illegal material and promoting and recognizing the creation of new codes of conduct.[299] Overall, Australia, albeit faced with issues concerning the application of the common law by courts, is thus seemingly moving in a direction similar to that of the EU and the UK, by establishing forms of increased responsibility for ISPs for the contents their infrastructures host and contribute to disseminating.

4.6. Conclusions

The comparative overview of the Chapter offers an insight into how different jurisdictions have variously addressed the subject of intermediary liability, especially with respect to the dissemination of hate speech content. These differences are in general the reflection of the value frameworks[300] in which those jurisdictions are imbued, thus creating a spectrum that goes from highly liberal approaches granting wide immunity to ISPs (e.g., the US) to much more authoritarian perspectives that raise, in fact, significant concerns in terms of the protection of fundamental rights – notably, freedom of expression.

Amidst these different perspectives, the EU's strategy against hate speech, deeply influenced *inter alia* by the relevant ECtHR case law, is specifically characterized at its core by its drive towards the promotion of the values of human dignity and, specifically, of the value of equality.[301] Against this backdrop, the challenge of the EU in the context of the regulation on platform governance and, especially, of online hate speech governance will therefore be that of being able to guarantee, foster, and promote such core principles *vis-à-vis* such a transnational phenomenon. This challenge is twofold: internally, EU institutions will have to deal with the different sensibilities characterizing the various Member States in this field; externally, the strategies adopted at the European level may likely clash with conflicting foreign legislations. Most notably, the most significant issue is arguably the relationship with the legal framework on the Western side of the Atlantic, especially in light of the fact that most platforms are, indeed, US-based.

Furthermore, the road towards the affirmation of an EU equality-driven approach towards the governance of hate speech will not only have to face the challenges represented by the international mosaic of jurisprudence on the matter. Indeed, a second order of challenges result from the technological and organizational systems adopted by the subjects

[297] Sharing of Abhorrent Violent Material Act 2019.

[298] Digital Industry Group, 'Disinformation Code' (*DIGI*) <https://digi.org.au/disinformation-code/> accessed 24 September 2023.

[299] Online Safety Act 2021.

[300] Oreste Pollicino, *Judicial Protection of Fundamental Rights on the Internet: A Road Towards Digital Constitutionalism?* (Hart 2021) 14–21; Pollicino (n 191) 6–9.

[301] See *supra*, §§2.3.2, 2.5.2.2.

of regulation themselves: that is, ISPs. The following Chapter shall therefore analyse this second order of challenges, focusing most notably on the issues connected to the adoption of automated systems of hate speech moderation.

.

5.

Platform Moderation and Hate Speech in the Algorithmic Age: Preserving Substantive Equality

5.1. Introduction

Whereas the previous Chapter considered how hate speech and intermediary legislation have developed both within and outside Europe and how the legislation of the EU, represented notably by the DSA, may relate to those frameworks, the purpose of the present Chapter shall be that of investigating how providers of intermediary services themselves have addressed the phenomenon of hate speech, both in terms of the policies adopted and in terms of the practical means of enforcement of those policies. Indeed, such an investigation represents a necessary starting point to identify what challenges still lie ahead in

the governance of the phenomenon of online hate speech and how the EU should strive to deal with such challenges.

The Chapter is structured into two parts. The first part, which includes sections 5.2 and 5.3, is focused on the private anti-hate speech strategies applied by major providers of intermediary services. Section 5.2, in particular, deals with the policies, standards, and terms and conditions formulated by these actors. Specific attention is given to the case of Meta platforms – taken as a paramount example also in light of the significant insights offered by the company's recently established Oversight Board (§5.2.1) – as well as to other platforms, namely X, YouTube, and TikTok (§5.2.2), with a view to identifying common patterns and features (§5.2.3.). Section 5.3, in turn, addresses the technical means through which hate speech is actually moderated (§5.3.1), focusing on the rise of AI detection systems (§5.3.2) and giving an overview of their functioning and limitations (§§5.3.3, 5.3.4.)

The second part of the Chapter addresses the challenges that the ways in which platforms moderate hate speech pose to the law and, specifically, to European hate speech governance. Section 5.4 underlines how the resort to AI systems for content moderation and content curation necessarily entail the presence of certain margins of error, thus requiring policymakers and lawmakers to define the limits of "acceptability" of error (§5.4.1), and suggests substantive equality as a proxy to determine the borders of acceptable errors in the context of hate speech moderation in Europe (§5.4.2). It also indicates some areas of action to be addressed with a view to mitigating the collateral effects of errors (§5.4.3). Section 5.5. underscores how the DSA may indeed serve as the baseline for such mitigating interventions within the European context (§5.5.1), while arguing that more specific guidelines could (and should) be adopted through a renovation of the EU CoC on Illegal Hate Speech (§5.5.2) and clarifying that the DSA is in fact set within a larger, developing, European framework on AI (§5.5.3).

Section 5.6, finally, presents some brief conclusions serving as a bridge for the final remarks of the present work.

5.2. Hate speech and providers: an overview of very large online platforms' terms and conditions

An analysis of the content moderation practices of providers of intermediary services and of the impact of such practices on fundamental rights and democratic values requires, first and foremost, an overview of those providers' content policies and standards, so as to identify common patterns and peculiar aspects, as well as possible continuities or discontinuities with the international and/or European framework.[1]

[1] In this respect, see most notably Eva Nave and Lottie Lane, 'Countering Online Hate Speech: How Does Human Rights Due Diligence Impact Terms of Service?' (2023) 51 Computer Law & Security Review 105884, 12–17, comparing the terms of service on hate speech of Meta platforms, Twitter, and YouTube with European legal standards. See also Richard Wilson and Molly Land, 'Hate Speech on Social Media: Content Moderation in Context' (2021) 52 Connecticut Law Review 1029, 1046–1063.

The first subsection addresses, specifically, the standards followed by Meta platforms, which represent, in this respect, a paramount case study. This is due, in particular, to the degree of transparency of information released circa their moderation practices, as well as to the company's choice to create an independent Oversight Board vested with the power to monitor the correct application of Meta's standards in compliance with human rights principles. The decisions of the Board, indeed, can give useful insights into how Meta platforms actually enforce their rules. The second subsection gives a brief overview of the terms and conditions adopted with respect to hate speech by other notable social media platforms: X, YouTube, and TikTok.

5.2.1. *Meta Platforms and the Oversight Board*

5.2.1.1. The definition of hate speech under Meta's standards

Meta, in its Transparency Center, justifies the choice to filter out hate speech from Facebook and Instagram on account of the fact that "people use their voice and connect more freely when they don't feel attacked on the basis of who they are" and that hate speech "creates an environment of intimidation and exclusion, and in some cases may promote offline violence".[2]

In this respect, the strategies implemented by Meta against hate speech aim to reflect and balance the declared core values of the company. Most notably, while Meta's central goal is to "create a place for expression and give people a voice" ("Voice"), such freedom of expression may be limited under the platforms' standards with a view to ensuring other essential values. These include, in particular, that of "Safety", which requires to "remove content that could contribute to a risk of harm to the physical security of persons" – and thus to avoid the presence of threatening content which "has the potential to intimidate, exclude or silence others" –, and that of "Dignity", which entails the belief that "all people are equal in dignity and rights" and are thus to be protected from any form of harassment or degradation by other users.[3]

In this sense, Meta gives a rather analytical definition of what "hate speech" actually is, both with respect to the subjective scope of victimized categories of people considered and with respect to the objective scope of the types of utterances that are considered to actually amount to hate speech. With respect to the first profile, Meta considers a very wide variety of prohibited grounds of discrimination which go much further than most governmental proscriptions: race; ethnicity; nationality; disability; religion; caste; sexual orientation; sex; gender identity; subjection to serious diseases; age, when it is referenced alongside other characteristics; the condition of refugee, migrant, immigrant, and asylum seeker; and, in some cases, to characteristics such as a person's occupation.

[2] Meta, 'Facebook Community Standards: Hate Speech' (*Transparency Center*) <https://transparency.fb.com/policies/community-standards/hate-speech/> accessed 8 December 2023.

[3] Meta, 'Facebook Community Standards' (*Transparency Center*) <https://transparency.fb.com/policies/community-standards/> accessed 8 December 2023.

Furthermore, with respect to the second profile, speech acts that are prohibited are "direct attacks", a category encompassing a wide range of conducts, going from the utterance of dehumanizing speech to harmful stereotypes, from statements of inferiority to expressions of contempt, disgust, or dismissal, and from cursing to calls for exclusion or segregation.[4]

Meta's community standards also include a categorization of hate speech cases into three tiers. Tier 1 encompasses violent speech or support to violence in written or visual form; dehumanizing speech or imagery;[5] and the mocking of the concept, as well as of events or victims, of hate crimes. Tier 2 is associated with generalizations stating physical, mental, or moral inferiority or deficiencies of protected groups (as well as other statements of inferiority), expressions of contempt; expressions of dismissal, expressions of disgust, and cursing. Tier 3 consists, finally, of calls for action, statements of intent, aspirational or conditional statements, or statements advocating or supporting segregation or exclusion of a person or group of people, as well as content describing or negatively targeting people with slurs, "where slurs are defined as words that inherently create an atmosphere of exclusion and intimidation against people … even when targeting someone who is not a member of the … group that the slur inherently targets".[6]

5.2.1.2. Hate speech in the "case law" of the Oversight Board

Between 2019 and 2020, Meta (which was then still called Facebook) established the Oversight Board (OB), vested with the task of assisting the company's platforms in protecting freedom of expression while balancing it with the above mentioned values of safety and dignity, as well as with those of authenticity and privacy.[7] According to its

[4] Meta, 'Facebook Community Standards: Hate Speech' (n 2). "We 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 dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. We also prohibit the use of harmful stereotypes, which we define as dehumanizing comparisons that have historically been used to attack, intimidate, or exclude specific groups, and that are often linked with offline violence. We consider age a protected characteristic when referenced along with another protected characteristic. We also protect refugees, migrants, immigrants and asylum seekers from the most severe attacks, though we do allow commentary and criticism of immigration policies. Similarly, we provide some protections for characteristics like occupation, when they're referenced along with a protected characteristic. Sometimes, based on local nuance, we consider certain words or phrases as frequently used proxies for PC groups".

[5] I.e., comparisons to insects or animals perceived as inferior; to filth, pathogens, disease; to feces; to subhuman groups; to sexual predators, violent criminals, or other criminals; to objects. In this respect see, e.g., *Planet of the Apes racism* [2023] 2023-035-FB-UA; *Dehumanizing Comments About People in Gaza* [2024] 2024-026-FB-UA. Furthermore, the standards also prohibit statements denying the existence of some protected groups, as well as harmful stereotypes that are historically linked to forms of discrimination and oppression (including Blackface and Holocaust denial).

[6] Meta, 'Facebook Community Standards: Hate Speech' (n 2).

[7] Meta Oversight Board, 'Oversight Board Charter' (*Oversight Board*, February 2023) 3 <https://oversightboard.com/attachment/494475942886876/> accessed 25 October 2023.

Charter, the OB is composed of a minimum of 11 and a maximum of 40 members,[8] appointed on the basis of criteria of "knowledge, competencies, diversity, and expertise" for a term of three years (and for a maximum of three terms).[9] The main task of the OB consists of the review of Meta's content moderation decisions (be they decisions to remove or to uphold content) upon request of the recipients of the service, provided that they have exhausted all appeals procedures offered by Meta's platforms.[10]

Presented from the beginning as an intended "Supreme Court"[11] for the company platforms' decisions on content moderation, the OB represents a remarkable experiment in the context of the private platform-based adjudication of freedom of expression on a global scale. As such, it has garnered significant attention from the academia as well as from society at large, with leading media outlets worldwide referring to the Board's decisions as a leading authority in the context of online content moderation governance.[12] In this respect, the OB has been described as a sign of a societal constitutionalization process in progress[13] and, generally, as a "laboratory to study the transnational challenges which the information society has raised to global (digital) constitutionalism".[14] However, at the same time, the OB has also been at the centre of many debates concerning, *inter alia*, its degree of independence with respect to Meta platforms, its transparency, and, in general, its practical effectiveness.[15]

Be that as it may, the OB offers important insights into the interpretation and operationalization of Meta's standards on prohibited content, including hate speech. Indeed, according to the Oversight Board Charter, the Board's decisions "will be binding and Meta will implement [them] promptly, unless implementation of a resolution could violate the law". Additionally, even if Meta is not actually bound to comply with any recommendation included in a decision or policy advisory opinion, it should nevertheless "take further action by analyzing the operational procedures required to implement the

[8] The number of appointed members has nevertheless been, up to now, lower than 30. See Evelyn Douek, 'The Meta Oversight Board and the Empty Promise of Legitimacy' (SSRN, 7 September 2023) 23 <https://papers.ssrn.com/abstract=4565180> accessed 25 October 2023.

[9] Meta Oversight Board (n 7) art 1, ss 1–3.

[10] ibid 2, s 1. The provision clarifies that the OB "has the discretion to choose which requests it will review and decide upon" and that, for this purpose, it shall "seek to consider cases that have the greatest potential to guide future decisions and policies".

[11] Casey Newton, 'Facebook Will Create an Independent Oversight Group to Review Content Moderation Appeals' (*The Verge*, 15 November 2018) <https://www.theverge.com/2018/11/15/18097219/facebook-independent-oversight-supreme-court-content-moderation> accessed 25 October 2023; Kate Klonick, 'The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression' (2020) 129 Yale Law Journal 2418, 2425.

[12] Douek (n 8) 4–5.

[13] Angelo Jr Golia, 'The Transformative Potential of Meta's Oversight Board: Strategic Litigation within the Digital Constitution?' (2023) 30 Indiana Journal of Global Legal Studies 325.

[14] Oreste Pollicino and Giovanni De Gregorio, 'Shedding Light on the Darkness of Content Moderation: The First Decisions of the Facebook Oversight Board' (*Verfassungsblog*, 5 February 2021) <https://verfassungsblog.de/fob-constitutionalism/> accessed 25 October 2023.

[15] Klonick (n 11); David Wong and Luciano Floridi, 'Meta's Oversight Board: A Review and Critical Assessment' (2023) 33 Minds and Machines 261; Douek (n 8).

recommendations, considering those recommendations in the formal policy development process of Meta, and transparently communicating about actions taken as a result".[16]

Furthermore, it is specified that all decisions "have precedential value and should be viewed as highly persuasive when the facts, applicable policies, or other factors are substantially similar".[17] Although, clearly, the inherent voluntary and self-regulatory nature of the OB implies that full compliance with the Charter's provisions is necessarily dependent on the good will of Meta itself,[18] the pronouncements of the Board have thus a great potential to influence the policies of the company in terms of content moderation – also in light of the harm to image that a failure to respect the decisions would represent.

A variety of decisions by the OB deal specifically with the topic of hate speech, suggesting desirable courses of action to address its spread across Meta platforms while attempting to guarantee the maximum possible degree of protection of freedom of expression. To achieve this goal, and consistently with Meta's declared intent[19] to strive to comply with the United Nations' Guiding Principles on Business and Human Rights (UNGPs),[20] the OB generally refers, on top of the platforms' community standards, to the international framework on human rights law – namely, to Articles 19-20 ICCPR and to the ICERD.[21] However, the possibility for the OB to rely on such a framework was not a given from the start.

As has been noted, the Board's Charter and Bylaws do not reference international human rights law as a basis for decisions. Nevertheless, "the Board has elevated [international human rights law] as its primary source of authority, citing it in every decision".[22] As a result, the Board's approach towards hate speech moderation follows *inter alia* the principles expressed in documents such as General Comment No. 34 of the Human Rights Committee,[23] the so-called "Rabat Plan of Action",[24] and General Recommendation No. 35 of the Committee on the Elimination of Racial Discrimination,[25] all of which offer an insight into the system of conditions and criteria to assess the conformity of measures

[16] Meta Oversight Board (n 7) art 4.

[17] ibid 2, s 2.

[18] In fact, the Charter does not include an enforcement mechanism in case a decision by the OB is not complied with by Meta. See Douek (n 8) 11.

[19] Miranda Sissons, 'Our Commitment to Human Rights' (*Meta*, 16 March 2021) <https://about.fb.com/news/2021/03/our-commitment-to-human-rights/> accessed 4 November 2023.

[20] Office of the High Commissioner for Human Rights, 'Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework' (United Nations 2011) HR/PUB/11/04 <https://www.ohchr.org/sites/default/files/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf> accessed 4 November 2023.

[21] See *supra*, §2.2.2.

[22] Douek (n 8) 26. In this respect, see also Klonick (n 11) 2478; Ruby O'Kane, 'Meta's Private Speech Governance and the Role of the Oversight Board: Lessons from the Board's First Decisions' (2022) 25 Stanford Technology Law Review 167, 177.

[23] Human Rights Committee, 'General Comment No. 34. Article 19: Freedom of Opinion and Expression' (United Nations 2011) CCPR/C/GC/34. See *supra*, §2.2.2.1.

[24] Human Rights Committee, 'Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred That Constitutes Incitement, to Discrimination, Hostility or Violence' (United Nations 2013) A/HRC/22/17/Add.4.

[25] Committee on the Elimination of Racial Discrimination, 'General Recommendation No. 35. Combating Racist Hate Speech' (United Nations 2013) CERD/C/GC/35. See *supra*, §2.2.2.2.

against hate speech to the international human rights framework on freedom of expression.[26] Based on such a framework, the Board deploys the traditional threefold test to assess the conformity with human rights law of Meta's anti-hate speech measures, by verifying the legality (i.e., pre-existence of a clear and accessible rule), legitimacy (i.e., pursuit of a "legitimate aim"), and necessity/proportionality thereof.

With respect to the latter aspect, specific importance is given to the context in which the utterance at stake is published. In the *Myanmar post about Muslims* case, for example, the OB considered a case where a user from Myanmar had uploaded a post in Burmese that compared the perpetration of killings in France, following the publication of offensive depictions of the Prophet Muhammad, with the alleged lack of a response against the treatment of Uyghurs in the People's Republic of China, suggesting that such an apparent incoherence in the reactions from the Muslim community would be indicative that "there is something wrong psychologically" with Muslim men. In this case, the OB overturned Facebook's decision to remove that post, focusing precisely on the context of such affirmations. According to the OB, on the one hand, the subject of mental issues did not represent a common Islamophobic argument in Myanmar. On the other hand, the intent behind the comment appeared to be a criticism of an inconsistent conduct rather than a direct attack on Muslims as such.[27]

In the *Russian poem* case, the OB dealt with a post featuring the image of the dead body – face down – of a person who had been shot in Bucha (Ukraine), which was associated with a text comparing Russians to Nazis and a quote from a poem by Soviet poet Konstantin Simonov: "Kill the fascist… Kill him! Kill him! Kill!". The post, once again, was considered by the OB not to represent a case of hate speech, precisely because, when put in context,

> the excerpts with violent language of the poem "Kill him!" … may be read as describing, not encouraging, a state of mind. When read together with the entire post, including the photographic image, the excerpts are part of a broader message warning of the potential for history to repeat itself in Ukraine. They are an artistic and cultural reference employed as a rhetorical device by the user to convey their message.[28]

Conversely, in other cases, the OB found that Meta had actual responsibilities, under the UNGPs read *de juncto* with international human rights law, to remove specific content. This was the case, for example, of a version of Disney's cartoon *The Pied Piper* which had been edited by Croatian users and associated Serbs with rats.[29] Similar conclusions were reached in a case concerning the dissemination in Ethiopia of the false news of alleged killings and raping of women and children by the Tigray people's Liberation Front,[30] as well as with respect to the publication on Instagram of a meme containing false

[26] References to these sources are present from the very first decisions of the OB addressing the subject of hate speech: see, among others, *Myanmar post about Muslims* [2021] 2020-002-FB-UA; *Armenians in Azerbaijan* [2021] 2020-003-FB-UA; *Depiction of Zwarte Piet* [2021] 2021-002-FB-UA; *South Africa slurs* [2021] 2021-011-FB-UA.

[27] *Myanmar post about Muslims* (n 26).

[28] *Russian poem* [2022] 2022-008-FB-UA.

[29] *Knin cartoon* [2022] 2022-001-FB-UA.

[30] *Alleged crimes in Raya Kobo* [2021] 2021-014-FB-UA.

and distorted claims about the Holocaust.[31] In these cases, the Board stressed Meta's due diligence duties to actively remove hate speech when similar content risks affecting the right to equality, non-discrimination, and life of targeted groups – duties that are heightened, of course, when such material is posted against the backdrop of recent ethnic conflicts and/or of ongoing violent conflicts.

5.2.1.3. Promoting equality and counternarratives

Meta's standards concerning hate speech, as also interpreted by the OB, thus attempt to strike a balance between the company's human rights responsibilities towards, on the one hand, freedom of expression and, on the other hand, victims' rights to equality, non-discrimination, dignity, life, and safety. Hate speech, therefore, should be removed as it pollutes the information shared on the platforms and negatively impacts this second set of rights. However, the assessment in fact of whether a certain content amounts to a case of hate speech must be grounded in a careful examination of all contextual aspects, so as to avoid an excessive restriction of users' "voice". In this respect, Meta's attitude is reminiscent, to a certain extent, of Europe's "militant" approach.[32]

Besides, against this backdrop, it is worth analysing whether, how, and to what extent Meta's standards and the decisions of the OB also contribute to actively promoting equal speech opportunities for marginalized communities and, thus, to helping create a digital landscape characterized by a flourishing of substantive equality values. In this respect, it is worth mentioning a specific exception to the general standard conditions on hate speech considered by Meta. Indeed, although the company clarifies that the use of slurs is generally prohibited, it recognizes nevertheless that such terms could in fact be used with a view to condemning them or raising awareness, or may even be employed "self-referentially or in an empowering way". In such cases, Meta's policies allow the use of those slurs, provided that users "clearly indicate their intent".[33]

Coupled with the declared conviction that hate speech should be prohibited especially because it directly affects its victims' ability to fruitfully participate in the community ("people use their voices and connect more freely when they don't feel attacked on the basis of who they are"),[34] such an exception represents an essential feature of Meta's standards, as it implicitly recognizes the need for heightened guarantees for the freedom of expression of those minority, vulnerable, discriminated, or marginalized communities that are generally targeted by hate speech, with a view to promoting their ability to develop appropriate forms of counter-speech.[35] To this extent, the standards seem to

[31] *Holocaust Denial* [2024] 2023-022-IG-UA.
[32] See *supra*, §2.3.2.
[33] Meta, 'Facebook Community Standards: Hate Speech' (n 2).
[34] ibid.
[35] In this respect, see also Meta, 'Counterspeech' (*Counterspeech*) <https://counterspeech.fb.com/en/> accessed 6 November 2023.

internalize at least the "participative" dimension of the concept of substantive equality as interpreted by Fredman.[36]

However, a variety of cases dealt with by the OB show how the provision has not always been applied consistently by Meta platforms. In the *Wampum belt* case,[37] for example, the OB found that Meta had failed to recognize as counter-speech the publication by an indigenous North American artist of an image portraying a "wampum belt" – i.e., a traditional art form where shells are woven together to form images – that depicted the story of a former residential school for indigenous children where many unmarked graves had been discovered in 2021. The piece of art, entitled *Kill the Indian/Save the Man* had precisely the goal of denouncing the violence that indigenous people had suffered in the past, but had nevertheless been removed.

Although Meta agreed that the removal decision had been a mistake, the OB stressed that the issue remained that "such an unambiguous error may indicate deeper problems of proportionality in Meta's automated and human review processes".[38] Moreover, the OB argued that the mistake also corroborated the criticisms moved by marginalized communities, who had been raising for several years "significant concerns about the rate and impact of false positive removals", and that, under the UNGPs, it was "incumbent on Meta to demonstrate that it has undertaken human rights due diligence to ensure its systems are operating fairly and are not exacerbating historical and ongoing oppression".[39]

Another highly significant episode, in this respect, is represented by the *Reclaiming Arabic words* case.[40] The decision concerned, specifically, the removal of a post published by a public Instagram account dedicated to the open discussion of queer narratives in Arabic countries. The post featured a carousel of images representing different words used in Arabic-speaking countries to indicate in a derogatory way men with "effeminate mannerisms". The purpose, as clarified explicitly by the profile administrator, was precisely that of reclaiming the power of such hurtful terms. In this case, the OB explicitly referred to *Wampum belt*, stating the following:

> Online spaces for expression are particularly important to groups that face persecution and their rights require heightened attention for protection from social media companies. This case also demonstrates the tension for Meta in seeking to protect minorities from hate speech, while also seeking to create a space where minorities can fully express themselves, including by reclaiming hateful slurs … Given the importance of reclaiming derogatory terms for LGBTQIA+ people in countering discrimination, the Board expects Meta to be particularly sensitive to the possibility of wrongful removal of the content in this case and similar content on Facebook and Instagram. As the Board noted in the "Wampum Belt" decision … effects on particular marginalized groups must be taken into account … For LGBTQIA+ people in countries which penalize their expression, social

[36] Sandra Fredman, 'Substantive Equality Revisited' (2016) 14 International Journal of Constitutional Law 712. See *supra*, §2.5.2.1.

[37] *Wampum belt* [2021] 2021-012-FB-UA.

[38] ibid.

[39] ibid.

[40] *Reclaiming Arabic words* [2022] 2022-003-IG-UA.

> media is often one of the only means through which they can still express themselves freely.[41]

These episodes show how, although an approach towards hate speech moderation that is driven by purposes related to the participative dimension of substantive equality may be present *in nuce*, such an aspiration, in practice, comes up against the failure of (human and automated) moderation practices to evaluate correctly the positive value of certain counternarratives.[42]

Besides, other dimensions inherently entrenched in the concept of substantive equality are seemingly absent from the purview of Meta's standards. For instance, Meta applies its anti-hate speech provisions following a strictly symmetric approach, meaning, for example, that it considers equally relevant gender-based hate speech irrespective of it being uttered against women or men, or racial-based hate speech against POCs or white people. Such a choice was explicitly confirmed by the company's Product Policy Forum in 2019, when it was decided not to apply different protections for men and women because of the fundamental implications that policy would entail in terms of the platforms' approach towards hate speech.[43]

To this extent, taking once again Fredman's quadripartition of the dimensions of substantive equality, Meta platforms' standards arguably distance themselves both from the "redistributive" dimension, as they do not take into account and tackle the specific and inherent detrimental consequences attached to a certain social status but follow a "colour-blind"[44] attitude towards the phenomenon of hate speech, and from its "transformative" dimension, as they do not promote a transformation of free speech infrastructures in such a way as to accommodate the needs of victimized groups.[45]

Overall, the approach of Meta platforms towards the governance of hate speech is arguably still dominated by a rather formalistic interpretation of the right to equality.[46] Such a perspective has, nevertheless, led to internal tensions within the OB itself, with some of its members sharing the position of the company and others being much more critical. Indeed, in a decision concerning the removal of two Instagram posts condemning gender-based violence, wherein the author accused men of murdering, raping, and abusing women mentally and physically and declared herself to be a "manhater", the OB declared:

---

[41] ibid. For a similar case, concerning an Instagram post celebrating Pride month while reclaiming a slur traditionally used against gay people, see *Heritage of Pride* [2023] 2023-058-IG-UA.

[42] Conversely, with respect to episodes where Meta had failed to correctly recognize hate speech contents as such, see *Media Conspiracy Cartoon* [2023] 2023-042-FB-UA; *Fictional Assault on Gay Couple* [2023] 2023-051-FB-UA; *Hateful Memes Video Montage* [2024] 2024-015-FB-UA; *Dehumanizing Comments About People in Gaza* (n 5).

[43] Louisa Bartolo, '"Eyes Wide Open to the Context of Content": Reimagining the Hate Speech Policies of Social Media Platforms through a Substantive Equality Lens' (2021) 29 Renewal 39, 44–45.

[44] For a critique of "colourblindness" when dealing with matters of discrimination, see Neil Gotanda, 'A Critique of Our Constitution Is Color-Blind' (1991) 44 Stanford Law Review 1; Kevin Brown, 'Critical Race Theory Explained by One of the Original Participants' (2023) 98 New York University Law Review Online 91.

[45] In this respect, see also *infra*, §5.5.2.3.

[46] Bartolo (n 43).

> For some Board Members, the global context of violence against women is also relevant to the analysis, as the content reflects and raises awareness of a broader worldwide societal phenomenon, further reinforcing that read within the context of the post, the statement was not an assertion that all men are rapists or murderers. On the other hand, other Board Members do not believe that such broad and contested sociological considerations such as root cause assessments or analysis of power differentials should be used to interpret the statement, believing that it could invite controversial interpretations of what constitutes hate speech. The majority of the Board, though cognizant of the societal phenomenon of violence against women and the debates around its root causes, did not rely on them in order to reach its conclusion that the statement was a "qualified" one.[47]

5.2.2. *Other platforms*

5.2.2.1. X's policies

As of April 2023, X's policies against "hateful conduct" state that users "may not directly attack other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease".[48] In this respect, behaviours that are considered to represent hateful "attacks" include: referencing forms of violence or violent events (e.g., genocides such as the Holocaust, lynchings) that have targeted protected categories, if such referencing is made with the intent to harass; inciting to fear (or spread of fearful stereotypes), to harassment, or discrimination; targeting others with repeated slurs, tropes, or content that is degrading or reinforcing negative or harmful stereotypes; dehumanizing a group on the basis of a protected characteristic; publishing hateful imagery (e.g., Nazi symbols, images depicting others as less than human); creating hateful profiles that use hateful images or symbols in the profile image or profile header.[49]

According to X, the rationale behind such a policy is directly related to the detrimental impact that hate speech itself has on the possibility for users to exercise their fundamental right to freedom of expression. Indeed, as underscored by the platform, "if people experience abuse on X, it can jeopardize their ability to express themselves". Furthermore, X also recognizes explicitly that "some groups are disproportionately targeted with abuse online" and that "for those who identify with multiple underrepresented groups, abuse may be more common, more severe in nature, and more harmful".[50]

The declared intentions of the platform thus seem to align in part with the reasoning followed by other platforms, such as, for instance, Meta, as the alleged objective for reducing the spread of hate speech is represented by the goal of promoting "more speech" of targeted groups, an objective which is, in fact, consistent with the participative dimension of equality. Furthermore, X also underlines the importance of evaluating context when assessing whether a certain content is or is not hateful. In particular, X recognizes that "members of a protected category may refer to each other using terms that are

[47] *Violence against women* [2023] 2023-002-IG-UA, 2023-005-IG-UA.

[48] X, 'X's Policy on Hateful Conduct' (*X Help Center*) <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy> accessed 8 November 2023.

[49] ibid.

[50] ibid.

typically considered as slurs" and that, in such cases, the use of those expressions might not be abusive but, rather, a "means to reclaim terms that were historically used to demean individuals".[51] In this respect, the approach followed by X appears to have, at least in theory, several points in common with the standards set by Meta: namely, the platform seems to recognize the need for hate speech policies to guarantee the promotion of the right to (equal) freedom of expression of those groups that have historically been the targets of hate speech.

However, the implementation of such policies by X has been subjected to criticisms, especially in the wake of its acquisition by Elon Musk. On the one hand, it has been argued that, after 2022, the platform has seen a rise in the presence of hate speech content and has failed to properly address it and remove it.[52] On the other hand, it has been observed that the current wording of the policies themselves, updated in April 2023, have "quietly" – but significantly – rolled back some clauses actively supporting victimized groups. Namely, not only did X remove the sentence that specified explicitly that the groups "disproportionately targeted" by hateful conducts are "women, people of color, lesbian, gay, bisexual, transgender, queer, intersex, asexual individuals, and marginalized and historically underrepresented communities", but it also removed the clause based on which the targeted misgendering or deadnaming of transgender individuals was prohibited as hateful conduct.[53] These choices, rather than being neutral, are indicative of a clear departure from the principles and dimensions of substantive equality.

5.2.2.2. YouTube's policies

The online video-sharing platform YouTube also features a specific anti-hate speech policy, pursuant to which it does not allow the sharing of content that promotes violence or hatred against individuals or groups pertaining to one of the following protected statuses: age; caste; disability; ethnicity; gender identity and expression; nationality; race; immigration status; religion; sex/gender; sexual orientation; victims of a major event and their kin; veteran status. On top of encouragement of violence and incitement to hatred, conducts considered to constitute forms of hate speech include, in particular, the dehumanization of protected groups and individuals, the praise or glorification of violence against them, the use of slurs and stereotypes, as well as the dissemination of claims of physical

[51] ibid.

[52] Center for Countering Digital Hate, 'X Content Moderation Failure: How Twitter/X Continues to Host Posts Reported for Extreme Hate Speech' (CCDH 2023) <https://counterhate.com/research/twitter-x-continues-to-host-posts-reported-for-extreme-hate-speech/#about> accessed 8 November 2023; Dan Milmo, 'Twitter Sues Anti-Hate Speech Group over "Tens of Millions of Dollars" in Lost Advertising' *The Guardian* (2 August 2023) <https://www.theguardian.com/technology/2023/aug/02/twitter-accuses-anti-hate-speech-group-over-tens-of-millions-of-dollars-in-lost-advertising> accessed 8 November 2023; Center for Technology and Society, 'Evaluating Twitter's Policies Six Months After Elon Musk's Purchase' (*Anti Defamation League*, 5 September 2023) <https://www.adl.org/resources/blog/evaluating-twitters-policies-six-months-after-elon-musks-purchase> accessed 8 November 2023.

[53] Center for Technology and Society (n 52).

or mental inferiority, the promotion of hateful supremacism, the spread of conspiratorial claims, and the denial or minimization of major violent events.[54]

An interesting aspect is that, although YouTube does not mention, clearly, the possibility that slurs and tropes may be used in a self-empowering manner by victimized groups, it nevertheless allows content containing forms of hate speech "if that content includes an educational, documentary, scientific, or artistic context" (e.g., a documentary about a hate group), especially if that hate speech is accompanied by additional content or context aimed at "condemning, refuting, including opposing views, or satirizing [it]".[55] In this respect, YouTube thus recognizes – although, admittedly, to a rather limited extent – the importance of promoting forms of counter-speech and counternarratives.

5.2.2.3. TikTok's policies

TikTok prohibits hateful behaviour, hate speech, and promotion of hateful ideologies – defined as "systems of beliefs that exclude, oppress, or otherwise discriminate against individuals" – based on the attributes of caste, ethnicity, national origin, race, religion, tribe, immigration status, gender, gender identity, sex, sexual orientation, disability, serious disease, and, in some cases, age.

Examples of prohibited conduct include, *inter alia*, the denial of well-documented historical events such as the Holocaust and the genocide against the Tutsi community in Rwanda, as well as the promotion or advertising of conversion therapy attempting to change a person's sexual orientation or gender identity and the intentional and targeted deadnaming or misgendering of transgender or gender non-conforming individuals. At the same time, TikTok explicitly allows, like Meta and X, the resort to "self-referential slurs used by a member of a particular protected attribute", as well as, like YouTube, the posting of "educational and documentary content raising awareness against hate speech".[56]

5.2.3. *Observations and conclusions*

The ways in which providers of intermediary services – namely, providers of very large online platforms such as the ones addressed in the previous subsections – tend to define and act upon hate speech clearly have deeply significant implications in terms of how recipients of their services are able to experience the Internet and, thus, enjoy the new avenues that technology offers them to engage in freedom of expression. In this respect, the triangular dynamics of free speech characterizing the Internet[57] are particularly

[54] YouTube, 'Hate Speech Policy' (*Google Help*) <https://support.google.com/youtube/answer/2801939?hl=en#zippy=%2Cother-types-of-content-that-violates-this-policy%2Ceducational-documentary-scientific-and-artistic-content%2Cmore-examples> accessed 10 November 2023.

[55] ibid.

[56] TikTok, 'Safety and Civility: Hate Speech and Hateful Behaviors' (*TikTok*, 8 March 2023) <https://www.tiktok.com/community-guidelines/en/safety-civility/> accessed 10 November 2023.

[57] Jack M Balkin, 'Free Speech Is a Triangle' (2018) 118 Columbia Law Review 2011. See *supra*, §2.5.3.

evident and require the law to take into direct account the role that these private actors play, as private regulators,[58] in the overall governance of freedom of expression.

When it comes to hate speech, most of the analysed online platforms have indeed adopted rules to counter its presence across their digital infrastructures, often as a response to criticisms concerning earlier permissive attitudes towards the phenomenon[59] and with a view to guaranteeing users a more enjoyable experience on their platforms.[60] As shown in the previous subsections, notable social media and social network platforms have in fact adopted standards and terms and conditions encompassing wide-ranging sets of conducts and long lists of protected grounds of discrimination.

In most cases, at the same time, such provisions recognize the paramount importance of investigating the context and intent of a certain utterance in order to be able to characterize it as hate speech or as acceptable content. In particular, platforms generally strive to strike an adequate balance, within their terms and conditions, between the goal of removing as much hateful content as possible and that of maintaining online counter-speech and counter-narratives. Nevertheless, the perspective adopted is still, in most cases, strongly driven by a deeply formalistic take on the principle of equality in the enjoyment of freedom of expression online,[61] which carries the inherent risk of a speech governance system that is overall oblivious to the practical power and dominance dynamics that are both the root and result of hate speech.[62]

Especially in the wake of the adoption of the DSA, the question arises as to whether and to what extent such an approach should be deemed to be sufficient in the light of the new human rights-related due diligence obligations of providers of intermediary services, including, notably, the obligation to enforce one's terms and conditions having due regard of users' fundamental rights[63] and the obligation, when choosing to adopt certain measures against systemic risks, to take into particular consideration precisely the collateral effects that those measures might in turn produce.[64] In other words, the challenge will be that of establishing whether the content moderation and content curation practices operated by VLOPs such as Facebook, Instagram, X, YouTube, and TikTok are in fact consistent with the European system of fundamental rights and of identifying which should be the direction that the EU should take to promote an approximation between those practices and the democratic values of equality and pluralism.

In order to better understand the terms of the problem, however, it is essential to investigate the practical and technical means by which platforms actually enforce their

[58] João Pedro Quintais, Giovanni De Gregorio and João C Magalhães, 'How Platforms Govern Users' Copyright-Protected Content: Exploring the Power of Private Ordering and Its Implications' (2023) 48 Computer Law & Security Review 105792.

[59] Wilson and Land (n 1) 1046.

[60] Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press 2018).

[61] Bartolo (n 43).

[62] See *supra*, §2.5.1.

[63] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act), OJ L 277/1 art 14, para 4.

[64] ibid 35, para 1.

policies, standards, and terms and conditions. The following section will thus focus more in-depth on the strategies followed, noting particularly the central role which has come to be played in this respect by AI. In this respect, besides, it is worth noting preemptively how

> algorithms used by online intermediaries effectively advance the intermediaries' own interpretation of legal norms. This process of translating legal mandates into code inevitably embodies particular choices as to how the law is interpreted, which may be affected by a variety of extrajudicial considerations, including the conscious and unconscious professional assumptions of program developers, as well as various private business incentives … Some disparity between the algorithmic representation of law and the law as it operates in practice is hence unavoidable.[65]

5.3. Artificial Intelligence and hate speech moderation

5.3.1. *The many forms of content moderation*

As mentioned in Chapter 2,[66] the term "content moderation" encompasses, *lato sensu*, a variety of different strategies, including both the activities aimed at detecting and removing content that is illegal or in violation of a provider's terms and conditions and, generally, at sanctioning the users uploading such contents (hard moderation or content moderation *stricto sensu*) and the activities aimed at ordering and organizing the distribution of content itself, as well as its demonetization (soft moderation or content curation).[67]

In order to counter the dissemination of hate speech content across their platforms, providers of online platforms generally resort to a combination of both content moderation and content curation practices. For instance, according to X's policies, the publication of hateful content might entail different tiers of sanctions, depending on "a number of factors including, but not limited to the severity of the violation and an individual's previous record of rule violations".[68] Measures adopted could, in this respect, include not only the removal of the controversial content and/or the suspension of the account of the poster, but also reactions at the layer of content curation, such as the reduction of visibility of the post, its downranking, a reduction in its discoverability, or the removal of advertisements adjacent to it.[69] Similarly, YouTube mentions as possible measures both the removal of videos and their demonetization.[70]

---

[65] Maayan Perel and Niva Elkin-Koren, 'Accountability in Algorithmic Copyright Enforcement' (2016) 19 Stanford Technology Law Review 473, 518.

[66] See *supra*, §2.4.3.

[67] James Grimmelmann, 'The Virtues of Moderation' (2015) 17 Yale Journal of Law and Technology 42; Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7 Big Data & Society 2053951719897945; ibid; Emma Llansó and others, 'Artificial Intelligence, Content Moderation, and Freedom of Expression' (TWG 2020) <https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf> accessed 13 December 2021; Tim Wu, 'Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-Ordering Systems' (2019) 119 Columbia Law Review 2001.

[68] X (n 48).

[69] ibid.

[70] YouTube (n 54).

As shown by Klonick,[71] the ways in which these measures can be implemented take, in practice, different forms. A first distinction can be made, based on a temporal criterion, between *ex ante* and *ex post* forms of moderation, depending on whether control is exerted before or after the publication of a specific content upon the platform. An example of *ex ante* tools for moderation can be, for example, "upload filters"[72] that are able to detect the presence of copyrighted material or known CSAM right from the moment when the user attempts to publish a certain content. Conversely, in the case of *ex post* moderation, the platform's approach might be either proactive, meaning that it actively seeks those contents it aims to act upon, or reactive, meaning that it takes a more passive attitude and simply responds to external inputs (e.g., notifications from third users or entities bringing to the platform's attention the possible presence of content in violation of the law or of its terms and conditions).

Another distinctive criterion concerns the actual entity that is vested with content moderation and content curation tasks. Thus, moderation and curation of content may be manual, automated, or hybrid, often based on the type of content with respect to which the provider's policies are intended to be deployed. For instance, whereas manual moderation may be a viable option especially in the case of small communities or media, such as a personal blog, purely automated systems may be implemented by large platforms to identify items infringing copyright, more easily and directly identifiable through the use, for example, of upload filter systems. In many cases, however, hybrid moderation systems have become the preferred avenues – especially for those cases, such as that of hate speech, where the categorization of a content as being in violation or not of the law or of the terms and conditions might open up to a range of grey areas. For the purposes of hard moderation, for instance, AI within hybrid systems usually operates a pre-emptive automated classification of content depending on whether it recognizes it as being acceptable or unacceptable, or, conversely, whether it considers that a human intervention is necessary.[73]

AI represents an invaluable tool for purposes related not only to hard content moderation, but also for purposes related to content curation. Recommender systems, which can be defined as "functions that take information about a user's preferences … as an input, and output a prediction about the rating that a user would give of the items under evaluation",[74] possibly represent, in this respect, the most powerful tools for providers of online platforms to govern the visibility and discoverability of content, as well as to define the ranking thereof. Recommender systems thus play a fundamental role in determining which contents are prioritized and thus presented to users. This is often done, furthermore, through the collection and analysis of data about, *inter alia*, users' preferences, interests,

[71] Kate Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech' (2017) 131 Harvard Law Review 1598, 1635–1649.

[72] Giovanni Sartor and Andrea Loreggia, 'The Impact of Algorithms for Online Content Filtering or Moderation. "Upload Filters"' (European Parliament 2020) JURI Committee PE 657.101.

[73] ibid 22–23.

[74] Silvia Milano, Mariarosaria Taddeo and Luciano Floridi, 'Recommender Systems and Their Ethical Challenges' (2020) 35 AI & Society 957, 957.

and consumption habits,[75] so that the final output is generally the result of complex interactions between the algorithm and the users themselves.[76]

5.3.2. *The rise of automated hate speech moderation*

The deployment of AI systems for the purpose of countering the dissemination of hate speech content across the Internet has become increasingly significant in recent years for a number of reasons. Indeed, through the resort to hybrid moderation systems, it is possible for providers of online platforms to face the extraordinary flow of information by using AI to filter out automatically the most egregious forms of hateful content as well as to help prioritize the items to be subjected to the evaluation of human moderators in cases of doubt. The benefits of AI are manyfold in the context of moderation of harmful content like hate speech and go beyond the simple possibility of addressing larger quantities of information across online platforms. Notably, AI can represent an exceptional instrument to improve the quality of life of human reviewers themselves by preventing them from being exposed to particularly harmful, and potentially traumatizing, content.[77]

Unsurprisingly, AI has come to play an increasingly central role also in the context of hate speech moderation.[78] To underscore how important automated moderation systems have become in this respect, the data disclosed by Meta's Transparency Center shall be

[75] Eleonora Maria Mazzoli and Damian Tambini, 'Prioritisation Uncovered: The Discoverability of Public Interest Content Online' (Council of Europe 2020) DGI(2020)19 38 <https://rm.coe.int/publication-content-prioritisation-report/1680a07a57> accessed 26 May 2022.

[76] "Firstly, users are responsible for uploading content from which content recommenders draw their recommendations. Secondly, users' behaviour provides feedback signals, including explicit feedback such as rating, following or subscribing, as well as implicit feedback such as scrolling and clicking. Since recommender systems commonly rely on machine-learning processes to optimize the algorithm, these user signals can also serve to shape the weighting of the algorithm over time. Conversely, the recommender system can also shape users' behaviour over time, in terms of their preferences, habits and expectations they form in relation to the service. These complex interactions between the recommendation algorithm and its users make for a recursive and unpredictable system, with the potential for unexpected feedback loops and path dependencies". Paddy Leerssen, 'The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems' (2020) 11 European Journal of Law and Technology 4 <https://ejlt.org/index.php/ejlt/article/view/786> accessed 13 November 2023.

[77] Indeed, as highlighted by Roberts, continuous exposure of human reviewers to particularly harmful content (e.g., particularly egregious violent content or hate speech, child pornography, etc.) can affect significantly their health and well-being, leading even to cases of post-traumatic stress disorder: see Sarah T Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media* (Yale University Press 2019). AI offers the possibility to reduce such a collateral effect in many ways, not only by operating itself as a preliminary filter, but also, for example, by allowing reviewers to ask questions about an image without having to view it directly ("visual question answering"): see Cambridge Consultants, 'Use of AI in Online Content Moderation' (Ofcom 2019) 8 <https://www.ofcom.org.uk/research-and-data/online-research/online-content-moderation> accessed 30 August 2023.

[78] See, in this respect, European Union Agency for Fundamental Rights, *Online Content Moderation: Current Challenges in Detecting Hate Speech* (Publications Office 2023) <https://fra.europa.eu/en/publication/2023/online-content-moderation> accessed 8 December 2023.

taken as an example, in light of the rather extensive degree of information rendered public by the company.[79]

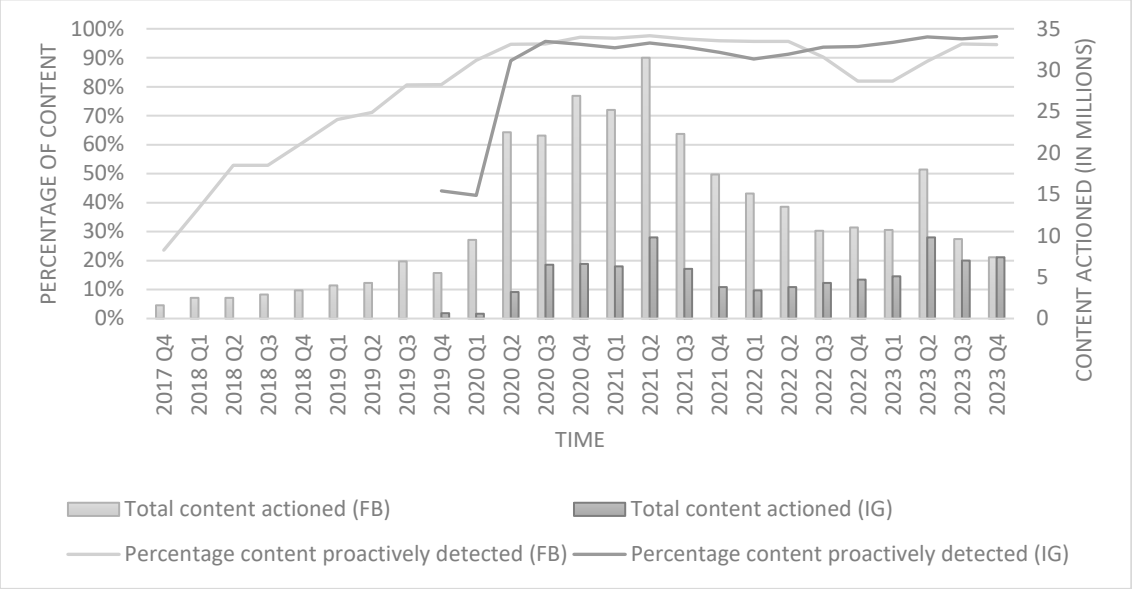

Figure 1. Proactive detection of hate speech on Meta platforms. Source: <https://transparency.fb.com/policies/community-standards/hate-speech/> accessed 28 April 2024.

According to the data disclosed by Meta, as shown in *Figure 1*, hate speech content today is detected in most cases in a proactive way directly by its platforms, mainly thanks to the development of "machine learning technology that automatically identifies content that might violate [their] standards".[80] With respect to Facebook, the increase in the proactive detection of hate speech content was particularly striking between the end of 2017 and the pandemic period of 2020-2021. Indeed, whereas hate speech content detected proactively by the platform represented only 23.6% of all hate speech content actioned in the last quarter of 2017, the figure had risen to 97.1% by the last quarter of 2020. Admittedly, such a percentage has slightly decreased since then: nevertheless, the figure is still significantly higher if compared to the pre-pandemic period. In the case of Instagram, which was acquired by Zuckerberg's company in 2019, hateful content was detected proactively in 44.10% of cases at the end of that year, whereas the percentage had risen to 94.7% by the end of 2020 and has remained particularly high (97.3% in the last quarter of 2023).

Furthermore, a comparison between the data concerning the rate of proactive detection of hate speech and that concerning the number of contents actually actioned by the platforms seems to suggest a partial correlation between the two sets (see *Table 1*).[81]

[79] Meta, 'Community Standards Enforcement: Hate Speech' (*Transparency Center*) <https://transparency.meta.com/reports/community-standards-enforcement/hate-speech/facebook/> accessed 28 April 2024.

[80] Meta, 'Proactive Rate' (*Transparency Center*, 22 February 2023) <https://transparency.fb.com/policies/improving/proactive-rate-metric/> accessed 8 December 2023.

[81] Meta, 'Community Standards Enforcement: Hate Speech' (n 79).

PERIOD	FACEBOOK		INSTAGRAM	
	Variation proactive rate (%)	Variation contents actioned (millions)	Variation proactive rate (%)	Variation contents actioned (millions)
2017 Q4 - 2018 Q1	+14,40%	+0,90		
2018 Q1 – 2018 Q2	+14,90%	+0,00		
2018 Q2 – 2018 Q3	+0,00%	+0,40		
2018 Q3 – 2018 Q4	+7,80%	+0,50		
2018 Q4 - 2019 Q1	+8,00%	+0,60		
2019 Q1 - 2019 Q2	+2,40%	+0,30		
2019 Q2 – 2019 Q3	+9,50%	+2,60		
2019 Q3 – 2019 Q4	+0,10%	-1,40		
2019 Q4 - 2020 Q1	+8,40%	+4,00	-1,60%	-0,07
2020 Q1 – 2020 Q2	+5,60%	+13,00	+46,50%	+2,62
2020 Q2 – 2020 Q3	+0,10%	-0,40	+6,70%	+3,30
2020 Q3 – 2020 Q4	+2,30%	+4,80	-1,00%	0,10
2020 Q4 - 2021 Q1	-0,40%	-1,70	-1,30%	-0,30
2021 Q1 - 2021 Q4	+0,90%	+6,30	+1,70%	+3,50
2021 Q2 – 2021 Q3	-1,10%	-9,20	-1,30%	-3,80
2021 Q3 – 2021 Q4	-0,60%	-4,90	-1,90%	-2,20
2021 Q4 - 2022 Q1	-0,30%	-2,30	-2,30%	-0,40
2022 Q1 - 2022 Q2	+0,00%	-1,60	+1,60%	+0,40
2022 Q2 - 2022 Q3	-5,40%	-2,90	+2,50%	+0,50
2022 Q3 – 2022 Q4	-8,30%	+0,40	+0,20%	+0,40
2022 Q4 - 2023 Q1	+0,10%	-0,30	+1,40%	+0,40
2023 Q1 – 2023 Q2	+6,80%	+7,30	+1,90%	+4,70
2023 Q2 – 2023 Q3	+6,00%	-8,40	-0,70%	-2,80
2023 Q3 – 2023 Q4	-0,30%	-2,20	+0,80%	+0,40

Table 1. Variation in proactive rates and in number of contents actioned. Source: <https://transparency.fb.com/policies/community-standards/hate-speech/> accessed 28 April 2024.

For example, a rise in the number of items acted upon by Facebook, from 10.7 million to 18 million, took place between the first and second quarter of 2023, when the proactive rate rose from 82% to 88.8%. Clearly, such a correlation presents its limits, as many other factors can play a significant role in influencing both figures. Exceptional events such as the COVID-19 pandemic likely led to an outstanding rise both of the quantity of content actioned, due to the alarming spread of hateful content in that period, and of the reliance on automated moderators over human moderators, as many of the latter had been sent home in the aftermath of the breakout of the pandemic and had to be substituted by the former.[82]

[82] Ghadah Alrasheed and Merlyna Lim, 'Beyond a Technical Bug: Biased Algorithms and Moderation Are Censoring Activists on Social Media' (*The Conversation*, 16 May 2021) <http://theconversation.com/beyond-a-technical-bug-biased-algorithms-and-moderation-are-censoring-activists-on-social-media-160669> accessed 18 November 2023.

Be that as it may, the figures appear nevertheless to show how the fine-tuning and heavy reliance on hate speech automated detection systems can deeply affect the overall quantity of content recognized as being in violation of the community's standards. Besides, although such results have been widely presented as key successes by Meta, the reported data do not shed much light on the actual accuracy of the detection systems' output.[83]

5.3.3. *An introduction to automated hate speech detection systems*

From a technical point of view, AI systems for content moderation can follow different approaches and techniques depending on the goal sought and on the type of illegal or harmful materials to be detected and, where necessary, subjected to sanctions.

For example, in some cases, such as the moderation of content infringing copyright or known CSAM, hash-matching strategies have proven to be rather effective. "Hashing" consists of the activity of extracting strings of data from files, each of which corresponds to a defined underlying content. By comparing the hashes of the items uploaded by users with those already stored in a repository, the algorithm can easily match the former with the latter and, where necessary, remove content automatically. However, although hash-matching – due to the fact that hashes are easy to compute – can be deployed on a large scale, it is nevertheless overly sensitive to the slightest changes in the content analysed (e.g., hash-matching systems comparing images may be tricked by simply altering few pixels).[84] Furthermore, matching can be effective in countering content the illegality or harmfulness of which has already been assessed. In the case of hate speech, therefore, other paths have been taken.

Automated hate speech detection generally relies upon classification systems which, as opposed to matching techniques, assess "newly uploaded content that has no corresponding previous version in a database", with the aim "to put new content into one of a number of categories", typically by involving machine-learning and, more recently, deep-learning systems.[85]

5.3.3.1. Classification systems: machine-learning, deep-learning, and natural language processing

Through machine-learning and deep-learning, automated moderation systems are trained to identify content based on statistical patterns. Machine-learning can be either supervised or unsupervised. In the first case, data is manually labelled by humans before being fed to the algorithm, which learns, this way, how to classify content based on the instructions received. In the second case, the machine is trained with unlabelled data and thus

[83] Tom Simonite, 'Facebook's AI for Hate Speech Improves. How Much Is Unclear' (*Wired*, 12 May 2021) <https://www.wired.com/story/facebook-ai-hate-speech-improves-unclear/> accessed 14 December 2021.

[84] Gorwa, Binns and Katzenbach (n 67) 4; Sartor and Loreggia (n 72) 40; Cambridge Consultants (n 77) 48.

[85] Gorwa, Binns and Katzenbach (n 67) 5.

autonomously identifies common patterns underlying that data. Additionally, reinforcement learning consists of giving the machine feedback – either through rewards or penalties – based on its outcomes, as a means to push it to maximize its score.[86]

In recent years, a new branch of machine-learning has undergone extraordinary developments, that of so-called "deep-learning". Deep-learning can process raw data, something which more traditional machine-learning systems struggled to do.[87] This is done, typically, via the deployment of multi-layered neural networks which, designed to mimic the functioning of human neural systems, consist of a series of nodes (neurons) arranged in layers and strictly interconnected. The number of hidden layers implemented impacts directly on the performance of the processing of inputs:[88]

> Deep-learning methods are representation-learning methods with multiple levels of representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level. With the composition of enough such transformations, very complex functions can be learned. For classification tasks, higher layers of representation amplify aspects of the input that are important for discrimination and suppress irrelevant variations … The key aspect of deep learning is that these layers of features are not designed by human engineers; they are learned from data using a general-purpose learning procedure.[89]

The possible applications of such technologies are numerous, encompassing among others image recognition and natural language processing (NLP). Indeed, state-of-the-art NLP classifiers resort to "word embeddings", that is, distributed word representation systems that are based precisely on neural networks and trained with non-annotated *corpora*.[90]

NLP represents an essential application of deep-learning for the purpose of detecting hate speech, especially in its textual form, as it consists of that "discipline of computer science that focuses on techniques for using computers to parse texts", with the typical goal to "predict something of the meaning of the text, such as whether it expresses a positive or negative opinion".[91] Put differently, NLP is "the set of methods for making human language accessible to computers".[92]

5.3.3.2. Training datasets

Two aspects are especially relevant with respect to the development of efficient NLP text classifiers: the datasets used for training and the feature extraction techniques.

---

[86] Sartor and Loreggia (n 72) 37.

[87] Yann LeCun, Yoshua Bengio and Geoffrey Hinton, 'Deep Learning' (2015) 521 Nature 436, 436.

[88] Jenna Burrell, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms' (2016) 3 Big Data & Society 2053951715622512, 5–7.

[89] LeCun, Bengio and Hinton (n 87) 436.

[90] Anna Schmidt and Michael Wiegand, 'A Survey on Hate Speech Detection Using Natural Language Processing' in Lun-Wei Ku and Cheng-Te Li (eds), *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (Association for Computational Linguistics 2017) 2.

[91] Natasha Duarte, Emma Llansó and Anna Loup, 'Mixed Messages? The Limits of Automated Social Media Content Analysis' (*Center for Democracy & Technology*, November 2017) 9 <https://cdt.org/wp-content/uploads/2017/11/2017-11-13-Mixed-Messages-Paper.pdf> accessed 14 December 2021.

[92] Jacob Eisenstein, *Introduction to Natural Language Processing* (MIT Press 2019) 1.

Because their training is generally supervised, meaning that the machine is fed *corpora* of labelled documents that are meant to teach it how to identify hateful content, the creation and use of adequate databases represents an essential aspect. Inadequate data can lead to low-quality training and, inevitably, to low-quality outputs ("garbage in, garbage out").[93] However, the creation of fully representative datasets for hate speech detection is not at all an easy feat, also because of the absence of a clear and universal definition of what hate speech exactly is and because of the existence of grey areas within the process of assessing whether a certain utterance constitutes hate speech or not. Literature has proposed in recent years several sets of labelled and annotated data in this regard, although such proposals still face significant limitations.[94]

For instance, many datasets tend to reflect the subjective perspectives of human annotators, with the consequence of opening the door to involuntary biases.[95] Another challenge arises with respect to the presence of wide gaps in the availability of datasets concerning hate speech in languages different from English and, in general, Western idioms. Furthermore, because data is generally collected, in the first place, directly from social media platforms (which may have different purposes and characteristics, and thus display different subtypes of hate speech),[96] issues of irregularity might emerge as regards, specifically, the percentage of hate speech examples as opposed to non-hate speech examples.[97] As explained by Schmidt and Wiegand,

> annotating hate speech in an extremely time consuming endeavour. There are much fewer hateful than benign comments present in randomly sampled data, and therefore a large number of comments have to be annotated to find a considerable number of hate speech instances. This skewed distribution makes it generally difficult and costly to build a corpus that is balanced with respect to hateful and harmless comments.[98]

5.3.3.3. Feature extraction techniques

Text classifiers can also follow different paths of feature extraction techniques. A feature, which can be defined as "the closed characteristic of an entity or a phenomenon",[99] is in essence an element which allows the automated system to extract insights and patterns in

[93] Indeed, "conclusions can only be as reliable (but also as neutral) as the data they are based on": see Brent Daniel Mittelstadt and others, 'The Ethics of Algorithms: Mapping the Debate' (2016) 3 Big Data & Society 205395171667967, 5.

[94] Francimaria RS Nascimento, George DC Cavalcanti and Márjory Da Costa-Abreu, 'Exploring Automatic Hate Speech Detection on Social Media: A Focus on Content-Based Analysis' (2023) 13 SAGE Open 21582440231181311, 5.

[95] Anusha Chhabra and Dinesh Kumar Vishwakarma, 'A Literature Survey on Multimodal and Multilingual Automatic Hate Speech Identification' (2023) 29 Multimedia Systems 1203, 1220; Ninareh Mehrabi and others, 'A Survey on Bias and Fairness in Machine Learning' (arXiv, 25 January 2022) 3–4 <http://arxiv.org/abs/1908.09635> accessed 21 November 2023.

[96] Schmidt and Wiegand (n 90) 7.

[97] Chhabra and Vishwakarma (n 95) 1221.

[98] Schmidt and Wiegand (n 90) 7. To face these issues, "some studies focused on techniques to deal with the class imbalance problem, some studies focused on techniques to deal with the class imbalance problem, such as oversampling and undersampling. The oversampling technique is applied in the training data to increase the minority class … while the undersampling technique reduces the majority class". Nascimento, Cavalcanti and Da Costa-Abreu (n 94) 5.

[99] Chhabra and Vishwakarma (n 95) 1208.

texts from which the classification of a content as hate speech may be inferred.[100] The choice concerning the feature extraction technique to be followed thus represents a highly influential element in the development of well-functioning detection systems. Besides, because "what differentiates a hateful speech utterance from a harmless one is probably not attributable to a single class of influencing aspects",[101] automated text classifiers may combine together a variety of different techniques.

More basic systems include the bag-of-words approach, by which the machine focuses on the documents composing the training *corpus* as simple collections of single words, so that the order or combination of the words is not taken into account,[102] but, rather, it is the frequency of words that represents the relevant feature in order to classify texts.[103] Such a technique, for instance, has been largely employed by spam filters, which compare the input text with the training documents labelled as spam in order to check for similarities in the words used. However, because the bag-of-words technique ignores the actual sequence of words, it is generally incapable of understanding the semantic and syntactic content of a certain utterance, making it vulnerable to misclassifications and to being eluded by scammers.[104]

To face such limitations, different approaches have been developed, such as N-gram analysis. N-grams are a combination of two or more words – for example, "free speech" is a bi-gram composed of the words "free" and "speech" –, the meaning of which may vary from that of the single words considered separately – thus, "free speech" conveys a whole set of ideas and principles which are much more complex than the simple definitions of the words "free" and "speech". By analysing such combinations of words, rather than the single terms, it is possible for the classifier to evaluate texts in a more nuanced manner, thus understanding, for instance, that the word "queer" may well have a positive connotation (rather than an offensive one) when it is part of the bi-gram "openly queer".[105]

N-gram analysis is often complemented by part-of-speech tagging – that is, the association of a word with its grammatical function within a sentence (e.g., noun, adverb, verb, adjective, etc.) – or parsing – that is, the identification of the syntactic structure of whole clauses and sentences – to improve the machine's syntactic understanding.[106] More advanced strategies, as mentioned above, also include word embeddings, which allow to

[100] See, among others, Nascimento, Cavalcanti and Da Costa-Abreu (n 94) 7. Thus Duarte and others: "Training corpora are pre-processed to numerically represent their features, such as the words, phrases, and grammatical structures that appear in the text. Machine-learning models use these features to learn patterns associated with the targeted content. For example, a spam detection model might learn which words occur more frequently in non-spam examples. Features can range from the simple (individual words) to the more complex (word embeddings …). Complexity here refers to how much of the information in a document the single feature can represent … Newer, state-of-the-art methods rely on more complex 'word embeddings' that take into account the entire sentence or document". Duarte, Llansó and Loup (n 91) 10.

[101] Schmidt and Wiegand (n 90) 2.

[102] "For example, the text 'Bob called Alice' would be represented as the bag of words 'Alice, Bob, called'". Duarte, Llansó and Loup (n 91) 10.

[103] Chhabra and Vishwakarma (n 95) 1209.

[104] ibid; Nascimento, Cavalcanti and Da Costa-Abreu (n 94) 9; Burrell (n 88) 8.

[105] Duarte, Llansó and Loup (n 91) 11.

[106] Nascimento, Cavalcanti and Da Costa-Abreu (n 94) 9; Sartor and Loreggia (n 72) 42.

map out how the words contained in a *corpus* "are related to one another based on the context in which they appear, including their place and function in a document". [107] This leads to more nuanced parsing of language and thus to improvements in the overall efficiency of the classifier.

Further feature extraction techniques can be undertaken, additionally, to enhance the semantic analysis and understanding of texts by NLP classifiers. Semantic analysis may attend to various tasks, including lexical semantics (i.e., ascertainment of the meaning of single words), topic categorization (i.e., ascertainment of the text's subject-matter), natural language understanding (i.e., ascertainment of the meaning of chunks of text), and sentiment analysis (i.e., ascertainment of the positive or negative attitude or polarity expressed).[108]

Sentiment analysis, often also called "opinion mining" is in particular "the field of study that analyzes people's opinions, sentiments, appraisals, attitudes, and emotions towards entities and their attributes expressed in written text".[109] It therefore has many applications in the context of social media, including for marketing as well as opinion monitoring purposes,[110] but may also have important applications in the field of hate speech detection. Indeed, sentiment analysis techniques allow automated hate speech detection systems to grapple directly with people's feelings, opinions, and emotions towards specific "elements" – such as members of protected groups. However, sentiment analysis itself presents its own limits, as it may be able to detect the presence of terms expressing negative feelings while not being able to set those terms within the wider context of a certain speech. Indeed, in this respect, it is important to note that the "presence of negative words or expressions, even in such sentences using the word 'hate', depending on the context, does not make them related to hate speech".[111]

5.3.3.4. Recent developments: large language models

Overall, developers of automated moderation systems have at their disposal a wide range of strategies and techniques which, especially when combined together, can represent an extraordinary asset in the detection and removal, or demotion, of hate speech content. More recently, large language models (LLMs) have also proven to represent highly innovative and effective tools for the detection of hate speech content. The main characteristic of LLMs is that they "are trained on massive amounts of text data and are able to generate

[107] Duarte, Llansó and Loup (n 91) 11.

[108] Sartor and Loreggia (n 72) 42.

[109] Bing Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions* (2nd edn, Cambridge University Press 2020) 1. Similarly, Pozzi and others argue that "the aim of sentiment analysis is to define automatic tools able to extract subjective information from texts in natural language, such as opinions and sentiments, so as to create structured and actionable knowledge to be used by either a decision support system or a decision maker". Federico Alberto Pozzi and others, 'Challenges of Sentiment Analysis in Social Networks: An Overview' in Federico Alberto Pozzi and others (eds), *Sentiment Analysis in Social Networks* (Morgan Kaufmann 2017) 1.

[110] Liu (n 109) 6–9.

[111] Nascimento, Cavalcanti and Da Costa-Abreu (n 94) 10–11.

human-like text, answer questions, and complete other language-related tasks with high accuracy".[112]

These systems clearly represent a double-edged sword in the context of the fight against hate speech across the Internet. On the one hand, LLMs could be abused with a view to augmenting the generation and spread of hate speech content across online platforms. On the other hand, LLMs could, conversely, be implemented to disseminate, rather, counter-speech. Furthermore, LLMs have been found to have important applications in terms of hate speech detection.

For instance, LLMs can help partly overcome issues connected to the classification of content in a variety of languages and, even more, have proven to be capable of recognizing and understanding the meaning of emojis,[113] thus improving the quality of classification *vis-à-vis* multi-modal content.[114] Moreover, LLMs can help develop, automatically, extensive datasets for the training of NLP classifiers, with positive effects, *inter alia*, on the capability of those systems to identify forms of implicit toxic or hate speech.[115]

5.3.4. *Challenges and limitations*

5.3.4.1. The challenges of multi-modality and context

Notwithstanding the extraordinary developments in the context of automated hate speech detection and moderation, the described tools present nevertheless some significant shortcomings, especially in light of the communication dynamics characterizing the Internet and of the paramount role played by context in determining whether a content actually constitutes hate speech or not.

As regards the first aspect, it is worth noting that, while the previous subsection has focused on the use of AI to detect hateful content in the form of texts, communication on the Internet – and particularly upon online platforms – generally takes more complex forms, often combining written, visual, and/or audiovisual materials. Multi-modality of

---

[112] Enkelejda Kasneci and others, 'ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education' (2023) 103 Learning and Individual Differences 102274, 1.

[113] Mithun Das, Saurabh Kumar Pandey and Animesh Mukherjee, 'Evaluating ChatGPT's Performance for Multilingual and Emoji-Based Hate Speech Detection' (arXiv, 22 May 2023) <http://arxiv.org/abs/2305.13276> accessed 23 November 2023.

[114] See *infra*, §5.3.4.1.

[115] An example is represented by Toxigen, a large-scale machine-generated dataset containing both benign and toxic statements: "Detecting implicit toxicity about minority groups … remains an elusive goal for NLP systems … One key challenge is that, in contrast to explicit toxicity, implicit toxicity is not marked by the use of profanity or swearwords, is sometimes positive in sentiment, and is generally harder to detect or collect at scale … A second challenge for detecting subtle toxicity about minority groups is that … minority mentions often co-occur with toxicity labels in datasets scraped from online platforms … With ToxiGen, we aim for generating a *large scale* dataset that represent *implicit* toxicity while *balancing* between toxic and benign statements, to address the gaps of previous work … While valuable, most previous work has relied on scraping data from online platforms, which leads to dataset imbalances with respect to minority-mentioning posts that are toxic vs. benign … In contrast, using large language models to generate our dataset allows us to control the minority groups mentioned in our statements, as well as their implicitness, at larger scale". Thomas Hartvigsen and others, 'ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection' (arXiv, 14 July 2022) 2–3 <http://arxiv.org/abs/2203.09509> accessed 24 November 2023.

content, by merging together such different elements, tends to complicate significantly the task of automated classifiers. Indeed, when a text is combined, for example, with an image or a video, the underlying meaning of a post, as a whole, may well be much different and more complex than the sum of the meanings of the various components. For example, whereas the words "love the way you smell today" and the image of a skunk would not *per se*, if taken singularly, have any negative meaning, the association of the two could result in deeply offensive content.[116] In this respect, the understanding of the semantic meaning of online content is especially burdensome with respect to phenomena typical of viral Internet such as so-called "memes".[117] Although significant progress has been made in this regard, the issue of multi-modality still represents a highly significant challenge from a technical point of view that can seriously affect the accuracy of automated classifiers.[118]

With respect to the second aspect, the paramount role of context for assessing whether a certain content constitutes hate speech or not has often been stressed not only by institutional actors, especially at the international and supranational levels,[119] but also by providers of online platforms themselves (as well as by Meta's Oversight Board).[120] Indeed, a certain utterance may well have different meanings and effects depending on a number of variables, including the identity of the speaker, the dimension and composition of the audience, the time, and the place:

> For instance, [the phrase "Put on a wig and lipstick and be who you really are"] may not be regarded as some form of hate speech when only read in isolation … However, when the context information is given that this utterance has been directed towards a boy on a social media site for adolescents, one could infer that this is a remark to malign the sexuality or gender identity of the boy being addressed … The above example shows that whether a message is hateful or benign can be highly dependent on world knowledge, and it is therefore intuitive that the detection of a phenomenon as complex as hate speech might benefit from including information on aspects not directly related to language.[121]

In order for an AI system to be able to consider such variables, a path that has been followed is that of augmenting the machine's knowledge base, so that more nuanced decisions may be taken. However, such approaches require the inclusion within the code of domain-specific assertions, making such augmentation quite burdensome.[122] Issues concerning the availability of sufficient knowledge bases, furthermore, seem to represent

---

[116] Douwe Kiela and others, 'The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes' [2021] arXiv:2005.04790 [cs] 1–2 <http://arxiv.org/abs/2005.04790> accessed 14 December 2021. See also Raul Gomez and others, 'Exploring Hate Speech Detection in Multimodal Publications', *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2020).

[117] Memes are multi-modal forms of expression, often characterized by a satirical or ironic intent, which consist precisely of a combination of images and text the understanding of which generally requires the reader to merge the two components and, often, to have a preliminary knowledge of the meme's structure (meme "literacy") in order to understand it. See Gabriele Marino, 'Semiotics of Spreadability: A Systematic Approach to Internet Memes and Virality' (2015) 1 Punctum 43, 43.

[118] Paulo Cezar de Q Hermida and Eulanda M dos Santos, 'Detecting Hate Speech in Memes: A Review' (2023) 56 Artificial Intelligence Review 12833.

[119] See *supra*, §2.2.2.

[120] See *supra*, §5.2.1.

[121] Schmidt and Wiegand (n 90) 4–5.

[122] ibid 5.

a limitation also of systems based on LLMs, such as ChatGPT, which has been found to have difficulties in distinguishing, for example, hate speech from instances of counter-speech.[123]

Additionally, the analysis of meta-information, especially that concerning the background and identity of users, has been considered to represent in theory a significant piece of information to help categorize correctly a certain content: for instance, the use of slurs relating to racial or sexual minorities will likely have different connotations based on whether the speaker is a member of those minorities or, rather, is a person connected to alt-right networks. However, the processing of such information clearly raises important concerns in terms of privacy and data protection rights. Moreover, it has been highlighted that training a system with such information could, in fact, lead to unwarranted consequences and lead an AI system to "solve the wrong puzzle or learn based on wrong knowledge from the data" and thus to originate biased outputs.[124]

5.3.4.2. Automated moderation and biases

In fact, also as a result of the inherent difficulties encountered by AI systems in evaluating the semantic meaning and context of content posted online, automated moderation strategies have often been found to lead to biased outcomes.[125] Most notably, research has shown that automated hate speech – or, more generally, "toxic" speech – detection systems can have the effect of removing precisely the content produced by minority groups or categories of people traditionally subject to discrimination and marginalization.[126]

Thus, for instance, many studies have shown how the African American community is more easily subject to having their contents being detected as either "toxic" or "hateful", and therefore to receiving moderation sanctions. This is in good part due to the fact that the datasets used to train algorithmic moderators are often not sufficiently representative of African American English, so that the machine is confused and misinterprets the use of terms and expressions that, without an appropriate understanding of the identity of the speaker and of their intentions, might be interpreted as being offensive.[127]

---

[123] Yiming Zhu and others, 'Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks' (arXiv, 22 April 2023) <http://arxiv.org/abs/2304.10145> accessed 23 November 2023; Das, Pandey and Mukherjee (n 113).

[124] Nascimento, Cavalcanti and Da Costa-Abreu (n 94) 11; Sean MacAvaney and others, 'Hate Speech Detection: Challenges and Solutions' (2019) 14 PLOS ONE e0221152, 7.

[125] In fact, AI technologies have been found in recent years to raise significant concerns in terms of the possibility of biased and discriminatory outcomes and have, as such, led to the need for the law to rethink its traditional anti-discrimination strategies. On the specific aspects characterizing AI-driven discrimination see, among others, Costanza Nardocci, 'Artificial Intelligence-Based Discrimination: Theoretical and Normative Responses. Perspectives from Europe' (2023) 60 DPCE Online 2367, 2372–2376.

[126] Oliver L Haimson and others, 'Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas' (2021) 5 Proceedings of the ACM on Human-Computer Interaction 466:1.

[127] Thomas Davidson, Debasmita Bhattacharya and Ingmar Weber, 'Racial Bias in Hate Speech and Abusive Language Detection Datasets' in Sarah T Roberts and others (eds), *Proceedings of the Third Workshop on Abusive Language Online* (Association for Computational Linguistics 2019); Maarten Sap and

Similarly, Oliva, Antonialli, and Gomes's experiment on the functioning of Google's Perspective API represents a clear example of how AI detection systems could easily misrepresent content produced by the LGBTQIA+ community. The authors, in particular, by comparing the results obtained when processing tweets published by famous US drag queens and those obtained when processing tweets by US far-right white supremacists, found that the algorithm recognized the tweets by the former as being "toxic" at a rather alarming rate.[128] Gender studies have shown that biased moderation can also affect disproportionately content produced by women.[129] Hate speech moderation systems have even been found to internalize ableist biases in language, thus charging with negative connotations any disability-related term – as a result, a machine will more likely consider as toxic phrases such as "I am a blind person" or "I am a deaf person" than statements like "I am a person" or "I am a tall person".[130]

Moreover, the reproduction of biases has been known to represent a challenge not only for hard content moderation practices, but also for content curation practices. As highlighted by Noble, biased AI can lead to the creation of proper "algorithms of oppression".[131] A clear example of this is the case of online search engines presenting images charged with sexual innuendo whenever queried with keywords such as "black girls", or the case of online market platforms reducing the visibility of African-American-owned businesses.[132] With respect to content curation on social media platforms, representatives of minority groups and of discriminated or marginalized communities have repeatedly argued that most platforms' recommendation systems often have the tendency to reduce

---

others, 'The Risk of Racial Bias in Hate Speech Detection' in Anna Korhonen, David Traum and Màrquez Lluís (eds), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics 2019) <https://www.aclweb.org/anthology/P19-1163> accessed 22 October 2021.

[128] Thiago Dias Oliva, Dennys Marcelo Antonialli and Alessandra Gomes, 'Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online' (2021) 25 Sexuality & Culture 700. As a matter of fact, queer linguistics and studies on communication styles typical of queer people have highlighted the typical resort to "mock impoliteness", i.e. to the use of expressions and terms which may appear as offensive at a first glance (e.g., the use of the F-word) but are ultimately accepted when used by the members of the LGBTQIA+ or drag community as a system to help each other "build a thick skin" and, to a certain extent, reduce the negative charge of an insulting term. These linguistic practices, often referred to as "reading", can easily confuse algorithms that may lack a sufficient knowledge base. See, with respect to the practices of "reading" within the LGBTQIA+ community (and especially within the drag community), Sean McKinnon, '"Building a Thick Skin for Each Other": The Use of "Reading" as an Interactional Practice of Mock Impoliteness in Drag Queen Backstage Talk' (2017) 6 Journal of Language and Sexuality 90.

[129] Carolina Are, 'How Instagram's Algorithm Is Censoring Women and Vulnerable Users but Helping Online Abusers' (2020) 20 Feminist Media Studies 741; Ysabel Gerrard and Helen Thornham, 'Content Moderation: Social Media's Sexist Assemblages' (2020) 22 New Media & Society 1266.

[130] Ben Hutchinson and others, 'Unintended Machine Learning Biases as Social Barriers for Persons with Disabilities' (*SIGACCES*, October 2019) <http://sigaccess.org/newsletter/2019-10/hutchinson.html> accessed 12 April 2022.

[131] Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York University Press 2018).

[132] ibid.

significantly the visibility of the counter-speech content they publish, sometimes leading even to forms of "shadowbanning".[133]

The issue concerning the possible biased shadowbanning of counter-narratives is, furthermore, complicated by the lack of transparency as to the functioning of recommender systems[134] and by the inherent hurdles hampering the possibility to actually map out the entity of the phenomenon. Indeed, because the reduction of visibility or discoverability of a certain content is a much less "evident" sanction than that of content removal, research often struggles with the acquisition of data offering clear insights into how much automated content curation impacts minority voices.

Admittedly, research in the field of AI has managed to develop a range of techniques aimed at mitigating the collateral effects connected to biased hate speech classifiers. Debiasing strategies can be implemented at various stages in the development of those systems and have contributed consistently to the improvement of automated hate speech detection. However, despite the encouraging results already yielded, debiasing tools still seem to face important challenges and shortfalls.[135]

5.4. Algorithmic errors and fundamental rights

5.4.1. *The inevitability of error*

The hate speech policies adopted by online platforms, as well as the ways in which those policies are enforced in practice, are clearly highly relevant with regard to the enjoyment of fundamental rights and liberties upon the Internet. In particular, the implementation of AI systems for hate speech detection necessarily entails the search for a balance between two poles of interests: removal of unwarranted content, on the one hand, and protection of freedom of expression (and the right to equality), on the other hand.

Because most automated classifiers, including hate speech detecting systems, are ultimately based on statistical and probabilistic factors, a certain error rate is always inevitable. Besides, errors include both false negatives – namely, hate speech content not perceived as such – and false positives – namely, content wrongly categorized as hate speech. As pinpointed by Sartor and Loreggia, it is generally possible to reduce the false negatives rate only by increasing the false positives rate, and vice versa, as the two rates are, in

[133] Gabriel Nicholas, 'Shedding Light on Shadowbanning' (*Center for Democracy and Technology*, April 2022) <https://cdt.org/insights/shedding-light-on-shadowbanning/> accessed 22 July 2022; Carolina Are, 'The Shadowban Cycle: An Autoethnography of Pole Dancing, Nudity and Censorship on Instagram' (2022) 22 Feminist Media Studies 2002.

[134] Leerssen (n 76).

[135] Ji Ho Park, Jamin Shin and Pascale Fung, 'Reducing Gender Bias in Abusive Language Detection', *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics 2018); Xuhui Zhou and others, 'Challenges in Automated Debiasing for Toxic Language Detection' in Paola Merlo, Jorg Tiedemann and Reut Tsarfaty (eds), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (Association for Computational Linguistics 2021); Tanmay Garg and others, 'Handling Bias in Toxic Speech Detection: A Survey' (2023) 55 ACM Computing Surveys 264:1.

practice, inversely proportional.[136] Therefore, platforms find themselves in the situation of having to choose between reducing the number of false negatives or, instead, the number of false positives. In the first case, the result would be that of under-blocking the uploading and dissemination of hate speech content, to the possible detriment of a variety of rights of targeted communities. In the second case, the main risk is that of the over-blocking of content, to the possible detriment, specifically, of the rights to freedom of expression and non-discrimination.[137]

As a result, as highlighted by Douek[138], the question is not so much whether AI systems should or should not be used *vis-à-vis* the concrete possibility of errors being made, but, rather, whether it is in fact possible to identify a threshold of "acceptability" of error. In other words, the challenge which platforms, and in turn the law, must face is specifically that of identifying to what extent it is possible to accept that a number of legitimate contents may be removed in order to ensure that a lower number of unlawful or harmful contents are removed or, vice versa, to what extent it is admissible that unlawful or harmful contents are erroneously maintained online due to the implementation of strategies more focused on limiting to the maximum extent possible the number of false positives.

> Content moderation systems must accept error. Indeed, they decide to get it wrong sometimes and must decide in which direction to do so. The choice to get it wrong in some portion of cases is the price of getting it right, within a reasonable timeframe (or at all), in the vast majority of cases … The question cannot be whether there are *instances* of false positives … or false negatives … the answer to that question will *always* be "yes". The more pertinent questions are: What error rates are acceptable in enforcing a ban …? … To be clear, there may be cases where error costs are too great, or where the probabilities cannot be made to fall within an acceptable range. Accepting errors does not mean accepting *all* errors.[139]

Clearly, the answers to questions concerning the threshold of acceptable error rates will vary significantly not only depending on the type of content the moderation system aims to remove from the Internet, but also on the value-framework adopted by the platform. Furthermore, the law can play a central role in this respect, as the governmental regulation of speech and content moderation inherently contributes to shaping platforms' attitudes towards illegal and/or harmful content and, therefore, to defining more or less directly those thresholds. As a result, different false positives rates may be perceived as overall acceptable depending on whether, for example, the wish is that of countering copyright infringement, disinformation, or child pornography, and on the relevant (value-laden)[140] constitutional framework.

---

[136] Sartor and Loreggia (n 72) 45.

[137] European Union Agency for Fundamental Rights, *Bias in Algorithms: Artificial Intelligence and Discrimination* (Publications Office 2022) 50 <https://data.europa.eu/doi/10.2811/25847> accessed 3 February 2023.

[138] Evelyn Douek, 'Governing Online Speech: From "Posts-as-Trumps" to Proportionality and Probability' (2021) 121 Columbia Law Review 759.

[139] ibid 808–809, 813.

[140] Oreste Pollicino, 'The Quadrangular Shape of the Geometry of Digital Power(s) and the Move towards a Procedural Digital Constitutionalism' (2023) 29 European Law Journal 10.

In light of such remarks, it is apparent that, in the context of hate speech governance, the paramount challenge ahead for regulators shall be precisely that of dealing with the inevitability of error in the algorithmic moderation of hateful content, by focusing on strategies to identify what should be considered as acceptable (or rather, conversely, as unacceptable) error. Following the arguments brought to the table throughout the present work and, especially, within Chapter 2,[141] the notion of substantive equality may arguably represent a notable lens to address this point.

5.4.2. *Acceptable errors and substantive equality*

Taking the promotion of substantive equality as the core target of hate speech governance can indeed serve as an insightful proxy to identify what types of errors – especially in terms of false positives – should be considered as being generally acceptable or unacceptable and thus as a purposeful tool to help design better hate speech policies.

Substantive equality encompasses, *inter alia*, the goal of further empowering traditionally marginalized and discriminated groups of people, notably by incentivizing their active participation in society and by encouraging the adaptation of social practices with a view to meeting their specific needs. As previously argued, applying these objectives in the context of the moderation of hateful content by platforms necessarily translates into the need to ensure that the removal of such content is accompanied by strategies specifically intended to guarantee that those categories of people that are mostly targeted by hate speech are, in turn, fully able to enjoy their right to freedom of expression and to engage, namely, in forms of counter-speech aimed at deconstructing the oppressive power dynamics historically suffered.[142]

That being the case, the necessary conclusion is that, when assessing what false positive rate is acceptable in the context of automated hate speech detection and moderation, a distinction ought to be made between the error rate affecting speech uttered by historically dominant groups and that affecting speech uttered by historically dominated groups, especially in cases where the latter address topics related, precisely, to the phenomenon of discrimination. Indeed, the wrongful censorship of speech published by this second category of people would translate into a short-circuit whereby the attempt to limit hate speech, which should be driven by the effort to protect victimized groups and individuals, ends up silencing and disempowering them. In this respect, platforms – and the law – should thus "consider adopting the non-discriminatory measures a priority" and giving special attention to "the way algorithms may disproportionately render minorities' speech toxic".[143]

Consistently, the bar of acceptance, when it comes to error rate acceptability, should be especially high when it comes to the protection of minority, discriminated, or

[141] See *supra*, §2.5.2.
[142] See *supra*, §2.5.3.
[143] Thiago Dias Oliva, 'Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression' (2020) 20 Human Rights Law Review 607, 639.

marginalized groups that are targeted by hate speech itself. Moreover, erroneous limitations affecting cases of counter-speech should be considered as particularly unacceptable.

As a matter of fact, such a perspective seems to be in line not only with the trends concerning hate speech governance that are emerging within the case law of the ECtHR and the legislative strategies of the EU,[144] but is also in good part consistent with the opinions delivered by Meta's OB for the *Wampum Belt*[145] and *Reclaiming Arabic Words*[146] cases.[147] Although online platforms are still largely invested in formalistic approaches to equality when it comes to the moderation of hate speech,[148] those decisions, coupled with the presence in most providers' terms and conditions of rules concerning, for example, the availability of exemptions for the self-referential use of slurs, indicate at least the awareness of private platforms of the importance of supporting the full online participation of minorities and discriminated or marginalized groups. As will be argued below, this aspect, loosely coinciding with the "participative dimension" of substantive equality, could represent a solid starting point for a cooperative collaboration between private digital platforms and EU institutions.

Furthermore, it must be stressed that these arguments do not apply only with regard to practices of "hard moderation". Rather, they should be extended also to the way automated content curation is deployed. Indeed, preventing targeted groups' contents from being easily discoverable and findable by users has detrimental effects that are often equivalent to those produced by the actual removal thereof. Therefore, if one were to take substantive equality as a proxy, unfair and biased outputs of content curation systems limiting the reachability of counter-speech items should be seen to constitute precisely instances of algorithmic error in overt contrast with the inherent goals of hate speech governance.

Besides, the remark that current automated systems of content moderation and content curation are inherently subject to error – and, in many cases, tend to replicate biases as well as ancient dynamics of oppression, discrimination, and silencing – should not lead to the conclusion that the deployment of those AI systems should be rejected outright. Nowadays, such a conclusion, given the sheer dimensions of the Internet phenomenon, would be utterly impracticable. Rather, recognizing through the lens of substantive equality that such mistakes are generally unacceptable entails, on the one hand, accepting that a certain number of errors may indeed take place and affect individual rights while, on the other hand, focusing on the need to design systems aimed specifically at reducing the negative effects that those errors might have at an aggregate and systemic level.[149]

---

[144] See *supra*, §2.5.2.2.

[145] *Wampum belt* (n 37).

[146] *Reclaiming Arabic words* (n 40).

[147] See *supra*, §5.2.1.3.

[148] See *supra*, §5.2.3.

[149] Indeed, as argued by Douek, the "sheer scale and diversity of online speech belies thinking through the traditional frame of each individual case. Content moderation is inherently systemic. Content moderation systems do not promise to get every individual speech decision right; they are designed to increase the probability that most decisions will be right most of the time and when the system errs, it does so in a

Such a perspective should orient the future actions not only of the private online platforms but also those of public policymakers and lawmakers. Indeed, as stressed by Alkiviadou, "at the very least, the principles and precepts of international human rights law and the thresholds attached to Article 20(2) ICCPR, as further interpreted by the Rabat Plan of Action, must inform and guide any effort in enhanced platform liability".[150] On top of those principles and precepts, the values and implications connected to the principle of substantive equality could and should play an important additional role.

5.4.3. *Mitigating the impact of errors: areas of action*

In order to address the concerns related to high rates of false positives caused by biased automated moderation systems, a multi-faceted approach should be followed, with a view to dealing both with the goal of reducing – at a systemic level – those rates and of ensuring that adequate redress systems are ensured to individual users to protect their fundamental rights.

First, an important issue is represented by the question of transparency in decision-making. As is well known, the matter of transparency is especially problematic when one deals with AI systems that are based on machine-learning or deep-learning techniques such as those usually employed for hate speech moderation. From a technical point of view, these systems generally work as black boxes, meaning that the way they interpret and classify inputs is not comprehensible to human beings (including even the programmers of the system themselves).[151] It may thus be particularly burdensome to obtain information concerning the reasons, and possibly the biases at play, behind specific automated moderation and curation decisions.

However, the effects of the use of certain AI moderation systems could be more effectively understood when the analysis is conducted at a more generalized and systemic level, for example through research based upon observational studies or practices such as "black box tinkering".[152]Attention should therefore be paid to fostering cooperation between private platforms, public institutions, and researchers with a view to promoting a better understanding at the collective, rather than merely individual, level of the impact of automated hate speech moderation systems upon the freedom of expression of minority, marginalized, and discriminated groups. Furthering transparency with regard to this aspect is a paramount objective which would contribute not only to ensuring that platforms are made accountable for the anti-equalitarian aggregated effects caused by the AI

preferred direction. a systemic approach accepts the inevitability of errors and factors them into governance design". Douek (n 138) 790–791.

[150] Natalie Alkiviadou, 'The Internet, Internet Intermediaries and Hate Speech: Freedom of Expression in Decline?' (2023) 20 SCRIPTed 243, 268.

[151] Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2015); Burrell (n 88); Giovanni De Gregorio, 'Democratising Online Content Moderation: A Constitutional Framework' (2020) 36 Computer Law & Security Review 105374, 14.

[152] Maayan Perel and Niva Elkin-Koren, 'Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement' (2017) 69 Florida Law Review 181.

systems they employ, but also to allowing fruitful research aimed at the technical fine-tuning and improvement of those systems.

A second aspect which is of primary importance is that concerning the necessary injection, within the moderation cycle, of principles connected to the rule of law and to the due process of law.

As regards matters concerning the rule of law, it has to be noted that the ways in which platforms govern speech on the Internet often lack consistency with basic tenets of democratic ideals: suffice it to bear in mind how, when it comes to content moderation, platforms bear in practice forms of unilateral (quasi-)legislative, executive, and judicial powers.[153] Overcoming the lack of incorporation of rule of law principles[154] and promoting forms of "democratization"[155] represent important challenges in the age of platform governance, as they represent an inescapable premise for ensuring the full protection of fundamental rights and freedoms within the digital sphere. When it comes to hate speech, the promotion of a more democratic approach could (and should) entail first and foremost the inclusion, within the governance of the phenomenon, of the members of those groups that are most often victimized by it. In this regard, for instance, platforms should not only ensure diversity within the teams responsible for the development of moderating systems, but also maintain a fruitful dialogue with research institutions and organizations advocating for the rights of social minorities.[156]

With regard to due process, it is essential not only that moderation practices are deployed following appropriate procedures and, where necessary, include the intervention of human beings, but also that efficient redress systems are put in place. Even though, inevitably, a certain threshold of errors may have to be accepted, this does not mean that the lack of adequate solutions to tackle those situations where an error has been made is, itself, acceptable:

> A corollary of error acceptance is the need for a way to challenge and rectify mistakes. Mistakes are inevitable, but not always acceptable. No doubt part of the reason platforms do not openly acknowledge their error choices is because they have failed to build adequately robust systems for error correction. The failure to provide adequate procedural checks is not separate but related to the dissatisfaction with the substantive rules. Mistakenly removing [legitimate content] might be more readily acceptable, for example, if a reliable process existed for ensuring such mistakes were indeed temporary rather than relying on media outrage to force reversals.[157]

In this respect, specific redress systems could be envisaged for the submission of complaints concerning the unwarranted restriction of speech of historically oppressed communities and, specifically, of cases of counter-speech. For example, forms of intervention by associations advocating for minority rights in support of complainants could be

[153] Quintais, De Gregorio and Magalhães (n 58).

[154] Nicolas P Suzor, 'Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms' (2018) 4 Social Media + Society 2056305118787812.

[155] De Gregorio, 'Democratising Online Content Moderation' (n 151).

[156] Oliva (n 143) 639.

[157] Douek (n 138) 825.

devised, with a view to helping mitigate the asymmetry of powers between the individual user and the private owner of digital platforms.

Overall, the inevitability and need to accept a certain margin of error in the processes of detection, moderation, and curation of hate speech content would require in turn platforms and the law to intervene with a view to supporting the replication of basic constitutional values – historically well-established in the relational dynamics between private individuals and the state – also within the relational dynamics between private individuals and the platforms themselves. In this respect, the paradigm of thought represented by "digital constitutionalism",[158] hereby intended as the recalibration of constitutional principles and values in light of the new (private) digital powers of the algorithmic age,[159] may well represent a starting point to help shape future policy strategies to deal with the important challenges ahead.

Admittedly, the adoption of legal strategies related to such constitutional aspirations may face significant hurdles in jurisdictions where lawmakers and courts have up to now proven to be sceptical with regard to the possibility of associating the power held by ISPs with that held by the state. In this respect, an emblematic case is certainly represented by the US framework where, as described throughout Chapter 4, the configurability of social media platforms as state actors has up to now been rejected[160] and where the First Amendment has been found by several courts to bar the introduction of statutory rules aimed at conforming providers' terms and conditions[161] or the ways in which those terms and conditions are actually enforced.[162]

Conversely, constitutional aspirations seem to emerge from a plurality of legislative initiatives across other jurisdictions, especially in Europe. Moves have arguably been made in this direction, for example, through the German NetzDG, the UK's OSA, and the EU's DSA, all of which set limits to the unilateral power of platforms to design and enforce terms and conditions of service.[163] The following Section will focus, precisely, on the EU framework in the wake of the DSA.

---

[158] On the many understandings of the term and on the many directions undergone by research on "digital constitutionalism", see among others Edoardo Celeste, 'Digital Constitutionalism: A New Systematic Theorisation' (2019) 33 International Review of Law, Computers & Technology 76. In fact, as highlighted by De Gregorio, "merging the expressions 'digital' and 'constitutionalism' does not lead to revolutionising the pillars of modern constitutionalism. Instead, it aims to understand how to interpret the (still hidden) role of constitutional law in the algorithmic society. Therefore, digital constitutionalism should be seen not as a monolith but as the expression of different constitutional approaches to digital technologies from an internal and external point of view". Giovanni De Gregorio, *Digital Constitutionalism in Europe: Reframing Rights and Powers in the Algorithmic Society* (Cambridge University Press 2022) 4–5.

[159] Indeed, as underscored by Pollicino, following the shift from the world of atoms to the world of bits, "power is relocated among different actors in the information society", and thus "constitutionalism becomes 'digital constitutionalism'". See Oreste Pollicino, *Judicial Protection of Fundamental Rights on the Internet: A Road Towards Digital Constitutionalism?* (Hart 2021) 190.

[160] *Prager University v Google LLC* 951 F3d 991 (9th Cir 2020). See *supra,* §4.4.3.

[161] *Volokh v James* 2023 WL 1991435 (SDNY 2023). See *supra*, §4.4.4.2.

[162] *NetChoice, LLC v Paxton* 49 F4th 439 (5th Cir 2022); see, *contra*, *NetChoice, LLC v Moody* 34 F4th 1196 (11th Cir 2022). See *supra*, §4.4.4.1.

[163] See *supra*, §§4.2.1, 4.3, 3.5.

5.5. Algorithmic hate speech moderation in Europe: constitutional challenges and substantive equality

5.5.1. *Constitutional aspirations of the Digital Services Act*

The inherent aspirations of digital constitutionalism, as interpreted above, are in fact deeply embedded in the recent wave of EU legislation concerning digital policies.[164] The DSA, in particular, represents a paradigmatic example of digital constitutionalism precisely because, by regulating content moderation practices, it directly aims to orient and set limitations to the power of private platforms to "shape the boundaries of freedom of expression", thus showcasing "the resilience of the European constitutional model reacting to the threats of private powers in the information society".[165]

Moreover, the new system envisaged by the DSA is at least in its intentions coherent with the lines of action suggested in the previous subsection, as the Regulation includes vast sets of rules that clearly seek to foster transparency, rule of law, and due process values within the domain of content moderation and content curation. Fundamentally, the focus of the DSA is to "constitutionalize" the moderation and curation practices of providers of hosting services and, especially, of online platforms, VLOPs, and VLOSEs. This having been dealt with more in depth throughout Chapter 3,[166] only a brief overview of the most relevant provisions will be given at this stage.

A first, clear example of the DSA's attempt to incorporate the constitutional and democratic principles within the action of all providers of intermediary services is represented by the duty to specify, within their terms and conditions – which must be "publicly available in an easily accessible and machine-readable format" (transparency) –, all relevant information concerning "any restrictions that they impose in relation to the use of their service in respect of information provided by the recipients of the service" (rule of law) and clarifying "procedures, measures and tools used for the purpose of content moderation, including algorithmic decision-making and human review, as well as the rules of procedure for their internal complaint handling system" (due process).[167] Indeed, through this provision, the DSA strives to ensure that the deployment of content moderation and curation activities is not arbitrary, but, rather, follows pre-determined rules that are enforced following pre-determined procedures with a view to respecting the "rights and legitimate interests of all involved, including the fundamental rights of the recipients of the service".[168]

In terms of transparency, moreover, these requirements are complemented by the obligation to publish transparency reports periodically, with a view to allowing public

[164] Jan Czarnocki, 'Saving EU Digital Constitutionalism through the Proportionality Principle and a Transatlantic Digital Accord' (2021) 20 European View 150, 152–153.

[165] Giovanni De Gregorio, 'The Digital Services Act: A Paradigmatic Example of European Digital Constitutionalism' (*Diritti Comparati*, 16 May 2021) <https://www.dirittocomparati.it/the-digital-services-act-a-paradigmatic-example-of-european-digital-constitutionalism/> accessed 4 December 2023.

[166] See *supra*, §3.5.3.

[167] DSA art 14, para 1.

[168] ibid 14, para 4.

scrutiny over the ways those terms and conditions are, in fact, applied.[169] Providers of online platforms, VLOPs, and VLOSEs must also set out in their terms and conditions information about the functioning of recommender systems used and give the recipients of services a certain degree of choice as regards the ways in which content is presented to them.[170] In the case of VLOPs and VLOSEs, furthermore, an important additional tool has been devised to foster transparency and cooperation between private platforms, public institutions, and the academia, that is, the explicit recognition of the possibility for national DSCs to request such platforms to provide either themselves or "vetted researchers" with data concerning their moderation activities.[171]

As regards due process requirements set as guarantees for service recipients' fundamental rights, the Regulation includes the obligation for all providers of hosting services to deliver a clear and specific statement of the reasons behind the adoption of any restrictive measure – including demotion and/or demonetization of content –, also indicating whether that measure was adopted through the use of AI.[172] This obligation is not intended only to further promote transparency but, rather, has the practical function of allowing Internet users to be able to raise substantiated complaints against the decision taken. This is confirmed, indeed, by the requirement that the statement of reasons contains information about the possible avenues for redress. As regards the latter, the DSA envisages, in particular, the obligation for online platforms, VLOPs, and VLOSEs to establish effective internal complaint-handling systems. Upon such complaints, providers of those platforms are required to deal with them in a timely, non-discriminatory, diligent and non-arbitrary manner, and are required to render decisions that are not taken solely on the basis of automated means.[173] Moreover, recipients of services shall be given the possibility to refer the matter to out-of-court dispute settlement bodies, free of charge.[174]

Finally, with respect to VLOPs and VLOSEs, the choice to insert within the approved text of the DSA a clause clarifying that the mechanisms employed for the purposes of assessing and mitigating systemic risks must have particular consideration of the actual collateral impact that mitigating measures could have on the fundamental rights of users[175] represents a further attempt to reconcile the DSA's aspirations to counter illegal and harmful content with the basic tenets of the rule of law.

5.5.2. *A renovated Code of Conduct on Hate Speech?*

As pointed out in Chapter 3,[176] a characteristic feature of the DSA is the choice to develop a legal framework on moderation that is applicable, horizontally, to all kinds of relevant illegal and harmful contents. Such a choice is to be welcomed as it allows for a systematic

[169] ibid 15, 24, 42.
[170] ibid 27, 38.
[171] ibid 40.
[172] ibid 17.
[173] ibid 20.
[174] ibid 21.
[175] ibid 35, para 1.
[176] See *supra*, §3.5.1.

and coherent regulation of the policies, strategies, and practices enforced by regulated ISPs. As a matter of fact, such a general scope of applicability further allows the range of provisions set by the DSA to truly reach its aspired constitutional dimension.

At the same time, however, the general and horizontal approach of the DSA necessarily entails a lack of specificity with regard to the particular regulatory needs each kind of "information bad"[177] entails. Indeed, depending on the type of content one seeks to counter, different principles and values necessarily enter into consideration. The removal of copyright-infringing material must be carried out taking into account ethical, legal, and technical premises that are much different from those to be considered when fighting, for example, the dissemination of CSAM or of terrorist content. This point is clear and evident to the lawmakers of the EU, as emerges from the adoption of legislative acts such as the DSM Copyright Directive and the TERREG, or from the CSAM Regulation proposal.[178] These acts complement the DSA by setting specific rules governing the moderation of specific types of illegal content to be removed. Similarly, an appropriate governance of the hate speech phenomenon would benefit from the definition of more specific criteria and principles.

However, the recognition of the importance of complementary sectorial tools does not necessarily mean, *per se*, that such tools should always be of a legislative nature. Rather, the DSA itself, through its novel rules on the role of codes of conduct, has opened the way for a more significant implementation of co-regulatory means.[179]

5.5.2.1. DSA, co-regulation, and hate speech

As described in Chapter 3,[180] Article 45 of DSA introduces indeed the possibility for the Commission to invite providers of VLOPs and VLOSEs, together with any other relevant providers of intermediary services or stakeholders, to participate in the drawing up of codes of conduct directed at addressing certain common systemic risks. These codes, in particular, shall set out clearly specific objectives, taking into account the interests of all parties affected, and identify key performance indicators to allow the measuring of results, under the control of the Commission and the newly created EBDS. By drawing up such codes, it is possible to introduce more specific and tailored commitments[181] which, while being designed not through a top-down regulatory strategy but, rather, via bottom-

[177] Sartor and Loreggia (n 72).

[178] Respectively, 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 (DSM Copyright Directive), OJ L 130/92; Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online, OJ L 172/79; European Commission, 'Communication of 11 May 2022, Proposal for a Regulation of the European Parliament and of the Council Laying down Rules to Prevent and Combat Child Sexual Abuse' COM(2022) 209 final. See *supra*, §3.4.3.2.

[179] DSA art 45.

[180] See *supra*, §3.5.3.5.

[181] Rachel Griffin and Carl Vander Maelen, 'Codes of Conduct in the Digital Services Act: Exploring the Opportunities and Challenges' (SSRN, 30 May 2023) 7 <https://papers.ssrn.com/abstract=4463874> accessed 14 June 2023.

up negotiations with the interested parties, tend to acquire, through the framework introduced by the DSA, a rather compelling normative force.

In other words, under the DSA, codes of conduct undergo a transformation from being mere self-regulatory tools to being co-regulatory tools, the compliance with which ultimately comes to represent a parameter to measure compliance with the new set of rules – especially those related to the assessment and mitigation of systemic risks.[182] In this respect, the new system established by the DSA has the effect of transforming the instrument of the code of conduct into a hybrid tool which, while being developed through a voluntary cooperation between private actors and public institutions at the stage of its writing, is conversely subjected to a process of "juridification" with respect to its implementation and enforcement. Codes thus become a proxy for the assessment of the accountability of providers.[183]

The new co-regulatory framework on codes of conduct, while helpful with regard to the fight against the dissemination of any type of harmful or illegal content, will potentially play a particularly important role when addressing "information bads" that are not directly included within the category of "illegal content", such as is the case of disinformation – subjected, indeed, to the first Code of Practice specifically drafted with a view to complementing the DSA.[184] As regards the case of hate speech, the resort to a dedicated code of conduct similarly appears to be the appropriate avenue for the EU to further specify the best practices that providers of intermediary services should follow to fully comply with the Regulation. Indeed, as mentioned in Chapter 2,[185] the current absence of a uniform framework on hate speech across Member States and of a basis within the Treaties allowing the EU to adopt harmonizing legislation – at least until Article 83, paragraph 1, TFEU is amended accordingly – would make it rather difficult, if not impossible, to adopt a regulatory instrument in this field.

[182] The new Regulation clarifies, indeed, that "adherence to and compliance with a given code of conduct by a very large online platform or a very large online search engine may be considered as an appropriate risk mitigating measure", while the "refusal without proper explanations by a provider of an online platform or of an online search engine of the Commission's invitation to participate in the application of such a code of conduct could be taken into account, where relevant, when determining whether the online platform or the online search engine has infringed the obligations laid down by this Regulation". DSA rec 104.

[183] Carl Vander Maelen, 'Hardly Law or Hard Law? Investigating the Dimensions of Functionality and Legislation of Codes of Conduct in Recent EU Legislation and the Normative Repercussions Thereof' (2022) 47 European Law Review 752. See more *supra*, §3.5.3.5.

[184] Strengthened Code of Practice on Disinformation 2022. Letters (h)-(g) of the Preamble to the Code clarify, indeed, that "actions under the Code will complement and be aligned with regulatory requirements and overall objectives in the Digital Services Act" and that the "Code of Practice aims to become a Code of Conduct under Article [45] of the DSA … regarding Very Large Online Platforms that sign up to its Commitments and Measures". At the same time, letter (k) also mentions the will to facilitate the participation in the Code of providers that do not qualify as VLOPs or VLOSEs, stressing that "they are encouraged to subscribe to Commitments that are relevant to their services and to implement them through measures that are proportionate in light of the size and nature of their services and the resources available to them".

[185] See *supra*, §2.2.3.2.

5.5.2.2. Renovating the scope of applicability of the Code of Conduct

Admittedly, an EU Code of Conduct on Illegal Hate Speech already exists. However, especially in light of its limited contents and scope of intervention, as well as its reportedly underwhelming results,[186] a renovation of that tool seems to be in order, starting from the title itself. Indeed, the new Code should first and foremost encompass not only those forms of hate speech that are "illegal" under EU or national domestic law – as this would hamper extensively its power to shape meaningfully the moderation and curation practices of VLOPs and VLOSEs – but, more generally, all harmful forms of hate speech.

The DSA sets the legal basis for such an extension of the scope of the Code. The mandatory risk assessment and mitigation mechanism required of VLOPs and VLOSEs concerns indeed not only those "systemic risks" connected to the dissemination of "illegal content" but also, more generally, those represented by the possibility of "actual or foreseeable negative effects" that affect the exercise of fundamental rights; affect the civic discourse and electoral processes, as well as public security; or are otherwise related to gender-based violence, the protection of public health and minors, and serious negative consequences to the person's physical and mental well-being.[187]

Most of these "negative effects" can in fact be the direct product of the dissemination of hate speech content,[188] meaning that the mentioned due diligence obligation likely extends to the systemic risk of dissemination not only of hate speech that is illegal, but also of hate speech that is, in fact, simply "harmful". That being the case, the DSA's rules on codes of conduct would be applicable also with respect to hate speech going beyond the limited scope of Council Framework Decision 2008/913/JHA.[189]

As a matter of fact, such an all-encompassing new Code of Conduct on Hate Speech may well benefit in many ways ISPs themselves, as it would allow them to focus their compliance efforts on the application of harmonized guidelines, rather than having to deal with the magmatic legal stance of hate speech across the various Member States.

5.5.2.3. Renovating the content of the Code of Conduct through the lens of substantive equality

Furthermore, the text of the 2016 CoC on Illegal Hate Speech is rather synthetic and mainly contains only vague commitments by the signatories. A more analytical text, such as that of the Strengthened Code of Practice on Disinformation, would be necessary to present providers with more appropriate and adequate instructions as to the best practices to follow. In this regard, the Strengthened Code of Practice on Disinformation represents a rather meaningful antecedent precisely because it does not only include clauses aimed at reducing the amount of disinformation, but also provisions oriented towards the

[186] See *supra*, §3.4.3.3.
[187] DSA art 34, para 1.
[188] See *supra*, §2.
[189] 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 2008 (OJ L 328/5). See *supra*, §2.2.3.2.

empowerment of users, the research community, and the fact-checking community.[190] In other words, it complements a "negative" approach towards the countering of disinformation, focused on the removal and reduction of such content, with more "positive" measures.

For instance, the Strengthened Code of Practice on Disinformation expressly requires signatories to strengthen their efforts "in the area of media literacy and critical thinking, also with the aim to include vulnerable groups"[191] – notably by designing and implementing, or continuing to maintain, tools "empowering users with context on the content visible on services or with guidance on how to evaluate it"[192] – as well as to further empower users *vis-à-vis* the use of automated content curation systems.[193] Additionally, providers should grant users further tools to identify disinformation, for example through instruments allowing them to assess the "factual accuracy of sources through fact-checks from fact-checking organisations that have flagged potential Disinformation, as well as warning labels from other authoritative sources".[194] Moreover, signatory parties commit to actively engage in cooperative efforts with researchers and fact-checkers, namely by providing them promptly with the necessary data and information,[195] while operating in full respect of the highest ethical and transparency standards possible.[196]

In a similar way, the new Code of Conduct on Hate Speech should not focus only on measures aimed at curbing the dissemination of hate speech across the Internet, but also on positive actions specifically designed to promote the further empowerment of victimized communities and allowing them to actively participate, also through the tool of counter-narratives and the reappropriation of traditionally disparaging terms, in the deconstruction of those ancient dominance dynamics that represent the core of the hate speech phenomenon.[197] However, the implementation of such a positive approach would require the drafters of the Code to bear clearly in mind what the main ideologies and purposes behind the governance of hate speech should, in fact, be.

It is against this backdrop that the principles and goals of substantive equality could serve as the starting point for the discussion, offering a paradigm of thought to interpret the phenomenon of online hate speech and the rationales supporting the fight against it. In other words, the value of substantive equality should play a guiding role when drafting a new Code of Conduct. This way, the Code would represent a further instrument for injecting constitutional values connected to transparency, rule of law, and due process of law into the platforms' practices of hate speech moderation, taking into account the specific and characteristic aspects of any particular "information bad". Focusing on the promotion of substantive equality would allow, notably, the orientation of choices regarding

[190] Strengthened Code of Practice on Disinformation s V–VII.
[191] ibid Commitment 17.
[192] ibid Measure 17.1.
[193] ibid Commitments 18-20.
[194] ibid Commitment 21.
[195] ibid Commitments 26-27, 31-32.
[196] ibid Commitments 28-30, 33.
[197] See *supra*, §2.5.1.

the design and implementation of dedicated AI systems and the direct mitigation of the collateral effects of algorithmic mistakes and, accordingly, help render the false positives error rate acceptable.

With regard to this point, another aspect should be clarified. As specified by the DSA and mentioned above, the new co-regulatory governance system is based on codes of conduct that are to be drawn up by all stakeholders involved, including the providers of VLOPs and VLOSEs themselves.[198] As a result, an important challenge would be represented by bringing the principles of substantive equality to a table composed of private actors that, as shown above,[199] still hold a view of the governance of hate speech that is in many ways anchored to a formalistic interpretation of the value of equality. Thus, the "colourblind" approach followed by platforms often seems to be at odds with what Fredman would refer to as the "redistributive" and "transformative" dimensions of substantive equality.[200]

In order to achieve the objective of producing a Code of Conduct whose contents are shared by all actors involved, a common terrain should therefore be identified as a necessary starting point. To overcome such an *impasse*, the common terrain hereby suggested is represented by another facet of substantive equality, namely, its participative dimension. Indeed, many providers of VLOPs and VLOSEs tend to clarify that their hate speech policies are adopted and enforced precisely with a view, *inter alia*, to promoting the right of all users to enjoy their free speech prerogatives which could be seriously hampered if they were made the victims of hate speech.

The need to guarantee and promote effectively the possibility for everyone – including members of minority, discriminated, or marginalized groups – to enjoy their right to freedom of expression could thus be taken as the founding (and shared) starting point of negotiations, ultimately serving as the lodestar to develop a co-regulatory framework that is able to fully foster the interests of victimized communities. Relevant measures could include, for instance, enhanced participation of associations and NGOs representing those communities at all stages of the drafting and enforcement of platforms' policy, as well as in the context of complaints made against the (automated) imposition of restrictions and sanctions that are potentially fruit of a machine's biased output. Simplified complaint-handling procedures could furthermore be envisaged in defence of minority, vulnerable, and discriminated groups.

Be that as it may, it is worth underscoring that the EU Commission itself has suggested the possibility to review the existing 2016 CoC on Illegal Hate Speech, in the wake of the unsatisfactory results observed at the end of 2022 during its seventh evaluation. At the time of writing, the proceedings for such a renovation are yet to be started. Nevertheless, such a review may well represent a turning point for the future of online hate speech governance within the European Union.

[198] DSA art 45, para 2.
[199] See *supra*, §5.2.3.
[200] See *supra*, §5.2.1.3.

5.5.3. *AI regulation beyond the Digital Services Act*

The Digital Services Act arguably represents the most important piece of legislation within the EU framework addressing the use of automated systems for content moderation and content curation purposes. However, it is not the only regulatory tool reflecting a European constitutional aspiration in the governance of digital technologies and automated decision-making systems. Indeed, the likelihood of collateral negative effects of AI on the fundamental rights of individuals has garnered increasing global attention and triggered the rise of legislative attempts calling for more human-centred and trustworthy AI systems.

In the EU, one of the first provisions adopted to address the use of automated decision-making is notoriously contained within the GDPR which states that data subjects "shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning [them] or significantly affects [them]",[201] clarifying that, when such automated decision-making is allowed because it is necessary for entering into, or performance of, a contract or because the data subjects have given their consent, "the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention … to express his or her point of view and to contest the decision".[202] The described provisions of the DSA, concerning the use of AI for content moderation and content curation, thus build directly upon the system designed by the GDPR, and complement the system designed by the latter with further transparency and procedural requirements.

The AI Act[203] represents, however, the most notable legislative attempt to regulate AI in the context of the EU. The Regulation, as is well known, follows a risk-based approach that is focused on four levels of risk:[204] unacceptable; high; limited; minimal or none. The high-risk category, in particular, includes – on top of systems related to product safety – also those systems that fall into Annex III of the Regulation, which contains, for this purpose, a range of relevant areas and fields of action. Those systems shall have to comply with a long and extensive series of requirements and procedures to ensure that they are safe, trustworthy, and respectful of individuals' fundamental rights. It has been correctly

[201] 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), OJ L/119 art 22, para 1.

[202] ibid 22, para 3.

[203] Regulation (EU) 2024/... of the European Parliament and of the Council of ... laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

[204] Giovanni De Gregorio and Pietro Dunn, 'The European Risk-Based Approaches: Connecting Constitutional Dots in the Digital Age' (2022) 59 Common Market Law Review 473, 488–493.

noted that, although the AI Act could potentially add another layer of protection,[205] the current absence of a reference to (hard) content moderation within Annex III means that, in fact, such an additional layer may not be available yet.[206]

Conversely, the AI Act has introduced important transparency obligations concerning those AI systems designed to interact with human beings as well as those generative AI systems capable of synthetizing or manipulating content, also with a view to countering the effects of the use of AI in the dissemination of unwarranted, illegal, or harmful material across the Internet. The main concern of the EU lawmakers was that of countering the collateral effects brought by the increased development and spread of bots and deepfakes in the digital informational ecosystem. Such an attempt, entailing the need for providers and deployers of those AI systems to adopt the necessary technical measures to allow individuals to be aware of the artificial nature of the entity they are interacting with or of the content they are viewing,[207] is certainly to be welcomed as it represents an important asset in the countering of phenomena such as disinformation and hate speech.

Moving beyond the EU, it is also worth mentioning the draft Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law, under development by the CoE Committee on Artificial Intelligence (CAI).[208] The latest version of the draft CAI Convention, as approved by the Committee on 14 March 2024 provides specifically that states parties shall adopt or maintain measures to protect democratic processes *vis-à-vis* the development and deployment of AI systems, including measures to ensure "individuals' fair access to and participation in public debate, as well as their ability to freely form opinions".[209] In this respect, the text of the draft Convention is very broad, encompassing first and foremost the possible use of AI systems, including generative AI systems, to create and disseminate across the Internet a variety of polluting contents – including not only hate speech, but also disinformation.[210] Nevertheless, the provision is also arguably applicable to the case of automated systems of content moderation

[205] Thus, for instance, the European Union Agency for Fundamental Rights: "The European Commission's proposal of April 2021 for an AI act … reflects the increased policy and legislative focus on AI. The proposal contains provisions relevant to the protection of fundamental rights. These provisions include requirements for risk management (Article 9), including with respect to fundamental rights, and a conformity assessment for high-risk AI systems (Article 43). Notably, with respect to the focus of this report, the proposed [AI Act] also includes a legal basis for the processing of sensitive data to detect, monitor and correct bias that could lead to discrimination (Article 10 (5))". European Union Agency for Fundamental Rights (n 137) 8.

[206] Anna Morandini, 'Recalibrating Platforms' AI Systems: EU advances' (*MediaLaws*, 10 July 2023) <https://www.medialaws.eu/recalibrating-platforms-ai-systems-eu-advances/> accessed 7 December 2023.

[207] AI Act art 50, paras 1-2, 4.

[208] Committee on Artificial Intelligence, 'Draft Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law' (Council of Europe 2024) CM(2024)52-prov1 <https://rm.coe.int/-1493-10-1b-committee-on-artificial-intelligence-cai-b-draft-framework/1680aee411> accessed 4 May 2024.

[209] ibid 5, para 2.

[210] Tate Ryan-Mosley, 'How Generative AI Is Boosting the Spread of Disinformation and Propaganda' (*MIT Technology Review*, 4 October 2023) <https://www.technologyreview.com/2023/10/04/1080801/generative-ai-boosting-disinformation-and-propaganda-freedom-house/> accessed 8 December 2023.

and curation, which have the inherent and inescapable quality of shaping directly the public debate and the freedom of individuals to express their thoughts and opinions.

Furthermore, the draft Convention explicitly emphasizes that equality and non-discrimination should represent guiding principles in the governance of AI systems and stresses that contracting states should therefore undertake "to adopt or maintain measures aimed at overcoming inequalities to achieve fair, just and equitable outcomes, in line with [their] applicable domestic and international human rights obligations, in relation to activities within the lifecycle of artificial intelligence systems".[211] Interpreting *de juncto* these provisions would thus lead to conclude that, for the purposes of the draft Convention, automated content moderation and curation should necessarily be informed directly by the principle of equality.

Although touching only marginally upon the matter of the automated moderation and curation of content online, regulatory tools such as the AI Act proposal and the draft CAI Convention will likely represent important frameworks also in the context of online speech governance, by setting further rules to ensure that the development, deployment, and use of AI responds to anthropocentric principles and to criteria of transparency, rule of law, due process, and equality.[212]

Whereas the DSA will still represent, in the EU, the core legislation when dealing with the application of AI to the monitoring and management of the informational flow across the Internet, the evolving European and international framework on AI regulation could play in the future a crucial role in reinforcing its rules and objectives.

5.6. Conclusions

The present Chapter has highlighted how the regulation and governance of online content, and especially of online hate speech, necessarily requires taking into consideration how private platforms actually interpret and deal with speech across their infrastructures. Policy and legislative options should consider not only how platforms tend to frame within their terms and conditions the rules applicable to hate speech, but also, and especially, the technical tools deployed to detect it and remove it. Indeed, the standards platforms establish, and the ways in which they enforce them, are not uninfluential when it comes to the pursuit of the goals and rationales supporting hate speech governance.

Most notably, the greatest challenge ahead is represented by the risk that the large employment of AI systems, coupled with hate speech policies that often do not fully reflect egalitarian aspirations, may lead to a short-circuit of the system, whereby those individuals and groups of people that should be protected by anti-hate speech policies are, in fact, silenced. A risk which is, besides, potentially enhanced by legislation aimed at increasing the liability and responsibilities of the providers of intermediary services themselves.

---

[211] ibid 10, para 2.

[212] With specific regard to the anti-discrimination dimensions of the AI Act and of the draft CAI Convention, see Nardocci (n 125) 2379–2388.

Within the EU framework, the DSA thus represents both a hazard and an asset for the future of hate speech governance in the Old Continent. It is a hazard precisely because the adoption of an extensive framework encompassing numerous new due diligence obligations inherently brings with it the risk of pushing platforms to over-removal. It is an asset because it couples those obligations with a similarly extensive range of provisions aimed at injecting, within the content moderation and content curation practices of regulated ISPs, a set of constitutional principles and values such as transparency, the rule of law, and the due process of law.

Additionally, the horizontal nature of the DSA, which allows it precisely to take on that (digital) "constitutional" dimension, also represents a limitation due to its lack of specificity. It is in this context that further action by the EU is needed, for example through a deep renovation of the 2016 CoC on Illegal Hate Speech. In order for such a renovation to be fully fruitful, nevertheless, it is essential that the drafters bear clearly in mind the core rationale justifying the governance and proscription of the phenomenon of hate speech itself. For this purpose, the present work suggests, as a possible lodestar to be followed, the principle of substantive equality.

6.

Concluding Remarks

Summary: 6.1. Main findings of the research: an overview. – 6.2. The challenges ahead.

6.1. Main findings of the research: an overview

The governance of online freedom of expression, and of hate speech in particular, represents a complex challenge for regulators. Many of the issues (still) to be confronted by policymakers and lawmakers have been critically examined in the previous Chapters of the present work. An overview of the main findings and arguments made throughout the research thus appears to be necessary.

The purpose of Chapter 2 was to address preliminary matters concerning the reasons behind, and the justifications for, the option to intervene through the instrument of the law to regulate and reduce the phenomenon of hate speech, both offline and online, notwithstanding the implications in terms of restricting the right to freedom of expression that such a choice necessarily entails. In this respect, the research questions considered in that Chapter referred to the constitutional stance of legislative limitations of hate speech utterances and to the possible outcomes of a balancing process between the interests protected through these limitations and the value of free speech. In other words, hate speech governance necessarily faces the constitutional challenge of defining to what extent the prohibition of hate speech is capable of pursuing public or individual interests that justify the abridgment of the fundamental human right to speak and express one's thoughts and opinions – a necessary premise for democracies themselves.

Identifying the correct response is not at all an easy task, especially because such a response inherently depends on the constitutional framework considered. Thus, for example, the First Amendment to the US Constitution, as interpreted by the SCOTUS throughout the twentieth century, has led to the consideration of free speech as an almost absolute right which cannot be limited with a view to reducing hate speech as such (as this would represent an impermissible viewpoint-based discrimination). Speech may be subject to limitations inasmuch as it amounts to low-value speech such as "true threats"[1]

[1] *Brandenburg v Ohio* 395 US 444 (1969); *Virginia v Black* 538 US 343 (2003).

or "fighting words".[2] In any case, what is relevant in these cases, according to US constitutional law, is not the discriminatory or disparaging content expressed therein but, rather, the modalities and potential consequences thereof.[3] Conversely, in Europe, hate speech bans have generally been considered to be compatible with, or even necessary for, the democratic asset of the state.[4] In many cases the utterance of hate speech has even been considered by the ECtHR to amount to an abuse of freedom of expression.[5]

Since the justifications put forward by the CoE and EU systems are strictly intertwined not only with the well-functioning of democratic societies as a whole, but also with the promotion and protection of the human dignity of persons targeted by hate speech, Chapter 2 argued in favour of the need to push forward the debate on hate speech governance in the European context. In particular, this work argued that policymakers should consider the dominant/dominated dynamics entailed by the phenomenon of hate speech and aim to offer a remedy against those dynamics, precisely with a view to promoting and fostering the principle of substantive equality. In the context, specifically, of the digital sphere, taking substantive equality as the ideal lodestar of hate speech governance would lead effectively to the development of strategies oriented not only towards the simple (negative) goal of removing such content, but also towards the (positive) goal of strengthening victimized groups and individuals and allowing them to actively participate in the public debate.

To understand how the principle of substantive equality may be injected into the governance of hate speech, the work subsequently explored how the law has in fact evolved in recent years with respect to the regulation of online freedom of expression and of content moderation and content curation practices. In this respect, the shift towards forms of liability-enhancing strategies, providing for increased duties of Internet intermediaries, has in recent years been particularly relevant. Bearing this point in mind, Chapter 3 considered specifically the European framework, addressing questions concerning the ways in which the ECtHR, the CJEU, and the EU lawmaker have dealt with those areas and what impacts this has had on the legal treatment of hate speech specifically. The findings have, in this respect, been plentiful.

As regards the case law of the ECtHR, following *Delfi*,[6] the Strasbourg judges have developed an approach towards intermediary liability for third-party illegal content that, while generally being quite lenient towards intermediaries, tends nevertheless to welcome

[2] *Chaplinsky v New Hampshire* 315 US 568 (1942).

[3] *RAV v City of St Paul* 505 US 377 (1992).

[4] *Gunduz v Turkey* [2003] ECtHR 35071/97, ECHR 2003-XI; *Beizaras and Levickas v Lithuania* [2020] ECtHR 41288/15; *Sanchez v France* [2023] ECtHR [GC] 45581/15, ECHR 2023.

[5] See, *ex multis*, *Garaudy v France* (dec) [2003] ECtHR 65831/01, ECHR 2003-IX; *Witzsch v Germany (2)* (dec) [2005] ECtHR 7485/03; *Norwood v the United Kingdom* (dec) [2004] ECtHR 23131/03, ECHR 2004-XI; *Pavel Ivanov v Russia* (dec) [2007] ECtHR 35222/04; *M'bala M'bala v France* (dec) [2015] ECtHR 25239/13, ECHR 2015-VIII..

[6] *Delfi AS v Estonia* [2015] ECtHR [GC] 64569/09, ECHR 2015.

the provision of obligations aimed specifically at combatting hate speech content.[7] With respect to the EU, Chapter 3 highlighted the shift from an overall liberal framework, mostly reflected by the adoption of the ECD, towards a progressively more interventionist approach. In particular, first through the (manipulative) intervention of the Luxembourg Court and, subsequently, with the rise of a new legislative season, the EU has turned precisely towards the enhancement of the liability, responsibility, and accountability of ISPs for their content moderation and content curation practices, with the DSA currently representing the most significant piece of legislation for the governance of online speech in general and hate speech in particular.

Although, due to its newness, the overall practical effects and impact of the DSA are yet to be assessed, Chapter 3 noted that the extent to which the new Regulation will in fact be applicable with respect to the moderation of hate speech still raises some doubts, especially in the light of the current lack of harmonization, across Member States, concerning the recognition of what is to be considered "illegal" hate speech. In this respect, *de iure condendo*, it is arguably desirable that further action is taken by the EU to establish common rules and/or guidelines, for example by including hate speech and hate crimes within the category of EU crimes under Article 83, paragraph 1, TFEU, or through the adoption of a new Code of Conduct on Hate Speech substituting the 2016 CoC on Illegal Hate Speech – the latter option currently being the most viable and swift.

In this respect, the implementation of complementary tools specifically oriented towards the appropriate governance of online hate speech would not simply be an important result in terms of pushing providers of intermediary services to address all relevant forms of hate speech present on the Internet. Rather, those tools could have another, probably even more important, effect, that is the introduction of specific guarantees aimed at ensuring that content moderation and content curation practices deployed against hate speech are fully sustainable in terms of protecting both the freedom of expression of the recipients of the services and the principle of substantive equality itself. Among other areas of intervention, specific attention should be given to the impact that the use of AI systems for hate speech moderation has on such fundamental rights and principles. Indeed, while the general framework established by the DSA – especially the risk assessment and mitigation system designed within Articles 34 and 35 of the Regulation – may well have the effect of further encouraging the deployment of those tools, additional attention should be given to the need to reduce the collateral effects necessarily associated with them.

Before addressing these challenges, the research addressed another equally important set of questions, concerning specifically the global context in which the described EU framework is actually set and the relationships that that framework has with other jurisdictions, both intra-EU and extra-EU. In particular, Chapter 4 investigated to what extent

[7] *Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v Hungary* [2016] ECtHR 22947/13; *Pihl v Sweden* (dec) [2017] ECtHR 74742/14; *Høiness v Norway* [2019] ECtHR 43624/14; *Jezior v Poland* [2020] ECtHR 31955/11; *Standard Verlagsgesellschaft Mbh v Austria (no 3)* [2021] ECtHR 39378/15; *Sanchez v France* (n 4).

EU law, and especially the DSA as applicable with respect to the phenomenon of hate speech, is consistent with the domestic legislations of EU Member States. Moreover, Chapter 4 gave an overview of the UK's OSA – possibly the closest piece of legislation to the DSA (closest not only in terms of time and space but also, to a certain extent, in terms of content) – and of the developing framework in the US, with a view to highlighting not only aspects of similarity but also, and even more importantly, the differences from the EU's approach. Chapter 4 also contained a brief overview of other legal and policy options taken on a global scale.

With regard to the relationship between the EU framework and the domestic legislations of EU Member States, Chapter 4 has stressed the challenges raised to a uniform fight against the phenomenon of online hate speech especially in the light of the different sensitivities characterizing each country. For instance, the enforcement of the DSA, considering its confirmed reliance upon the country-of-origin principle[8] and of an emerging shift of regulatory prerogatives from Berlin to Brussels, could lead to significant clashes with the German NetzDG. Another significant example is represented by the trend, characterizing specifically Eastern European Member States, represented by the adoption of "memory laws" that, rather than being aimed at protecting minority or discriminated groups, tend to be focused on protecting the population majority (thus following an approach which is opposite to the pursuit of substantive equality).

At the same time, the analysis of some extra-EU jurisdictions, especially the US framework, has demonstrated the significant difficulties that the application of the DSA could face with respect to its application in the light of its potential clashes with foreign legislations. A paramount example of this would be, were they to be considered admissible under the First Amendment, the cases of Texas' HB 20 and of Florida's SB 7072 proposals. Conversely, other jurisdictions have taken an approach towards online speech governance that is in many ways inspired by aspirations for a safe and transparent digital environment similar to those characterizing the DSA, as demonstrated for example by the choice of the UK to enact the OSA.

The international scenario is, in other words, quite composite and will require the EU to take this aspect into detailed account when enforcing its new set of rules. At the same time, the EU will have to keep bearing in mind its own constitutional stance in terms of hate speech governance. In other words, while acknowledging the need to come to terms and cooperate with different sensitivities and legislations, it will be necessary for the European lawmakers and policymakers to have a clear understanding of the objectives to be pursued in light of the European constitutional understanding of the regulation of hate speech. It is in this respect that the principle of substantive equality may represent a lodestar principle for the further development of policy strategies in this field.

Chapter 5 thus finally considered the role that substantive equality could play in this respect. Precisely because of the increasing focus on the enhancement of intermediary

[8] See, in this regard, Case C-376/22, *Google Ireland Limited and Others v Kommunikationsbehörde Austria* [2023] ECLI:EU:C:2023:467, Opinion of AG Szpunar [8]; Case C-376/22, *Google Ireland Limited and Others v Kommunikationsbehörde Austria* [2023] ECLI:EU:C:2023:835 [64].

liability that is showcased by EU law (but also by the jurisdictions of several other countries), attention must be given to the ways in which those intermediaries actually fight illegal or harmful content, including in particular hateful content. Chapter 5, therefore, dealt with the practical effects of private content moderation policies and practices on the fundamental rights of users and on the overall pursuit of democratic values and principles. Particular attention was given to the notion and concept of hate speech under major platforms' terms and conditions, with a view to understanding the rationales motivating those private actors' actions, and to the actual technical instruments used to enforce their terms and conditions.

With this in mind, Chapter 5 highlighted how the greatest challenge ahead is represented by the risk that the large scale employment of AI systems, coupled with hate speech policies not paying sufficient attention to the egalitarian goals of the fight against such a phenomenon, may ultimately lead to a short-circuit of the system whereby those individuals and groups of people that should be protected by anti-hate speech policies face, in fact, the risk of being silenced themselves. In the EU, the DSA, while presenting in this respect some hazardous elements due to its heavy reliance on private moderation, may contain nevertheless the seeds to counter that short-circuit, provided that future actions are driven by adequate goals and objectives.

Indeed, the DSA represents in this respect a potential asset because, on top of due diligence obligations oriented towards the removal of unwarranted content, it also foresees a whole range of duties focused, rather, on imposing constraints on the unlimited power of private platforms, with a view to promoting, *inter alia*, the principles of transparency, due process, and rule of law. It is with regard to these provisions that the DSA reflects its "digital constitutional" dimension, a dimension which, general in nature, may serve as a framework to inject constitutional values and principles into more specific anti-hate speech practices.

It is in this respect that Chapter 5, building on the framework established by the new Regulation, suggested as a way forward a long-overdue renovation of the 2016 CoC on Illegal Hate Speech, which is today in many ways unfit to serve its original purpose. The Chapter also argued that, to reach its goals, such a renovation could (and should) nevertheless be oriented towards the promotion and enforcement of the principle of substantive equality, interpreted *in primis* under its participative dimension. Such an instrument would further strengthen the EU's stance on hate speech governance with regard to its relation both to the international legal scenario on a global scale and to the private owners of digital infrastructures, while helping counter the collateral effects of contemporary content moderation and content curation practices.

6.2. The challenges ahead

As highlighted throughout this work, the governance of speech in the context of the contemporary "algorithmic society",[9] and especially the governance of unlawful and/or harmful speech such as is the case of hate speech, requires addressing both those old questions that have traditionally been posed with regard to offline freedom of expression and the new challenges deriving from the new dynamics arising from the self-imposition of powerful private digital actors within the speech market.

Among the first set of questions, the most significant and relevant one is possibly the one concerning the extent to which it is acceptable, under constitutional and human rights law, to restrict the utterance of certain expressions with a view to limiting, precisely, the dissemination of certain ideas and thoughts. How is the proscription of hate speech justifiable? What is the limit of freedom of expression itself, beyond which it is possible for the state to intervene? In other words, borrowing the formula contained within the ECHR, is governmental action against hate speech "necessary in a democratic society" or, alternatively, what is the threshold beyond which a hateful utterance amounts in fact to an "abuse" of the right protected under Article 10 of the Convention? Conversely, in the language of the US First Amendment jurisprudence, under what terms may hate speech diverge from the grounds of fully protected free speech and actually translate into actionable "fighting words" or "true threats"? Ultimately, the question consists of striking the appropriate balance, acceptable in constitutional terms, between the interests related to the protection of freedom of expression and the interests related to the rights to dignity and equality.

The second order of challenges, arising from the affirmation of the algorithmic society, requires us to consider the impact and role of private Internet intermediaries in the context of speech governance and to pay attention, specifically, to the ways in which they administer their own computational power. In an age where private platforms have the capacity to silence, by deplatforming him, a former President of the US (as shown by the case of Donald Trump),[10] public institutions have had to come to terms with such powers and will have to coherently address the (also) constitutional implications those powers entail. In this respect, the EU has shown to be both alarmed and fascinated by the rise of Internet intermediaries. On the one hand, the EU, aware of the opportunities offered by platforms' computational resources, has increasingly resorted to attempting to harness those powers for the pursuit of its own public interests. On the other hand, it has recognized the dangers and implications posed by such private power in terms of the protection of fundamental

[9] Jack M Balkin, 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation' (2018) 51 U.C. Davis Law Review 1149. See *supra*, §1.1.1.

[10] Andrew Ross Sorkin and others, 'The Deplatforming of President Trump' *The New York Times* (8 January 2021) <https://www.nytimes.com/2021/01/08/business/dealbook/trump-facebook-twitter-deplatforming.html> accessed 10 January 2024. For a comment on the comparative constitutional stance of such a power, both in Europe and in the US, see among others Aleksandra Kuczerawy, 'Does Twitter Trump Trump?' (*Verfassungsblog*, 29 January 2021) <https://verfassungsblog.de/twitter-trump-trump/> accessed 10 January 2024.

rights and principles and has, therefore, attempted to develop a legislative framework setting limits and constraints to its unlimited and unregulated exercise by platforms. In the context of speech governance in general – and of hate speech governance in particular – the challenge is thus once again that of striking a balance: in this case, between the goal of exploiting the resources of Internet intermediaries and the need to set adequate legal limitations to reduce the collateral impact on the constitutional rights of individuals.

The present work has suggested the principle of substantive equality – especially in its participative dimension – as a key value and a proxy to address both balancing challenges. First, substantive equality can serve as an adequate justification, at least within the European framework, for the regulation of hate speech. In this sense, it may also serve as a reasonable standard for the identification of appropriate legislative and policy measures both in the online and offline environment. Second, when defining clearer standards and guidelines regarding the moderation of the phenomenon of hate speech on the Internet, substantive equality may be adopted as an orienting principle to better set the boundaries between what intermediaries can (and should) remove for the benefit of public interests and what should be left untouched in order to preserve fundamental individual prerogatives and to fully remedy the relations of dominance entailed by hate speech.

The main challenge ahead is clearly represented by the inherent difficulties of adequately incorporating such principles into policy and legislative documents. This work pointed towards the instrument of codes of conduct, explicitly recognized by the DSA, as a possible arena where the deployment of the principle of substantive equality could take place. However, in order to pursue such a goal, a cooperative approach between public and private institutions is going to be necessary, in order to identify common grounds of departure and to define the practical means and tools through which the described constitutional interests may, in fact, be promoted.

References

Bibliography and online resources

Agarwal S and Chowdary CR, 'Combating Hate Speech Using an Adaptive Ensemble Learning Model with a Case Study on COVID-19' (2021) 185 Expert Systems with Applications 115632.

Alexy R, 'The Responsibility of Internet Portal Providers for Readers' Comments. Argumentation and Balancing in the Case of Delfi AS v. Estonia' in María Elósegui, Alina Miron and Iulia Motoc (eds), *The Rule of Law in Europe: Recent Challenges and Judicial Responses* (Springer 2021).

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

——, 'The Internet, Internet Intermediaries and Hate Speech: Freedom of Expression in Decline?' (2023) 20 SCRIPTed 243.

Allegri MR, *Ubi Social, Ibi Ius: Fondamenti Costituzionali Dei* Social Network *e Profili Giuridici Della Responsabilità Dei* Provider (Franco Angeli 2018).

Alrasheed G and Lim M, 'Beyond a Technical Bug: Biased Algorithms and Moderation Are Censoring Activists on Social Media' (*The Conversation*, 16 May 2021) <http://theconversation.com/beyond-a-technical-bug-biased-algorithms-and-moderation-are-censoring-activists-on-social-media-160669> accessed 18 November 2023.

Amar AR, 'The Case of the Missing Amendments: R.A.V. v. City of St. Paul' (1992) 106 Harvard Law Review 124.

Amnesty International, 'The Social Atrocity: Meta and the Right to Remedy for the Rohingya' (Amnesty International 2022) ASA 16/5933/2022 <https://www.amnesty.org/en/wp-content/uploads/2022/09/ASA1659332022ENGLISH.pdf> accessed 26 January 2023.

Andreoli E, 'Continuities and Discontinuities. First Amendment and Digital Free Speech in US Constitutionalism' (2023) 56 DPCE Online 261.

Angelopoulos C, 'MTE v Hungary: A New ECtHR Judgment on Intermediary Liability and Freedom of Expression' (2016) 11 Journal of Intellectual Property Law & Practice 582.

Angelopoulos C and Quintais JP, 'Fixing Copyright Reform' (2019) 10 Journal of Intellectual Property, information Technology and Electronic Commerce Law 147.

Are C, ‘How Instagram’s Algorithm Is Censoring Women and Vulnerable Users but Helping Online Abusers’ (2020) 20 Feminist Media Studies 741.

——, ‘The Shadowban Cycle: An Autoethnography of Pole Dancing, Nudity and Censorship on Instagram’ (2022) 22 Feminist Media Studies 2002.

Article 19, ‘The Camden Principles on Freedom of Expression and Equality’ (April 2009) <https://www.article19.org/data/files/pdfs/standards/the-camden-principles-on-freedom-of-expression-and-equality.pdf> accessed 27 December 2022.

——, ‘Towards an Interpretation of Article 20 of the ICCPR: Thresholds for the Prohibition of Incitement to Hatred’ (Regional expert meeting on article 20, Vienna, 9/02 2010) <https://www2.ohchr.org/english/issues/opinion/articles1920_iccpr/docs/CRP7Callamard.pdf> accessed 27 December 2022.

——, ‘Prohibiting Incitement to Discrimination, Hostility or Violence’ (2012) <https://www.article19.org/data/*files*/medialibrary/3548/ARTICLE-19-policy-on-prohibition-to-incitement.pdf> accessed 28 December 2022.

——, ‘Germany: Responding to “Hate Speech”’ (2018) 19 <https://www.article19.org/resources/germany-responding-to-hate-speech/> accessed 12 July 2023.

——, ‘Hungary: Responding to “Hate Speech”’ (2018) 20 <https://www.article19.org/resources/hungary-responding-hate-speech/> accessed 16 August 2023.

Arun C and Nayak N, ‘Preliminary Findings on Online Hate Speech and the Law in India’ (2016) Berkman Klein Center Research Publication No. 2016-19 <https://cyber.harvard.edu/publications/2016/HateSpeechIndia> accessed 25 September 2023.

Austin JL, *How to Do Things with Words: The William James Lectures Delivered at Harvard University in 1955* (Clarendon Press, Oxford University Press 1962).

Baghdasaryan M, ‘Standard Verlagsgesellschaft MBH v. Austria (No. 3): Is the ECtHR Standing up for Anonymous Speech Online?’ (*Strasbourg Observers*, 25 January 2022) <https://strasbourgobservers.com/2022/01/25/standard-verlagsgesellschaft-mbh-v-austria-no-3-is-the-ecthr-standing-up-for-anonymous-speech-online/> accessed 6 May 2023.

Baistrocchi PA, ‘Liability of Intermediary Service Providers in the EU Directive on Electronic Commerce’ (2002) 19 Santa Clara Computer and High Technology Law Journal 111.

Balkin JM, ‘Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society’ (2004) 79 New York University Law Review 1.

——, ‘Old-School/New-School Speech Regulation’ (2014) 127 Harvard Law Review 2296.

——, ‘Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation’ (2018) 51 U.C. Davis Law Review 1149.

——, ‘Free Speech Is a Triangle’ (2018) 118 Columbia Law Review 2011.

Bán M and Belavusau U, 'Memory Laws' (SSRN, 9 May 2022) <https://papers.ssrn.com/abstract=4104552> accessed 14 August 2023.

Bandura A, *Moral Disengagement: How People Do Harm and Live with Themselves* (Worth Publishers, Macmillan Learning 2016).

Barata J, 'The Digital Services Act and Its Impact on the Right to Freedom of Expression: Special Focus on Risk Mitigation Obligations' (*DSA Observatory*, 27 July 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/> accessed 3 December 2021.

Barata J and others, 'Unravelling the Digital Services Act Package' (European Audiovisual Observatory 2021) 2021–1 <https://rm.coe.int/iris-special-2021-01en-dsa-package/1680a43e45> accessed 9 September 2022.

Barnes JA, 'Class and Committees in a Norwegian Island Parish' (1954) 7 Human Relations 39.

Barnes R and others, 'Supreme Court Considers If Google Is Liable for Recommending ISIS Videos' *Washington Post* (21 February 2023) <https://www.washingtonpost.com/technology/2023/02/21/gonzalez-v-google-section-230-supreme-court/> accessed 19 September 2023.

Bartolo L, '"Eyes Wide Open to the Context of Content": Reimagining the Hate Speech Policies of Social Media Platforms through a Substantive Equality Lens' (2021) 29 Renewal 39.

Bassini M, 'Fundamental Rights and Private Enforcement in the Digital Age' (2019) 25 European Law Journal 18.

——, *Internet e Libertà Di Espressione: Prospettive Costituzionali e Sovranazionali* (Aracne 2019).

——, 'Libertà Di Espressione E Social Network, Tra Nuovi "Spazi Pubblici" E "Poteri Privati". Spunti Di Comparazione' (2021) 2 Rivista di Diritto dei Media 67.

——, 'Social Networks as New Public Forums? Enforcing the Rule of Law in the Digital Environment' (2022) 1 The Italian Review of International and Comparative Law 311.

Bassini M and De Gregorio G, 'The Implementation of the Copyright Directive in Italy and the Proper Understanding of the "best Efforts" Clause' (MediaLaws) <https://www.medialaws.eu/wp-content/uploads/2021/04/Policy-paper-ML-Article-17-and-best-efforts-5.pdf> accessed 18 May 2023.

Bassini M and Vigevani GE, 'Primi Appunti Su *Fake News* e Dintorni' (2017) 1 Rivista di Diritto dei Media 11.

Bechtold E, 'Terrorism, the Internet, and the Threat to Freedom of Expression: The Regulation of Digital Intermediaries in Europe and the United States' (2020) 12 Journal of Media Law 13.

Belavusau U, ‘Fighting Hate Speech through EU Law’ (2012) 4 Amsterdam Law Forum 20.

——, ‘Mnemonic Constitutionalism and Rule of Law in Hungary and Russia’ (2020) 1 The Interdisciplinary Journal of Populism 16.

——, ‘The *NH* Case: On the “Wings of Words” in EU Anti-Discrimination Law’ (2020) 5 European Papers 1001.

——, ‘Law and the Politics of Memory’ in Maria Mälksoo (ed), *Handbook on the Politics of Memory* (Edward Elgar Publishing 2023).

Belavusau U, Gliszczyńska-Grabias A and Mälksoo M, ‘Memory Laws and Memory Wars in Poland, Russia and Ukraine’ (2021) 69 Jahrbuch des öffentlichen Rechts 95.

Benesch S, ‘Contribution to OHCHR Initiative on Incitement to National, Racial, or Religious Hatred’ (UN OHCHR 2011 Expert workshop on the prohibition of incitement to national, racial or religious hatred, Vienna, February 2011) <https://www2.ohchr.org/english/issues/opinion/articles1920_iccpr/docs/ContributionsOthers/S.Benesch.doc> accessed 26 December 2022.

Benkler Y, *The Wealth of Networks: How Social Production Transforms Markets and Freedom* (Yale University Press 2006).

Bloch-Wehba H, ‘Global Platform Governance: Private Power in the Shadow of the State’ (2019) 72 SMU Law Review 27.

Boerefijn I and Oyediran J, ‘Article 20 of the International Covenant on Civil and Political Rights’ in Sandra Coliver (ed), *Striking a Balance. Hate Speech, Freedom of Expression and Non-Discrimination* (Article 19 1992).

Bollinger LC, The Tolerant Society: *Freedom of Speech and Extremist Speech in America* (Oxford University Press 1988).

Bonis É and Peltier V, ‘Chronique de droit pénal et de procédure pénale: (janvier 2020 à juin 2020)’ (2020) 5 Titre VII 112.

Bortone R and Cerquozzi F, ‘L’Hate Speech al Tempo Di Internet’ (2017) 68 Aggiornamenti Sociali 818.

Boyd DM and Ellison NB, ‘Social Network Sites: Definition, History, and Scholarship’ (2007) 13 Journal of Computer-Mediated Communication 210.

Breckheimer II PJ, ‘A Haven for Hate: The Foreign and Domestic Implications of Protecting Internet Hate Speech under the First Amendment’ (2001) 75 Southern California Law Review 1493.

Brown A, *Hate Speech Law: A Philosophical Examination* (Routledge 2015).

——, ‘What Is Hate Speech? Part 1: The Myth of Hate’ (2017) 36 Law and Philosophy 419.

——, ‘What Is so Special about Online (as Compared to Offline) Hate Speech?’ (2018) 18 Ethnicities 297.

Brown A and Sinclair A, *The Politics of Hate Speech Laws* (Routledge 2019).

Brown K, 'Critical Race Theory Explained by One of the Original Participants' (2023) 98 New York University Law Review Online 91.

Brun Pereira M and others, 'Nuevas Posibilidades de Comunicación, Nuevos Peligros, Nuevos Desafíos: La Libertad de Expresión y El Discurso de Odio En Internet' (2022) 75 Revista IIDH 101.

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

Bucholc M, 'Commemorative Lawmaking: Memory Frames of the Democratic Backsliding in Poland After 2015' (2019) 11 Hague Journal on the Rule of Law 85.

Bukovská B, 'The European Commission's Code of Conduct for Countering Illegal Hate Speech Online' (TWG 2019) <https://www.ivir.nl/publicaties/download/Bukovska.pdf> accessed 22 January 2023.

Burrell J, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms' (2016) 3 Big Data & Society 2053951715622512.

Busch C, 'Regulating the Expanding Content Moderation Universe: A European Perspective on Infrastructure Moderation' (2022) 27 UCLA Journal of Law & Technology 32.

Buyse A, 'Dangerous Expressions: The ECHR, Violence and Free Speech' (2014) 63 International & Comparative Law Quarterly 491.

Caielli M, 'Punire l'omofobia: (Non) Ce Lo Chiede l'Europa. Riflessioni Sulle Incertezze Giurisprudenziali e Normative in Tema Di *Hate Speech*' (2015) 1 GenIUS 54.

Cambridge Consultants, 'Use of AI in Online Content Moderation' (Ofcom 2019) <https://www.ofcom.org.uk/research-and-data/online-research/online-content-moderation> accessed 30 August 2023.

Canaan I, 'NetzDG and the German Precedent for Authoritarian Creep and Authoritarian Learning' (2022) 28 Columbia Journal of European Law 101.

Cannie H and Voorhoof D, 'The Abuse Clause and Freedom of Expression in the European Human Rights Convention: An Added Value for Democracy and Human Rights Protection?' (2011) 29 Netherlands Quarterly of Human Rights 54.

Caruso C, 'L'Hate Speech a Strasburgo: Il Pluralismo Militante Del Sistema Convenzionale' (2017) 4 Quaderni costituzionali 963.

Castellaneta M, 'La Corte Europea Dei Diritti Umani e l'applicazione Del Principio Dell'abuso Del Diritto Nei Casi Di *Hate Speech*' (2017) 11 Diritti umani e diritto internazionale 745.

——, 'Responsabilità Del Politico per Commenti Altrui Su Facebook: Conforme Alla Convenzione Europea La "Tolleranza Zero" Nei Casi Di Messaggi d'odio' (2021) 3 Rivista di Diritto dei Media 311.

Cauffman C and Goanta C, ‘A New Order: The Digital Services Act and Consumer Protection’ (2021) 12 European Journal of Risk Regulation 758.

Cavaliere P, ‘Digital Platforms and the Rise of Global Regulation of Hate Speech’ (2019) 8 Cambridge International Law Journal 282.

Celeste E, ‘Digital Constitutionalism: A New Systematic Theorisation’ (2019) 33 International Review of Law, Computers & Technology 76.

Center for Countering Digital Hate, ‘X Content Moderation Failure: How Twitter/X Continues to Host Posts Reported for Extreme Hate Speech’ (CCDH 2023) <https://counterhate.com/research/twitter-x-continues-to-host-posts-reported-for-extreme-hate-speech/#about> accessed 8 November 2023.

Center for Technology and Society, ‘Evaluating Twitter’s Policies Six Months After Elon Musk’s Purchase’ (*Anti Defamation League*, 5 September 2023) <https://www.adl.org/resources/blog/evaluating-twitters-policies-six-months-after-elon-musks-purchase> accessed 8 November 2023.

Chen G, ‘How Equalitarian Regulation of Online Hate Speech Turns Authoritarian: A Chinese Perspective’ (2022) 14 Journal of Media Law 159.

Cheong I, ‘Freedom of Algorithmic Expression’ (2023) 91 University of Cincinnati Law Review 680.

Chhabra A and Vishwakarma DK, ‘A Literature Survey on Multimodal and Multilingual Automatic Hate Speech Identification’ (2023) 29 Multimedia Systems 1203.

Cinelli M and others, ‘The Echo Chamber Effect on Social Media’ (2021) 118 Proceedings of the National Academy of Sciences of the United States of America e2023301118.

——, ‘Dynamics of Online Hate and Misinformation’ (2021) 11 Scientific Reports 22083.

Citron DK, *Hate Crimes in Cyberspace* (Harvard University Press 2014).

Citron DK and Franks MA, ‘The Internet as a Speech Machine and Other Myths Confounding Section 230 Reform What’s the Harm? The Future of the First Amendment’ (2020) 2020 University of Chicago Legal Forum 45.

Citron DK and Norton H, ‘Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age’ (2011) 91 Boston University Law Review 1435.

Citron DK and Wittes B, ‘The Internet Will Not Break: Denying Bad Samaritans Sec. 230 Immunity’ (2017) 86 Fordham Law Review 401.

Clasen A, ‘Digital Services Act: Germany Proposes Creation of Advisory Board’ (*www.euractiv.com*, 9 May 2023) <https://www.euractiv.com/section/platforms/news/digital-services-act-germany-proposes-creation-of-advisory-board/> accessed 7 August 2023.

Coe P, ‘The Draft Online Safety Bill and the Regulation of Hate Speech: Have We Opened Pandora’s Box?’ (2022) 14 Journal of Media Law 50.

——, ‘Hate Speech, Free Speech and Draft Online Safety Bill’ (*Birmingham Law School Research Blog*, 12 December 2022) <https://blog.bham.ac.uk/lawresearch/2022/12/hate-speech-free-speech-and-draft-online-safety-bill/> accessed 20 August 2023.

——, ‘Is the New Online Safety Bill Built to Fail?’ (*University of Birmingham*, 18 January 2023) <https://www.birmingham.ac.uk/news/2023/is-the-new-online-safety-bill-built-to-fail> accessed 23 August 2023.

Cole MD, Etteldorf C and Ullrich C, *Cross-Border Dissemination of Online Content* (Nomos 2020).

Couture S and Toupin S, ‘What Does the Notion of “Sovereignty” Mean When Referring to the Digital?’ (2019) 21 New Media & Society 2305.

Culliford E, ‘Rohingya Refugees Sue Facebook for $150 Billion over Myanmar Violence’ *Reuters* (8 December 2021) <https://www.reuters.com/world/asia-pacific/rohingya-refugees-sue-facebook-150-billion-over-myanmar-violence-2021-12-07/> accessed 26 January 2023.

Czarnocki J, ‘Saving EU Digital Constitutionalism through the Proportionality Principle and a Transatlantic Digital Accord’ (2021) 20 European View 150.

D’Amico M, ‘Odio *On Line*: Limiti Costituzionali e Sovranazionali’ in Marilisa D’Amico and Cecilia Siccardi (eds), *La Costituzione non odia: Conoscere, prevenire e contrastare l’*hate speech on line (Giappichelli 2021).

Daniele L, ‘Disputing the Indisputable: Genocide Denial and Freedom of Expression in *Perincek v. Switzerland*’ (2016) 25 Nottingham Law Journal 141.

Das M, Pandey SK and Mukherjee A, ‘Evaluating ChatGPT’s Performance for Multilingual and Emoji-Based Hate Speech Detection’ (arXiv, 22 May 2023) <http://arxiv.org/abs/2305.13276> accessed 23 November 2023.

Davidson T, Bhattacharya D and Weber I, ‘Racial Bias in Hate Speech and Abusive Language Detection Datasets’ in Sarah T Roberts and others (eds), *Proceedings of the Third Workshop on Abusive Language Online* (Association for Computational Linguistics 2019).

De Gregorio G, ‘Expressions on Platforms: Freedom of Expression and ISP Liability in the European Digital Single Market’ (2018) 2 European Competition and Regulatory Law Review 203.

——, ‘From Constitutional Freedoms to the Power of the Platforms: Protecting Fundamental Rights Online in the Algorithmic Society’ (2019) 11 European Journal of Legal Studies 65.

——, ‘*Google v. CNIL* and *Glawischnig-Piesczek v. Facebook*: content and data in the algorithmic society’ (2020) 1 Rivista di Diritto dei Media 249.

——, ‘Democratising Online Content Moderation: A Constitutional Framework’ (2020) 36 Computer Law & Security Review 105374.

——, ‘The Digital Services Act: A Paradigmatic Example of European Digital Constitutionalism’ (*Diritti Comparati*, 16 May 2021) <https://www.dirittocomparati.it/the-digital-services-act-a-paradigmatic-example-of-european-digital-constitutionalism/> accessed 4 December 2023.

——, ‘The Rise of Digital Constitutionalism in the European Union’ (2021) 19 International Journal of Constitutional Law 41.

——, *Digital Constitutionalism in Europe: Reframing Rights and Powers in the Algorithmic Society* (Cambridge University Press 2022).

De Gregorio G and Dunn P, ‘The European Risk-Based Approaches: Connecting Constitutional Dots in the Digital Age’ (2022) 59 Common Market Law Review 473.

——, ‘Artificial Intelligence and Freedom of Expression’ in Alberto Quintavalla and Jeroen Temperman (eds), *Artificial Intelligence and Human Rights* (Oxford University Press 2023).

De Gregorio G and Stremlau N, ‘Platform Governance at the Periphery: Moderation, Shutdowns and Intervention’ in Judit Bayer and others (eds), *Perspectives on Platform Regulation. Concepts and Models of Social Media Governance Across the Globe* (Nomos 2021).

——, ‘Inequalities and Content Moderation’ (2023) 14 Global Policy 870.

De Vos M, ‘Substantive Formal Equality in EU Non-Discrimination Law’ in Thomas Giegerich (ed), *The European Union as Protector and Promoter of Equality* (Springer 2020).

——, ‘The European Court of Justice and the March towards Substantive Equality in European Union Anti-Discrimination Law’ (2020) 20 International Journal of Discrimination and the Law 62.

Delgado R, ‘Words That Wound: A Tort Action for Racial Insults, Epithets, and Name-Calling’ (1982) 17 Harvard Civil Rights-Civil Liberties Law Review 133.

Delgado R and Stefancic J, ‘Four Observations about Hate Speech’ (2009) 44 Wake Forest Law Review 353.

——, *Critical Race Theory: An Introduction* (3rd edn, New York University Press 2017).

——, *Must We Defend Nazis? Why the First Amendment Should Not Protect Hate Speech and White Supremacy* (New York University Press 2018).

Di Rosa A, Hate Speech *e Discriminazione: Un’analisi Performativa Tra Diritti Umani e Teorie Della Libertà* (Mucchi Editore 2020).

Díaz Hernández M, *Discurso de Odio En América Latina: Tendencias de Regulación, Rol de Los Intermediarios y Riesgos Para La Libertad de Expresión* (Derechos

Digitales América Latina 2020) <https://www.derechosdigitales.org/wp-content/uploads/discurso-de-odio-latam.pdf> accessed 27 September 2023.

Douek E, 'Governing Online Speech: From "Posts-as-Trumps" to Proportionality and Probability' (2021) 121 Columbia Law Review 759.

——, 'The Meta Oversight Board and the Empty Promise of Legitimacy' (SSRN, 7 September 2023) <https://papers.ssrn.com/abstract=4565180> accessed 25 October 2023.

Duarte N, Llansó E and Loup A, 'Mixed Messages? The Limits of Automated Social Media Content Analysis' (*Center for Democracy & Technology*, November 2017) <https://cdt.org/wp-content/uploads/2017/11/2017-11-13-Mixed-Messages-Paper.pdf> accessed 14 December 2021.

Dunn P, 'Moderazione Automatizzata e Discriminazione Algoritmica: Il Caso dell'*Hate Speech'* in Laura Abba, Adriana Lazzaroni and Marina Pietrangelo (eds), *La Internet Governance e le Sfide della Trasformazione Digitale* (Editoriale Scientifica 2022).

——, 'L'anonimato degli utenti quale forma mediata della libertà di stampa: Il caso *Standard Verlagsgesellschaft mbH c. Austria*' (2022) 1 Rivista di Diritto dei Media 291.

——, 'Carattere Eccezionale Dell'"Hate Speech" e Nuove Forme Di Responsabilità per Contenuti Di Terzi Nella Giurisprudenza EDU. Nota a C.Edu, Sanchez c. Francia, 15 Maggio 2023' (2023) 6 Osservatorio Costituzionale 238.

Durante M, *Computational Power: The Impact of ICT on Law, Society and Knowledge* (Routledge 2021).

Edwards L (ed), *The New Legal Framework for E-Commerce in Europe* (Hart 2005).

Efroni Z, 'The Digital Services Act: Risk-Based Regulation of Online Platforms' (*Internet Policy Review*, 16 November 2021) <https://policyreview.info/articles/news/digital-services-act-risk-based-regulation-online-platforms/1606> accessed 8 June 2023.

Eifert M and others, 'Taming the Giants: The DMA/DSA Package' (2021) 58 Common Market Law Review 987.

Einwiller SA and Kim S, 'How Online Content Providers Moderate User-Generated Content to Prevent Harmful Online Communication: An Analysis of Policies and Their Implementation' (2020) 12 Policy & Internet 184.

Eisenstein J, *Introduction to Natural Language Processing* (MIT Press 2019).

Ellis E and Watson P, *EU Anti-Discrimination Law* (2nd edn, Oxford University Press 2012).

Engle E, 'Third Party Effect of Fundamental Rights (*Drittwirkung*)' (2009) 5 Hanse Law Review 165.

European Union Agency for Fundamental Rights, *Bias in Algorithms: Artificial Intelligence and Discrimination* (Publications Office 2022) <https://data.europa.eu/doi/10.2811/25847> accessed 3 February 2023.

——, *Online Content Moderation: Current Challenges in Detecting Hate Speech* (Publications Office 2023) <https://fra.europa.eu/en/publication/2023/online-content-moderation> accessed 3 February 2023.

Farrior S, 'Molding The Matrix: The Historical and Theoretical Foundations of International Law Concerning Hate Speech' (1996) 14 Berkeley Journal of International Law 1.

Fertmann M and Kettemann MC (eds), *Viral Information: How States and Platforms Deal with Covid-19-Related Disinformation; an Exploratory Study of 20 Countries* (Verlag Hans-Bredow-Institut 2021).

Fiano N, 'Antisemitismo E Negazionismo. Un Fenomeno Ancora Attuale' in Marilisa D'Amico and Cecilia Siccardi (eds), *La Costituzione non odia: Conoscere, prevenire e contrastare l'*hate speech on line (Giappichelli 2021).

——, 'Il Linguaggio Dell'Odio in Germania: Tra *Wehrhafte Demokratie* e *Netzwerkdurchsetzungsgesetz*' in Marilisa D'Amico and Cecilia Siccardi (eds), *La Costituzione non odia: Conoscere, prevenire e contrastare l'*hate speech on line (Giappichelli 2021).

Floridi L, *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality* (Oxford University Press 2014).

——, 'The Fight for Digital Sovereignty: What It Is, and Why It Matters, Especially for the EU' (2020) 33 Philosophy & Technology 369.

Franks MA, 'How the Internet Unmakes Law Distinguished Lecture Series on the State of Internet Law' (2020) 16 Ohio State Technology Law Journal 10.

Frantziou E, 'The Horizontal Effect of the Charter: Towards an Understanding of Horizontality as a Structural Constitutional Principle' (2020) 22 Cambridge Yearbook of European Legal Studies 208.

Fredman S, 'Emerging from the Shadows: Substantive Equality and Article 14 of the European Convention on Human Rights' (2016) 16 Human Rights Law Review 273.

——, 'Substantive Equality Revisited' (2016) 14 International Journal of Constitutional Law 712.

Frosio G, 'Why Keep a Dog and Bark Yourself? From Intermediary Liability to Responsibility' (2018) 26 International Journal of Law and Information Technology 1.

Frosio G and Husovec M, 'Accountability and Responsibility of Online Intermediaries' in Giancarlo Frosio (ed), *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020).

Gagliardone I and others, *Countering Online Hate Speech* (UNESCO Publishing 2015).

Gardbaum S, 'The "Horizontal Effect" of Constitutional Rights' (2003) 102 Michigan Law Review 387.

Garg T and others, 'Handling Bias in Toxic Speech Detection: A Survey' (2023) 55 ACM Computing Surveys 264:1.

Gellert R, ‘Understanding the Notion of Risk in the General Data Protection Regulation’ (2018) 34 Computer Law & Security Review 279.

——, *The Risk-Based Approach to Data Protection* (Oxford University Press 2020).

Gerards J, ‘Non-Discrimination, the European Court of Justice and the European Court of Human Rights: Who Takes the Lead?’ in Thomas Giegerich (ed), *The European Union as Protector and Promoter of Equality* (Springer 2020).

Gerrard Y and Thornham H, ‘Content Moderation: Social Media’s Sexist Assemblages’ (2020) 22 New Media & Society 1266.

Gillespie T, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press 2018).

Gliszczyńska-Grabias A and Jabłoński M, ‘Is One Offended Pole Enough to Take Critics of Official Historical Narratives to Court?’ (*Verfassungsblog*, 12 October 2019) <https://verfassungsblog.de/is-one-offended-pole-enough-to-take-critics-of-official-historical-narratives-to-court/> accessed 16 August 2023.

Goldman E, ‘The Ten Most Important Section 230 Rulings’ (2017) 20 Tulane Journal of Technology and Intellectual Property 1.

——, ‘Why Section 230 Is Better than the First Amendment’ (2019) 95 Notre Dame Law Review Reflection 33.

——, ‘An Overview of the United States’ Section 230 Internet Immunity’ in Giancarlo Frosio (ed), *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020).

Golia AJ, ‘L’Antifascismo Della Costituzione Italiana Alla Prova Degli Spazi Giuridici Digitali. Considerazioni Su Partecipazione Politica, Libertà D’Espressione *Online* E Democrazia (Non) Protetta In *CasaPound c. Facebook* E *Forza Nuova c. Facebook*’ (2020) 18 Federalismi.it 134.

——, ‘The Transformative Potential of Meta’s Oversight Board: Strategic Litigation within the Digital Constitution?’ (2023) 30 Indiana Journal of Global Legal Studies 325.

Gomez R and others, ‘Exploring Hate Speech Detection in Multimodal Publications’, *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2020).

Gonçalves ME, ‘The Risk-Based Approach under the New EU Data Protection Regulation: A Critical Perspective’ (2020) 23 Journal of Risk Research 139.

Gorwa R, Binns R and Katzenbach C, ‘Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance’ (2020) 7 Big Data & Society 2053951719897945.

Gotanda N, ‘A Critique of Our Constitution Is Color-Blind’ (1991) 44 Stanford Law Review 1.

Grandinetti O, '*Facebook vs. CasaPound e Forza Nuova*, Ovvero La Disattivazione Di Pagine Social E Le Insidie Della Disciplina Multilivello Dei Diritti Fondamentali' (2021) 1 Rivista di Diritto dei Media 173.

Greenberg MH, 'A Return to Lilliput: The *LICRA v. Yahoo!* Case and the Regulation of Online Content in the World Market' (2003) 18 Berkeley Technology Law Journal 1191.

Griffin R, 'Rethinking Rights in Social Media Governance: Why fundamental rights are not enough to remedy the injustices of contemporary social media' (*Verfassungsblog*, 25 February 2022) <https://verfassungsblog.de/rethinking-rights/> accessed 8 June 2023.

Griffin R and Vander Maelen C, 'Codes of Conduct in the Digital Services Act: Exploring the Opportunities and Challenges' (SSRN, 30 May 2023) <https://papers.ssrn.com/abstract=4463874> accessed 14 June 2023.

Grimmelmann J, 'The Virtues of Moderation' (2015) 17 Yale Journal of Law and Technology 42.

Guardian (The), 'Polish Appeals Court Overturns Ruling against Holocaust Historians' *The Guardian* (16 August 2021) <https://www.theguardian.com/world/2021/aug/16/polish-appeals-court-overturns-ruling-against-holocaust-historians> accessed 16 August 2023.

Gupta I and Srinivasan L, 'Evolving Scope of Intermediary Liability in India' (2023) 37 International Review of Law, Computers & Technology 294.

Haas J, 'Freedom of the Media and Artificial Intelligence' (Global Conference for Media Freedom, 16 November 2020) <https://www.international.gc.ca/world-monde/assets/pdfs/issues_development-enjeux_developpement/human_rights-droits_homme/policy-orientation-ai-ia-en.pdf> accessed 2 August 2022.

Hackmann J, 'Defending the "Good Name" of the Polish Nation: Politics of History as a Battlefield in Poland, 2015-18' (2018) 20 Journal of Genocide Research 587.

Haggart B and Iglesias Keller C, 'Democratic Legitimacy in Global Platform Governance' (2021) 45 Telecommunications Policy 102152.

Haimson OL and others, 'Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas' (2021) 5 Proceedings of the ACM on Human-Computer Interaction 466:1.

Halmai G, 'Memory Politics in Hungary: Political Justice without Rule of Law' (*Verfassungsblog*, 10 January 2018) <https://verfassungsblog.de/memory-politics-in-hungary-political-justice-without-rule-of-law/> accessed 16 August 2023.

Hare I, 'Extreme Speech Under International and Regional Human Rights Standards' in Ivan Hare and James Weinstein (eds), *Extreme Speech and Democracy* (Oxford University Press 2009).

Hartvigsen T and others, 'ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection' (arXiv, 14 July 2022) <http://arxiv.org/abs/2203.09509> accessed 24 November 2023.

Haupt CE, 'Regulating Speech Online: Free Speech Values in Constitutional Frames' (2021) 99 Washington University Law Review 751.

He H, 'Online Intermediary Liability for Defamation under Chinese Laws' (2013) <https://www.law.uw.edu/media/1403/china-intermediary-liability-of-isps-defamation.pdf> accessed 25 September 2023.

Heinze E, 'Wild-West Cowboys versus Cheese-Eating Surrender Monkeys: Some Problems in Comparative Approaches to Hate Speech' in Ivan Hare and James Weinstein (eds), *Extreme Speech and Democracy* (Oxford University Press 2009).

Helberger N and others, 'A Freedom of Expression Perspective on AI in the Media – with a Special Focus on Editorial Decision Making on Social Media Platforms and in the News Media' (2020) 11 European Journal of Law and Technology 1.

——, 'Regulation of News Recommenders in the Digital Services Act: Empowering David against the Very Large Online Goliath' (*Internet Policy Review*, 26 February 2021) <https://policyreview.info/articles/news/regulation-news-recommenders-digital-services-act-empowering-david-against-very-large> accessed 13 June 2023.

Helberger N, Karppinen K and D'Acunto L, 'Exposure Diversity as a Design Principle for Recommender Systems' (2018) 21 Information, Communication & Society 191.

Helm T, 'Labour Pledges to Toughen "Weakened and Gutted" Online Safety Bill' *The Observer* (1 January 2023) <https://www.theguardian.com/technology/2023/jan/01/labour-pledges-toughen-online-safety-bill> accessed 20 August 2023.

Henley J, 'Poland Provokes Israeli Anger with Holocaust Speech Law' *The Guardian* (1 February 2018) <https://www.theguardian.com/world/2018/feb/01/poland-holocaust-speech-law-senate-israel-us> accessed 16 August 2023.

——, 'Fears for Polish Holocaust Research as Historians Ordered to Apologise' *The Guardian* (9 February 2021) <https://www.theguardian.com/world/2021/feb/09/fears-polish-holocaust-research-historians-ordered-apologise> accessed 16 August 2023.

Hermida PC de Q and dos Santos EM, 'Detecting Hate Speech in Memes: A Review' (2023) 56 Artificial Intelligence Review 12833.

Hong M, 'Regulating Hate Speech and Disinformation Online While Protecting Freedom of Speech as an Equal and Positive Right – Comparing Germany, Europe and the United States' (2022) 14 Journal of Media Law 76.

Horsman G, 'The Challenges Surrounding the Regulation of Anonymous Communication Provision in the United Kingdom' (2016) 56 Computers & Security 151.

Howe A, 'Supreme Court Skeptical of Texas, Florida Regulation of Social Media Moderation' (*SCOTUSblog*, 26 February 2024)

<https://www.scotusblog.com/2024/02/supreme-court-skeptical-of-texas-florida-regulation-of-social-media-moderation/> accessed 26 April 2024.

Huhn WR, 'The State Action Doctrine and the Principle of Democratic Choice' (2006) 34 Hofstra Law Review 1379.

Human Rights Watch, 'Germany: Flawed Social Media Law' (*Human Rights Watch*, 14 February 2018) <https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law> accessed 12 July 2023.

Hutchinson B and others, 'Unintended Machine Learning Biases as Social Barriers for Persons with Disabilities' (SIGACCES, October 2019) <http://sigaccess.org/newsletter/2019-10/hutchinson.html> accessed 12 April 2022.

Husovec M and Roche Laguna I, 'Digital Services Act: A Short Primer' (SSRN, 5 July 2022) 1 <https://papers.ssrn.com/abstract=4153796> accessed 6 June 2023.

Hutter BM, 'Risk, Regulation, and Management' in Peter Taylor-Gooby and Jens O Zinn (eds), *Risk in Social Science* (Oxford University Press 2006).

Jahn J, 'Strong on Hate Speech, Too Strict on Political Debate: The ECtHR Rules on Politicians' Obligation to Delete Hate Speech on Facebook Page' (*Verfassungsblog*, 25 May 2023) <https://verfassungsblog.de/strong-on-hate-speech-too-strict-on-political-debate/> accessed 1 June 2023.

Jaursch J, 'Platform Oversight: Here Is What a Strong Digital Services Coordinator Should Look Like' in Joris van Hoboken and others (eds), *Putting the DSA into Practice: Enforcement, Access to Justice, and Global Implications* (Verfassungsbooks 2023).

Jougleux P, *Facebook and the (EU) Law: How the Social Network Reshaped the Legal Framework* (Springer 2022).

Kaplan AM and Haenlein M, 'Users of the World, Unite! The Challenges and Opportunities of Social Media' (2010) 53 Business Horizons 59.

Kasneci E and others, 'ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education' (2023) 103 Learning and Individual Differences 102274.

Keane D, 'Attacking Hate Speech under Article 17 of the European Convention on Human Rights' (2007) 25 Netherlands Quarterly of Human Rights 641.

Keller D, 'Facebook Filters, Fundamental Rights, and the CJEU's *Glawischnig-Piesczek* Ruling' (2020) 69 GRUR International 616.

Kendrick L, 'Content Discrimination Revisited' (2012) 98 Virginia Law Review 231.

Kiela D and others, 'The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes' in Hugo Larochelle and others (eds), *Advances in Neural Information Processing Systems* (Curran Associates, Inc 2021).

Kiska R, ‘Hate Speech: A Comparison between the European Court of Human Rights and the United States Supreme Court Jurisprudence’ (2012) 25 Regent University Law Review 10.

Klonick K, ‘The New Governors: The People, Rules, and Processes Governing Online Speech’ (2017) 131 Harvard Law Review 1598.

——, ‘The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression’ (2020) 129 Yale Law Journal 2418.

Koltay A, *New Media and Freedom of Expression: Rethinking the Constitutional Foundations of the Public Sphere* (Hart 2019).

Kosseff J, *The Twenty-Six Words That Created the Internet* (Cornell University Press 2019).

Kovács Z, ‘Portrayal and Promotion – Hungary’s LGBTQI+ Law Explained’ *Euractiv* (24 June 2021) <https://www.euractiv.com/section/non-discrimination/news/portrayal-and-promotion-hungarys-latest-anti-lgbt-law-explained/> accessed 16 August 2023.

Kruzel J and Chung A, ‘US Supreme Court Weighs If Public Officials Can Block Critics on Social Media’ *Reuters* (31 October 2023) <https://www.reuters.com/legal/us-supreme-court-decide-if-public-officials-can-block-critics-social-media-2023-10-31/> accessed 27 December 2023.

Kuczerawy A, ‘General Monitoring Obligations: A New Cornerstone of Internet Regulation in the EU?’ in Centre for IT & IP Law (ed), *Rethinking IT and IP law: Celebrating 30 years CiTiP* (Intersentia 2020).

——, ‘From “Notice and Takedown” to “Notice and Stay Down”: Risks and Safeguards for Freedom of Expression’ in Giancarlo Frosio (ed), *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020).

——, ‘Does Twitter Trump Trump?’ (*Verfassungsblog*, 29 January 2021) <https://verfassungsblog.de/twitter-trump-trump/> accessed 10 January 2024.

Kukliš Ľ, ‘Video-Sharing Platforms in AVMSD: A New Kind of Content Regulation’ in Pier Luigi Parcu and Elda Brogi (eds), *Research Handbook on EU Media Law and Policy* (Edward Elgar Publishing 2021).

Laidlaw EB, *Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibility* (Cambridge University Press 2015).

Lakier G, ‘The Invention of Low-Value Speech’ (2015) 128 Harvard Law Review 2166.

Lamanuzzi M, ‘Il “Lato Oscuro Della Rete”: Odio e Pornografia Non Consensuale. Ruolo e Responsabilità Delle Piattaforme Social Oltre La *Net Neutrality’* (2021) 2 La Legislazione Penale 254.

Langton R, ‘Beyond Belief: Pragmatics in Hate Speech and Pornography’ in Ishani Maitra and Mary Kate McGowan (eds), *Speech & Harm: Controversies Over Free Speech* (Oxford University Press 2012).

——, ‘The Authority of Hate Speech’ in John Gardner, Leslie Green and Brian Leiter (eds), *Oxford Studies in Philosophy of Law*, vol 3 (Oxford University Press 2018).

Langton R, Haslanger S and Anderson L, ‘Language and Race’ in Gillian Russell and Delia Graff Fara (eds), *The Routledge Companion to Philosophy of Language* (Routledge 2012).

Lara Gálvez JC, ‘La Defensa de La Libertad de Expresión, La Ciberseguridad, y El Derecho a Una Información Veraz Frente a Las Fake News y La Neutralidad de Internet’ in Renata Ávila and others, *Derechos digitales en Iberoamérica: situación y perspectivas* (Fundación Carolina 2023).

LeCunY, Bengio Y and Hinton G, ‘Deep Learning’ (2015) 521 Nature 436.

Leerssen P, ‘The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems’ (2020) 11 European Journal of Law and Technology <https://ejlt.org/index.php/ejlt/article/view/786> accessed 13 November 2023.

——, ‘An End to Shadow Banning? Transparency Rights in the Digital Services Act between Content Moderation and Curation’ (2023) 48 Computer Law & Security Review 105790.

Leiser MR, ‘Analysing the European Union’s Digital Services Act Provisions for the Curtailment of Fake News, Disinformation, & Online Manipulation’ (SSRN, 24 April 2023) <https://papers.ssrn.com/abstract=4427493> accessed 14 June 2023.

Leiter B, ‘Cleaning Cyber-Cesspools: Google and Free Speech’ in Saul Levmore and Martha C Nussbaum (eds), *The Offensive Internet: Privacy, Speech, and Reputation* (Harvard University Press 2010) 155.

Lewis A, *Freedom for the Thought That We Hate* (Basic Books 2008).

Liptak A, ‘Supreme Court to Decide Whether Officials Can Block Critics on Social Media’ *The New York Times* (24 April 2023) <https://www.nytimes.com/2023/04/24/us/elected-officials-social-media-supreme-court.html> accessed 18 September 2023.

——, ‘Biden Asks Supreme Court to Lift Limits on Contacts With Social Media Sites’ *The New York Times* (14 September 2023) <https://www.nytimes.com/2023/09/14/us/politics/supreme-court-social-media-misinformation.html> accessed 18 September 2023.

——, ‘Supreme Court to Hear Challenges to State Laws on Social Media’ *The New York Times* (29 September 2023) <https://www.nytimes.com/2023/09/29/us/supreme-court-social-media-first-amendment.html> accessed 1 October 2023.

Liu B, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions* (2nd edn, Cambridge University Press 2020).

Llansó E and others, ‘Artificial Intelligence, Content Moderation, and Freedom of Expression’ (TWG 2020) <https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf> accessed 13 December 2021.

Lobba P, ‘Holocaust Denial before the European Court of Human Rights: Evolution of an Exceptional Regime’ (2015) 26 European Journal of International Law 237.

——, ‘From Introduction to Implementation: First Steps of the EU Framework Decision 2008/913/JHA against Racism and Xenophobia’ in Paul Behrens, Nicholas Terry and Olaf Jensen (eds), *Holocaust and Genocide Denial* (Routledge 2017).

Loewenstein K, ‘Militant Democracy and Fundamental Rights, I’ (1937) 31 The American Political Science Review 417.

Lomas N, ‘Meta Urged to Pay Reparations for Facebook’s Role in Rohingya Genocide’ (*TechCrunch*, 29 September 2022) <https://techcrunch.com/2022/09/29/amnesty-report-facebook-rohingya-reparations/> accessed 26 January 2023.

——, ‘Elon Musk Takes Twitter out of the EU’s Disinformation Code of Practice’ (*TechCrunch*, 27 May 2023) <https://techcrunch.com/2023/05/27/elon-musk-twitter-eu-disinformation-code/> accessed 14 June 2023.

Lulz S and Riegner M, ‘Freedom of Expression and Hate Speech’ in Philipp Dann and Arun K Thiruvengadam (eds), *Democratic Constitutionalism in India and the European Union: Comparing the Law of Democracy in Continental Polities* (Edward Elgar Publishing 2021).

MacAvaney S and others, ‘Hate Speech Detection: Challenges and Solutions’ (2019) 14 PLOS ONE e0221152.

Macenaite M, ‘The “Riskification” of European Data Protection Law through a Two-Fold Shift’ (2017) 8 European Journal of Risk Regulation 506.

MacKinnon CA, ‘Pornography as Defamation and Discrimination’ (1991) 71 Boston University Law Review 793.

——, ‘Substantive Equality: A Perspective’ (2011) 96 Minnesota Law Review 1.

——, ‘Substantive Equality Revisited: A Reply to Sandra Fredman’ (2016) 14 International Journal of Constitutional Law 739.

MacKinnon R and others, *Fostering Freedom Online: The Role of Internet Intermediaries* (UNESCO Publishing 2014).

Maitra I, ‘Subordinating Speech’ in Ishani Maitra and Mary Kate McGowan (eds), *Speech & Harm: Controversies Over Free Speech* (Oxford University Press 2012).

Mälksoo M, ‘Militant Memocracy in International Relations: Mnemonical Status Anxiety and Memory Laws in Eastern Europe’ (2021) 47 Review of International Studies 489.

Marino G, ‘Semiotics of Spreadability: A Systematic Approach to Internet Memes and Virality’ (2015) 1 Punctum 43.

Maroni M, ‘The Liability of Internet Intermediaries and the European Court of Human Rights’ in Bilyana Petkova and Tuomas Ojanen (eds), *Fundamental Rights Protection Online: The Future Regulation of Intermediaries* (Edward Elgar Publishing 2020).

Marrey Moncau LF and Werneck Arguelhes D, 'The Marco Civil Da Internet and Digital Constitutionalism' in Giancarlo Frosio (ed), *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020).

Marsden CT, *Internet Co-Regulation: European Law, Regulatory Governance and Legitimacy in Cyberspace* (Cambridge University Press 2011).

Martin C, 'Striking the Right Balance: Hate Speech Laws in Japan, the United States, and Canada' (2018) 45 Hastings Constitutional Law Quarterly 455.

Matsuda MJ, 'Public Response to Racist Speech: Considering the Victim's Story' (1989) 87 Michigan Law Review 2320.

——, 'Dissent in a Crowded Theater' (2019) 72 SMU Law Review 441.

Matsuda MJ and others (eds), *Words That Wound: Critical Race Theory, Assaultive Speech, And The First Amendment* (Westview Press 1993).

Mazzoli EM and Tambini D, 'Prioritisation Uncovered: The Discoverability of Public Interest Content Online' (Council of Europe 2020) DGI(2020)19 <https://rm.coe.int/publication-content-prioritisation-report/1680a07a57> accessed 26 May 2022.

Mazzoni V, 'Far Right Extremism on Telegram: A Brief Overview' (*European Eye on Radicalization*, 14 March 2019) <https://eeradicalization.com/far-right-extremism-on-telegram-a-brief-overview/> accessed 7 June 2023.

McGonagle T, 'The Council of Europe against Online Hate Speech: Conundrums and Challenges' (Council of Europe 2013) MCM(2013)005.

——, 'General Recommendation 35 on Combating Racist Hate Speech' in David Keane and Annapurna Waughray (eds), *Fifty Years of the International Convention on the Elimination of all Forms of Racial Discrimination: A Living Instrument* (Manchester University Press 2017).

McKinnon S, '"Building a Thick Skin for Each Other": The Use of "Reading" as an Interactional Practice of Mock Impoliteness in Drag Queen Backstage Talk' (2017) 6 Journal of Language and Sexuality 90.

Mehrabi N and others, 'A Survey on Bias and Fairness in Machine Learning' (arXiv, 25 January 2022) <http://arxiv.org/abs/1908.09635> accessed 21 November 2023.

Meiklejohn A, *Free Speech And Its Relation to Self-Government* (1st edn, Harper & Brothers 1948).

Meta, 'Proactive Rate' (*Transparency Center*, 22 February 2023) <https://transparency.fb.com/policies/improving/proactive-rate-metric/> accessed 8 December 2023.

——, 'Community Standards Enforcement: Hate Speech' (*Transparency Center*) <https://transparency.meta.com/reports/community-standards-enforcement/hate-speech/facebook/> accessed 28 April 2024.

——, ‘Counterspeech’ (*Counterspeech*) <https://counterspeech.fb.com/en/> accessed 6 November 2023.

——, ‘Facebook Community Standards’ (*Transparency Center*) <https://transparency.fb.com/policies/community-standards/> accessed 8 December 2023.

——, ‘Facebook Community Standards: Hate Speech’ (*Transparency Center*) <https://transparency.fb.com/policies/community-standards/hate-speech/> accessed 8 December 2023.

Meta Oversight Board, ‘Oversight Board Charter’ (*Oversight Board*, February 2023) <https://oversightboard.com/attachment/494475942886876/> accessed 25 October 2023.

Milano S, Taddeo M and Floridi L, ‘Recommender Systems and Their Ethical Challenges’ (2020) 35 AI & Society 957.

Miles T, ‘U.N. Investigators Cite Facebook Role in Myanmar Crisis’ *Reuters* (12 March 2018) <https://www.reuters.com/article/us-myanmar-rohingya-facebook-idUSKCN1GO2PN> accessed 26 January 2023.

Milkaite I, ‘A Picture of a Same-Sex Kiss on Facebook Wreaks Havoc: Beizaras and Levickas v. Lithuania’ (*Strasbourg Observers*, 7 February 2020) <https://strasbourgobservers.com/2020/02/07/a-picture-of-a-same-sex-kiss-on-facebook-wreaks-havoc-beizaras-and-levickas-v-lithuania/> accessed 16 January 2023.

Milmo D, ‘Rohingya Sue Facebook for £150bn over Myanmar Genocide’ *The Guardian* (6 December 2021) <https://www.theguardian.com/technology/2021/dec/06/rohingya-sue-facebook-myanmar-genocide-us-uk-legal-action-social-media-violence> accessed 26 January 2023.

——, ‘Twitter Sues Anti-Hate Speech Group over “Tens of Millions of Dollars” in Lost Advertising’ *The Guardian* (2 August 2023) <https://www.theguardian.com/technology/2023/aug/02/twitter-accuses-anti-hate-speech-group-over-tens-of-millions-of-dollars-in-lost-advertising> accessed 8 November 2023.

Mittelstadt BD and others, ‘The Ethics of Algorithms: Mapping the Debate’ (2016) 3 Big Data & Society 205395171667967.

Monti M, ‘La proposta del ddl Zanda-Filippin sul contrasto alle fake news sui social network: profili problematici’ (*Diritti Comparati*, 7 December 2017) <https://www.dirittocomparati.it/la-proposta-del-ddl-zanda-filippin-sul-contrasto-alle-fake-news-sui-social-network-profili-problematici/> accessed 11 August 2023.

Morais Carvalho J, Arga e Lima F and Farinha M, ‘Introduction to the Digital Services Act, Content Moderation and Consumer Protection’ (2021) 3 Revista de Direito e Tecnologia 71.

Morandini A, ‘Recalibrating Platforms’ AI Systems: EU advances’ (*MediaLaws*, 10 July 2023) <https://www.medialaws.eu/recalibrating-platforms-ai-systems-eu-advances/> accessed 7 December 2023.

Moreno Belloso N and Petit N, 'The EU Digital Markets Act (DMA): A Competition Hand in a Regulatory Glove' (2023) 48 European Law Review 391.

Mouron P, 'Du Sénat Au Conseil Constitutionnel: Adoption Des Lois de Lutte Contre La Manipulation de l'information' (2019) 49 Revue européenne des médias et du numérique 9.

Mozur P, 'A Genocide Incited on Facebook, With Posts From Myanmar's Military' *The New York Times* (15 October 2018) <https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html> accessed 26 January 2023.

Nardocci C, 'L'Odio Razziale E Religioso' in Marilisa D'Amico and Cecilia Siccardi (eds), *La Costituzione non odia: Conoscere, prevenire e contrastare l'*hate speech on line (Giappichelli 2021).

——, 'Dalla Parola Che Discrimina Alla Parità Nel Linguaggio: La Dimensione Sovranazionale (E Comparata)' in Marina Brambilla and others (eds), *Genere, disabilità, linguaggio. Progetti e prospettive a Milano* (Franco Angeli 2022).

——, 'Artificial Intelligence-Based Discrimination: Theoretical and Normative Responses. Perspectives from Europe' (2023) 60 DPCE Online 2367.

Nascimento FRS, Cavalcanti GDC and Da Costa-Abreu M, 'Exploring Automatic Hate Speech Detection on Social Media: A Focus on Content-Based Analysis' (2023) 13 SAGE Open 21582440231181311.

Nash V, 'Revise and Resubmit? Reviewing the 2019 Online Harms White Paper' (2019) 11 Journal of Media Law 18.

Nash V and Felton L, 'Treating the Symptoms or the Disease? Analysing the UK Online Safety Bill's Approach to Digital Regulation' (SSRN, 2 June 2023) <https://papers.ssrn.com/abstract=4467382> accessed 23 August 2023.

Nave E and Lane L, 'Countering Online Hate Speech: How Does Human Rights Due Diligence Impact Terms of Service?' (2023) 51 Computer Law & Security Review 105884.

Newton C, 'Facebook Will Create an Independent Oversight Group to Review Content Moderation Appeals' (*The Verge,* 15 November 2018) <https://www.theverge.com/2018/11/15/18097219/facebook-independent-oversight-supreme-court-content-moderation> accessed 25 October 2023.

Nicholas G, 'Shedding Light on Shadowbanning' (*Center for Democracy and Technology,* April 2022) <https://cdt.org/insights/shedding-light-on-shadowbanning/> accessed 22 July 2022.

Noble SU, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York University Press 2018).

Nunziato DC, 'The Digital Services Act and the Brussels Effect on Platform Content Moderation' (2023) 24 Chicago Journal of International Law 115.

O'Connell R, 'Cinderella Comes to the Ball: Art 14 and the Right to Non-Discrimination in the ECHR' (2009) 29 Legal Studies 211.

Oetheimer M, 'Protecting Freedom of Expression: The Challenge of Hate Speech in the European Court of Human Rights Case Law Symposium: Comparative Law of Hate Speech' (2009) 17 Cardozo Journal of International and Comparative Law 427.

O'Kane R, 'Meta's Private Speech Governance and the Role of the Oversight Board: Lessons from the Board's First Decisions' (2022) 25 Stanford Technology Law Review 167.

Oliva TD, 'Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression' (2020) 20 Human Rights Law Review 607.

Oliva TD, Antonialli DM and Gomes A, 'Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online' (2021) 25 Sexuality & Culture 700.

O'Reilly T, 'What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software' (2007) 1 Communications & Strategies 17.

Pagallo U, Casanovas P and Madelin R, 'The Middle-out Approach: Assessing Models of Legal Governance in Data Protection, Artificial Intelligence, and the Web of Data' (2019) 7 The Theory and Practice of Legislation 1.

Pappalardo K and Suzor NP, 'The Liability of Australian Online Intermediaries' in Giancarlo Frosio (ed), *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020).

Pariser E, *The Filter Bubble: What the Internet Is Hiding From You* (Penguin 2011).

Park JH, Shin J and Fung P, 'Reducing Gender Bias in Abusive Language Detection', *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics 2018).

Park KS, 'From Liability Trap to the World's Safest Harbour: Lessons from China, India, Japan, South Korea, Indonesia, and Malaysia' in Giancarlo Frosio (ed), *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020).

Partsch KJ, 'Racial Speech and Human Rights: Article 4 of the Convention on the Elimination of All Forms or Racial Discrimination' in Sandra Coliver (ed), *Hate Speech, Freedom of Expression and Non-Discrimination* (Article 19 1992).

Pasquale F, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2015).

——, *New Laws of Robotics: Defending Human Expertise in the Age of AI* (The Belknap Press of Harvard University Press 2020).

Perel M and Elkin-Koren N, 'Accountability in Algorithmic Copyright Enforcement' (2016) 19 Stanford Technology Law Review 473.

——, ‘Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement’ (2017) 69 Florida Law Review 181.

Peršak N, ‘Criminalising Hate Crime and Hate Speech at EU Level: Extending the List of Eurocrimes Under Article 83(1) TFEU’ (2022) 33 Criminal Law Forum 85.

Perset K, ‘The Economic and Social Role of Internet Intermediaries’ (OECD 2010) <https://www.oecd-ilibrary.org/content/paper/5kmh79zzs8vb-en> accessed 13 April 2023

Pettit P, *Republicanism: A Theory of Freedom and Government* (Clarendon Press, Oxford University Press 1997).

Pitruzzella G and Pollicino O, *Disinformation and Hate Speech* (Bocconi University Press 2020).

Pitruzzella G, Pollicino O and Quintarelli S, *Parole e Potere: Libertà d’Espressione, Hate Speech e Fake News* (Egea 2017).

Pohle J and Thiel T, ‘Digital Sovereignty’ (2020) 9 Internet Policy Review 1.

Pollicino O, ‘Fake News, Internet and Metaphors (to Be Handled Carefully)’ (2017) 1 Rivista di Diritto dei Media 23.

——, ‘Judicial Protection of Fundamental Rights in the Transition from the World of Atoms to the World of Bits: The Case of Freedom of Speech’ (2019) 25 European Law Journal 155.

——, ‘L’“Autunno Caldo” Della Corte Di Giustizia in Tema Di Tutela Dei Diritti Fondamentali in Rete e Le Sfide Del Costituzionalismo Alle Prese Con i Nuovi Poteri Privati in Ambito Digitale’ (2019) 19 Federalismi.it 1.

——, *Judicial Protection of Fundamental Rights on the Internet: A Road Towards Digital Constitutionalism?* (Hart 2021).

——, ‘Potere Digitale’ in Marta Cartabia and Marco Ruotolo (eds), *Enciclopedia del Diritto*, vol. *Potere e Costituzione* (Giuffrè 2023).

——, ‘The Quadrangular Shape of the Geometry of Digital Power(s) and the Move towards a Procedural Digital Constitutionalism’ (2023) 29 European Law Journal 10.

Pollicino O and Bassini M, ‘Free Speech, Defamation and the Limits to Freedom of Expression in the EU: A Comparative Analysis’ in Andrej Savin and Jan Trzaskowski (eds), *Research Handbook on EU Internet Law* (Edward Elgar Publishing 2014).

Pollicino O, Bassini M and De Gregorio G, *Internet Law and Protection of Fundamental Rights* (Bocconi University Press 2022).

Pollicino O and De Gregorio G, ‘A Constitutional-Driven Change of Heart: ISP Liability and Artificial Intelligence in the Digital Single Market’ in Giuliana Ziccardi Capaldo (ed), *The Global Community Yearbook of International Law and Jurisprudence 2018* (Oxford University Press 2019).

——, ‘Constitutional Law in the Algorithmic Society’ in Amnon Reichman and others (eds), *Constitutional Challenges in the Algorithmic Society* (Cambridge University Press 2021).

——, ‘Shedding Light on the Darkness of Content Moderation: The First Decisions of the Facebook Oversight Board’ (*Verfassungsblog*, 5 February 2021) <https://verfassungsblog.de/fob-constitutionalism/> accessed 25 October 2023.

Popper K, *The Open Society and Its Enemies*, vol I: *The Spell of Plato* (Routledge 1945).

Porter J, ‘The UK’s Tortured Attempt to Remake the Internet, Explained’ *The Verge* (4 May 2023) <https://www.theverge.com/23708180/united-kingdom-online-safety-bill-explainer-legal-pornography-age-checks> accessed 20 August 2023.

Pozzi FA and others, ‘Challenges of Sentiment Analysis in Social Networks: An Overview’ in Federico Alberto Pozzi and others (eds), *Sentiment Analysis in Social Networks* (Morgan Kaufmann 2017).

Quelle C, ‘Enhancing Compliance under the General Data Protection Regulation: The Risky Upshot of the Accountability- and Risk-Based Approach’ (2018) 9 European Journal of Risk Regulation 502.

Quintais JP, ‘The New Copyright in the Digital Single Market Directive: A Critical Look’ (2020) 42 European Intellectual Property Review 28.

Quintais JP, Appelman N and Fahy RÓ, ‘Using Terms and Conditions to Apply Fundamental Rights to Content Moderation’ (2023) 24 German Law Journal 881.

Quintais JP, De Gregorio G and Magalhães JC, ‘How Platforms Govern Users’ Copyright-Protected Content: Exploring the Power of Private Ordering and Its Implications’ (2023) 48 Computer Law & Security Review 105792.

Quintel T and Ullrich C, ‘Self-Regulation of Fundamental Rights? The EU Code of Conduct on Hate Speech, Related Initiatives and Beyond’ in Bilyana Petkova and Tuomas Ojanen (eds), *Fundamental Rights Protection Online: The Future Regulation of Intermediaries* (Edward Elgar Publishing 2020).

Redish MH, ‘The Content Distinction in First Amendment Analysis’ (1981) 34 Stanford Law Review 113.

Reidenberg JR, ‘Yahoo and Democracy on the Internet’ (2001) 42 Jurimetrics 261.

Resta G, ‘Anonimato, Responsabilità, Identificazione: Prospettive Di Diritto Comparato’ (2014) 2 Il diritto dell’informazione e dell’informatica 171.

Roberts ST, *Behind the Screen: Content Moderation in the Shadows of Social Media* (Yale University Press 2019).

Rosenfeld M, ‘Hate Speech in Constitutional Jurisprudence: A Comparative Analysis’ (2002) 24 Cardozo Law Review 1523.

Ryan-Mosley T, ‘How Generative AI Is Boosting the Spread of Disinformation and Propaganda’ (*MIT Technology Review*, 4 October 2023)

<https://www.technologyreview.com/2023/10/04/1080801/generative-ai-boosting-disinformation-and-propaganda-freedom-house/> accessed 8 December 2023.

Sap M and others, 'The Risk of Racial Bias in Hate Speech Detection' in Anna Korhonen, David Traum and Màrquez Lluís (eds), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics 2019) <https://www.aclweb.org/anthology/P19-1163> accessed 22 October 2021.

Sartor G and Loreggia A, 'The Impact of Algorithms for Online Content Filtering or Moderation. "Upload Filters"' (European Parliament 2020) JURI Committee PE 657.101.

Schauer F, 'The Exceptional First Amendment' in Michael Ignatieff (ed), *American Exceptionalism and Human Rights* (Princeton University Press 2005).

Schmidt A and Wiegand M, 'A Survey on Hate Speech Detection Using Natural Language Processing' in Lun-Wei Ku and Cheng-Te Li (eds), *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (Association for Computational Linguistics 2017).

Schulz W, 'Regulating Intermediaries to Protect Personality Rights Online - The Case of the German NetzDG' in Marion Albers and Ingo Wolfgang Sarlet (eds), *Personality and Data Protection Rights on the Internet: Brazilian and German Approaches* (Springer 2022).

Searle JR, 'What Is a Speech Act?' in Maurice Black (ed), *Philosophy in America* (Allen and Unwin 1965).

——, 'Austin on Locutionary and Illocutionary Acts' (1968) 77 The Philosophical Review 405.

——, *Expression and Meaning: Studies in the Theory of Speech Acts* (Cambridge University Press 1979).

——, 'J.L. Austin (1911-1960)' in Aloysius Patrick Martinich and David Sosa (eds), *A Companion to Analytic Philosophy* (Blackwell 2001).

Semenzin S and Bainotti L, 'The Use of Telegram for Non-Consensual Dissemination of Intimate Images: Gendered Affordances and the Construction of Masculinities' (2020) 6 Social Media + Society 2056305120984453.

Senden LAJ and others, 'Mapping Self- and Co-Regulation Approaches in the EU Context: Explorative Study for the European Commission, DG Connect' (Utrecht University Repository, 2015) <https://dspace.library.uu.nl/handle/1874/327305> accessed 10 July 2023.

Senftleben M, 'Bermuda Triangle: Licensing, Filtering and Privileging User-Generated Content Under the Directive on Copyright in the Digital Single Market' (2019) 41 European Intellectual Property Review 480.

Severino C, 'La Democrazia Francese e Le Sfide Del Digitale: Tra Opportunità e Rischi' (2021) 3 Rivista Gruppo di Pisa 33

Siccardi C, ‘La *Loi Avia*. La Legge Francese Contro l’Odio *On Line* (O Quello Che Ne Rimane)’ in Marilisa D’Amico and Cecilia Siccardi (eds), *La Costituzione non odia: Conoscere, prevenire e contrastare l’*hate speech on line (Giappichelli 2021).

Siegel AA, ‘Online Hate Speech’ in Joshua A Tucker and Nathaniel Persily (eds), *Social Media and Democracy: The State of the Field, Prospects for Reform* (Cambridge University Press 2020).

Simonite T, ‘Facebook’s AI for Hate Speech Improves. How Much Is Unclear’ (*Wired*, 12 May 2021) <https://www.wired.com/story/facebook-ai-hate-speech-improves-unclear/> accessed 14 December 2021.

Sissons M, ‘Our Commitment to Human Rights’ (*Meta*, 16 March 2021) <https://about.fb.com/news/2021/03/our-commitment-to-human-rights/> accessed 4 November 2023.

Smolla RA, ‘The Meaning of the “Marketplace of Ideas” in First Amendment Law’ (2019) 24 Communication Law and Policy 437.

Sorkin AR and others, ‘The Deplatforming of President Trump’ *The New York Times* (8 January 2021) <https://www.nytimes.com/2021/01/08/business/dealbook/trump-facebook-twitter-deplatforming.html> accessed 10 January 2024.

Spano R, ‘Intermediary Liability for Online User Comments under the European Convention on Human Rights’ (2017) 17 Human Rights Law Review 665.

Spigno I, *Discorsi d’odio. Modelli Costituzionali a Confronto* (Giuffrè 2018).

Stone GR, ‘Content Regulation and the First Amendment’ (1983) 25 William & Mary Law Review 189.

——, ‘The Origins of the Bad Tendency Test: Free Speech in Wartime’ (2002) 2002 Supreme Court Review 411.

Sunstein CR, *#Republic: Divided Democracy in the Age of Social Media* (Princeton University Press 2017).

Suzor NP, ‘Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms’ (2018) 4 Social Media + Society 2056305118787812.

——, *Lawless: The Secret Rules That Govern Our Digital Lives* (Cambridge University Press 2019).

Taddeo M and Floridi L (eds), *The Responsibilities of Online Service Providers* (Springer 2017).

Tambini D, ‘The Differentiated Duty of Care: A Response to the Online Harms White Paper’ (2019) 11 Journal of Media Law 28.

Tan R, ‘Social Media Platforms Duty of Care – Regulating Online Hate Speech’ (2022) 37 Australasian Parliamentary Review 143.

Temperman J, ‘Blasphemy versus Incitement: An International Law Perspective’ in Christopher S Grenda, Chris Beneke and David Nash (eds), *Profane: Sacrilegious Expression in a Multicultural Age* (University of California Press 2014).

Than K, ‘Hungary Vows to Fight in EU Court to Defend Anti-LGBT Law’ *Reuters* (9 March 2023) <https://www.reuters.com/world/europe/hungary-vows-fight-eu-court-defend-anti-lgbt-law-2023-03-09/> accessed 16 August 2023.

Thornberry P, *The International Convention on the Elimination of All Forms of Racial Discrimination: A Commentary* (Oxford University Press 2016).

TikTok, ‘Safety and Civility: Hate Speech and Hateful Behaviors’ (*TikTok*, 8 March 2023) <https://www.tiktok.com/community-guidelines/en/safety-civility/> accessed 10 November 2023.

Tondo L, ‘“Disgraceful”: Italy’s Senate Votes down Anti-Homophobic Violence Bill’ (*The Guardian*, 27 October 2021) <https://www.theguardian.com/world/2021/oct/27/italy-senate-votes-down-anti-homophobic-violence-bill> accessed 14 June 2023.

Tourkochoriti I, ‘The Digital Services Act and the EU as the Global Regulator of the Internet’ (2023) 24 Chicago Journal of International Law 129.

Trengove M and others, ‘A Critical Review of the Online Safety Bill’ (2022) 3 Patterns 100544.

Tribe LH, *American Constitutional Law* (2nd edn, Foundation Press 1988).

Tsesis A, ‘Hate in Cyberspace: Regulating Hate Speech on the Internet’ (2001) 38 San Diego Law Review 817.

——, ‘Dignity and Speech: The Regulation of Hate Speech in a Democracy Articles & Essays’ (2009) 44 Wake Forest Law Review 497.

Tulkens F, ‘When to Say Is To Do: Freedom of Expression and Hate Speech in the Case-Law of the European Court of Human Rights’ (Seminar on Human Rights for European Judicial Trainers, Strasbourg, 7 July 2015).

Tushnet R, ‘Power Without Responsibility: Intermediaries and the First Amendment’ (2008) 76 The George Washington Law Review 986.

Tworek H and Leerssen P, ‘An Analysis of Germany’s NetzDG Law’ (TWG 2019) 1 <https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf> accessed 12 July 2023.

Van Blarcum CD, ‘Internet Hate Speech: The European Framework and the Emerging American Haven’ (2005) 62 Washington and Lee Law Review 781.

Van de Kerkhof J, ‘Good Faith in Article 6 Digital Services Act (Good Samaritan Exemption)’ (*The Digital Constitutionalist*, 15 February 2023) <https://digicon.org/good-faith-in-article-6-digital-services-act-good-samaritan-exemption/> accessed 24 December 2023.

Van der Heijden J, 'Risk as an Approach to Regulatory Governance: An Evidence Synthesis and Research Agenda' (2021) 11 SAGE Open <https://doi.org/10.1177/21582440211032202> accessed 11 April 2022.

Van Eecke P, 'Online Service Providers and Liability: A Plea for a Balanced Approach' (2011) 48 Common Market Law Review 1455.

Vander Maelen C, 'Hardly Law or Hard Law? Investigating the Dimensions of Functionality and Legislation of Codes of Conduct in Recent EU Legislation and the Normative Repercussions Thereof' (2022) 47 European Law Review 752.

Vander Maelen C and Griffin R, 'Twitter's Retreat from the Code of Practice on Disinformation Raises a Crucial Question: Are DSA Codes of Conduct Really Voluntary?' (*DSA Observatory*, 12 June 2023) <https://dsa-observatory.eu/2023/06/12/twitters-retreat-from-the-code-of-practice-on-disinformation-raises-a-crucial-question-are-dsa-codes-of-conduct-really-voluntary/> accessed 14 June 2023.

Vermeule A, *The Constitution of Risk* (Cambridge University Press 2013).

Vigevani GE, 'Anonimato, Responsabilità e Trasparenza Nel Quadro Costituzionale Italiano' (2014) 2 Il diritto dell'informazione e dell'informatica 207.

——, 'Informazione e Potere' in Marta Cartabia and Marco Ruotolo (eds), *Enciclopedia del Diritto*, vol. *Potere e Costituzione* (Giuffrè 2023) 219.

Villaschi P, 'I Progetti Di Legge In Discussione In Italia: Analisi Critica' in Marilisa D'Amico and Cecilia Siccardi (eds), *La Costituzione non odia: Conoscere, prevenire e contrastare l'*hate speech on line (Giappichelli 2021).

——, 'La (Non) Regolamentazione Dei *Social Network* E Del *Web*' in Marilisa D'Amico and Cecilia Siccardi (eds), *La Costituzione non odia: Conoscere, prevenire e contrastare l'*hate speech on line (Giappichelli 2021).

Voorhoof D, 'Internet and the Right of Anonymity' in Jelena Surculija (ed), Proc*eedings of the conference Regulating the Internet, Belgrade, 2010* (Center for Internet Development 2011).

——, 'Qualification of News Portal as Publisher of Users' Comment May Have Far-Reaching Consequences for Online Freedom of Expression: Delfi AS v. Estonia' (*Strasbourg Observers*, 25 October 2013) <https://strasbourgobservers.com/2013/10/25/qualification-of-news-portal-as-publisher-of-users-comment-may-have-far-reaching-consequences-for-online-freedom-of-expression-delfi-as-v-estonia/> accessed 26 April 2023.

——, 'Delfi AS v. Estonia: Grand Chamber Confirms Liability of Online News Portal for Offensive Comments Posted by Its Readers' (*Strasbourg Observers*, 18 June 2015) <https://strasbourgobservers.com/2015/06/18/delfi-as-v-estonia-grand-chamber-confirms-liability-of-online-news-portal-for-offensive-comments-posted-by-its-readers/> accessed 26 April 2023.

——, ‘Blog Symposium “Strasbourg Observers Turns Ten” (2): The Court’s Subtle Approach of Online Media Platforms’ Liability for User-Generated Content since the “Delfi Oracle”’ (*Strasbourg Observers*, 10 April 2020) <https://strasbourgobservers.com/2020/04/10/the-courts-subtle-approach-of-online-media-platforms-liability-for-user-generated-content-since-the-delfi-oracle/> accessed 6 May 2023.

Wachter S, Mittelstadt B and Russell C, ‘Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-Discrimination Law’ (2020) 123 West Virginia Law Review 735.

Waldron J, *The Harm in Hate Speech* (Harvard University Press 2012).

Walker S, *Hate Speech: The History of an American Controversy* (University of Nebraska Press 1994).

Weber A, *Manual on Hate Speech* (Council of Europe Publishing 2009).

Williams SH, ‘Content Discrimination and the First Amendment’ (1991) 139 University of Pennsylvania Law Review 615.

Wilman F, ‘Between Preservation and Clarification: The Evolution of the DSA’s Liability Rules in Light of the CJEU’s Case Law’ in Joris van Hoboken and others (eds), *Putting the DSA into Practice: Enforcement, Access to Justice, and Global Implications* (Verfassungsbooks 2023).

Wilson R and Land M, ‘Hate Speech on Social Media: Content Moderation in Context’ (2021) 52 Connecticut Law Review 1029.

Wischmeyer T, ‘What Is Illegal Offline Is Also Illegal Online: The German Network Enforcement Act 2017’ in Bilyana Petkova and Tuomas Ojanen (eds), F*undamental Rights Protection Online: The Future Regulation of Intermediaries* (Edward Elgar Publishing 2020).

Wong D and Floridi L, ‘Meta’s Oversight Board: A Review and Critical Assessment’ (2023) 33 Minds and Machines 261.

Woods L, ‘The Duty of Care in the Online Harms White Paper’ (2019) 11 Journal of Media Law 6.

Wu T, ‘Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-Ordering Systems’ (2019) 119 Columbia Law Review 2001.

X, ‘X’s Policy on Hateful Conduct’ (*X Help Center*) <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy> accessed 8 November 2023.

Yannopoulos GN, ‘The Immunity of Internet Intermediaries Reconsidered?’ in Mariarosaria Taddeo and Luciano Floridi (eds), *The Responsibilities of Online Service Providers* (Springer 2017).

YouTube, ‘Hate Speech Policy’ (*Google Help*) <https://support.google.com/youtube/answer/2801939?hl=en#zippy=%2Cother-types-of-content-that-violates-this-policy%2Ceducational-documentary-scientific-and-artistic-content%2Cmore-examples> accessed 10 November 2023.

Ypma P and others, *Study to Support the Preparation of the European Commission's Initiative to Extend the List of EU Crimes in Article 83 of the Treaty on the Functioning of the EU to Hate Speech and Hate Crime: Final Report* (Publications Office of the European Union 2021) <https://data.europa.eu/doi/10.2838/04029> accessed 9 April 2024.

Zecca D, 'Tutela Dell'Integrità Dell'Informazione e Della Comunicazione in Rete: Obblighi per Le Piattaforme Digitali Fra Fonti Comunitarie e Disciplina Degli Stati Membri' (2019) 37 DPCE Online 889.

Zhou X and others, 'Challenges in Automated Debiasing for Toxic Language Detection' in Paola Merlo, Jorg Tiedemann and Reut Tsarfaty (eds), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (Association for Computational Linguistics 2021).

Zhu Y and others, 'Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks' (arXiv, 22 April 2023) <http://arxiv.org/abs/2304.10145> accessed 23 November 2023.

Ziccardi G, *L'Odio Online: Violenza Verbale e Ossessioni in Rete* (Raffaello Cortina 2016).

——, *Online Political Hate Speech in Europe: The Rise of New Extremisms* (Edward Elgar Publishing 2020).

Zingales N, 'Intermediary Liability in Africa: Looking Back, Moving Forward?' in Giancarlo Frosio (ed), *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020).

——, 'The DSA as a Paradigm Shift for Online Intermediaries' Due Diligence: Hail To Meta-Regulation' in Joris van Hoboken and others (eds), *Putting the DSA into Practice: Enforcement, Access to Justice, and Global Implications* (Verfassungsbooks 2023).

Zuleta L and Burkal R, 'Hate Speech in the Public Online Debate' (The Danish Institute for Human Rights 2017).

Zurth P, 'The German NetzDG as Role Model or Cautionary Tale? Implications for the Debate on Social Media Liability' (2021) 31 Fordham Intellectual Property, Media and Entertainment Law Journal 1084.

Institutional sources

AUSTRALIA

Digital Industry Group, 'Disinformation Code' (*DIGI*) <https://digi.org.au/disinformation-code/> accessed 24 September 2023.

COUNCIL OF EUROPE

Committee of Ministers of the Council of Europe, 'Recommendation No. R (97) 20 of the Committee of Ministers to Member States on "Hate Speech"' (Council of Europe 1997) CM/Rec(97)20.

——, 'Recommendation No. R (2010) 5 of the Committee of Ministers to Member States on Measures to Combat Discrimination on Grounds of Sexual Orientation or Gender Identity' (Council of Europe 2010) CM/Rec(2010)5.

——, 'Recommendation No. R (2018) 2 of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries' (Council of Europe 2018) CM/Rec(2018)2

——, 'Recommendation No. R (2022) 16 of the Committee of Ministers to Member States on Combating Hate Speech' (Council of Europe 2022) CM/Rec(2022)16.

Committee on Artificial Intelligence, 'Draft Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law' (Council of Europe 2024) CM(2024)52-prov1 <https://rm.coe.int/-1493-10-1b-committee-on-artificial-intelligence-cai-b-draft-framework/1680aee411> accessed 4 May 2024.

European Commission against Racism and Intolerance, 'General Policy Recommendation No. 15 on Combating Hate Speech' (Council of Europe 2015) CRI(2016)5.

European Court of Human Rights, 'Guide on Article 17 of the European Convention on Human Rights – Prohibition of Abuse of Rights' (Council of Europe 2022) <https://www.echr.coe.int/Documents/Guide_Art_17_ENG.pdf> accessed 6 April 2023.

EUROPEAN UNION

European Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. A Digital Single Market Strategy for Europe' COM(2015) 192 final.

——, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Online Platforms and the Digital Single Market Opportunities and Challenges for Europe' COM(2016) 288 final.

——, 'Delivery of Comments Pursuant to Article 5(2) of Directive (EU) 2015/1535 of 9 September 2015. Law Aimed at Combating Hate Content on the Internet' C(2019) 8585 final.

——, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. A Union of Equality: Gender Equality Strategy 2020-2025' COM(2020) 152 final.

——, ‘Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Union of Equality: LGBTIQ Equality Strategy 2020-2025’ COM(2020) 698 final.

——, ‘Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions On the European Democracy Action Plan’ COM(2020) 790 final

——, ‘Communication of 15 December 2020, 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.

——, ‘Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Union of Equality: Strategy for the Rights of Persons with Disabilities 2021-2030’ COM(2021) 101 final.

——, ‘Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, A Renewed EU Strategy 2011-14 for Corporate Social Responsibility’ COM(2021) 681 final.

——, ‘Communication from the Commission to the European Parliament and the Council. A More Inclusive and Protective Europe: Extending the List of EU Crimes to Hate Speech and Hate Crime’ COM(2021) 777 final.

——, ‘Communication of 11 May 2022, Proposal for a Regulation of the European Parliament and of the Council Laying down Rules to Prevent and Combat Child Sexual Abuse’ COM(2022) 209 final.

——, ‘EU Code of Conduct against Online Hate Speech: Latest Evaluation Shows Slowdown in Progress’ (*European Commission*, 24 November 2022) <https://ec.europa.eu/commission/presscorner/detail/en/ip_22_7109> accessed 15 June 2023.

——, ‘Digital Services Act: Commission Designates First Set of Very Large Online Platforms and Search Engines’ (*European Commission*, 25 April 2023) <https://ec.europa.eu/commission/presscorner/detail/en/IP_23_2413> accessed 12 June 2023.

——, ‘The Digital Services Act Package’ (*European Commission*, 12 May 2023) <https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package> accessed 2 June 2023.

——, ‘Commission Designates Second Set of Very Large Online Platforms under the Digital Services Act’ (*European Commission*, 20 December 2023) <https://digital-strategy.ec.europa.eu/en/news/commission-designates-second-set-very-large-online-platforms-under-digital-services-act> accessed 28 December 2023.

——, ‘The EU Code of Conduct on Countering Illegal Hate Speech Online’ (*European Commission*) <https://commission.europa.eu/strategy-and-policy/policies/justice-and-

fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en> accessed 30 May 2023.

European Parliament, 'Report on Extending the List of EU Crimes to Hate Speech and Hate Crime' (2023) 2023/2068(INI).

Reynders D, '7th Evaluation of the Code of Conduct' (European Commission 2022) <https://commission.europa.eu/system/files/2022-12/Factsheet%20-%207th%20monitoring%20round%20of%20the%20Code%20of%20Conduct.pdf> accessed 30 May 2023.

FRANCE

Assemblée Nationale, '*Proposition de Loi n 1785 Visant à Lutter Contre Les Contenus Haineux Sur Internet*' (*Assemblée Nationale*, 20 March 2019) <https://www.assemblee-nationale.fr/dyn/15/textes/l15b1785_proposition-loi#D_non_amendable_0> accessed 10 August 2023.

ORGANIZATION FOR SECURITY AND COOPERATION IN EUROPE

Office for Democratic Institutions and Human Rights, 'Hate Crime Laws: A Practical Report' (2nd edn, OSCE 2022) <https://www.osce.org/files/f/documents/1/4/523940.pdf> accessed 9 January 2023.

UNITED NATIONS

Committee on the Elimination of Racial Discrimination, 'Positive Measures Designed to Eradicate All Incitement to, or Acts of, Racial Discrimination: Implementation of the International Convention on the Elimination of All Forms of Racial Discrimination, Article 4' (United Nations 1986) CERD/2.

——, 'General Recommendation No. 35. Combating Racist Hate Speech' (United Nations 2013) CERD/C/GC/35.

De Varennes F, 'Recommendations Made by the Forum on Minority Issues at Its Thirteenth Session on the Theme "Hate Speech, Social Media and Minorities"' (United Nations 2021) A/HRC/46/58.

Human Rights Committee, 'General Comment No. 11. Prohibition of Propaganda for War and Inciting National, Racial or Religious Hatred (Art. 20)' (United Nations 1983).

——, 'General Comment No. 34. Article 19: Freedom of Opinion and Expression' (United Nations 2011) CCPR/C/GC/34.

——, 'Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred That Constitutes Incitement, to Discrimination, Hostility or Violence' (United Nations 2013) A/HRC/22/17/Add.4.

Izsák R, 'Report of the Special Rapporteur on Minority Issues' (United Nations 2015) A/HRC/28/64.

Kaye D, ‘Comment on the Social Networks Bill (Netzdurchführungsgesetz)’ (Office of the High Commissioner for Human Rights 2017) OL DEU 1/2017 < https://www.ohchr.org/en/special-procedures/sr-freedom-of-opinion-and-expression/comments-legislation-and-policy> accessed 12 July 2023.

——, ‘Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression’ (Human Rights Council 2018) A/HRC/38/35.

Office of the High Commissioner for Human Rights, ‘Guiding Principles on Business and Human Rights: Implementing the United Nations “Protect, Respect and Remedy” Framework’ (United Nations 2011) HR/PUB/11/04 <https://www.ohchr.org/sites/default/files/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf> accessed 4 November 2023.

United Nations, ‘Countering COVID-19 Hate Speech’ (*United Nations Secretary-General*, 2020) <https://www.un.org/sg/en/node/251827> accessed 15 December 2021.

United Nations High Commissioner for Human Rights, ‘Incitement to Racial and Religious Hatred and the Promotion of Tolerance’ (United Nations 2006) A/HRC/2/6.

UNITED KINGDOM

Department for Digital, Culture, Media and Sport, and the Home Office, ‘Online Harms White Paper’ (HM Government 2019) CP 57.

Law Commission, ‘Hate Crime Laws: A Consultation Paper’ (2020) Law Com CP 250.

Case Law

AUSTRALIA

Clarke v Nationwide News Pty Ltd (2012) 289 ALR 345.

Silberberg v Builders Collective of Australia Inc (2007) 164 FCR 475.

EUROPEAN COURT OF JUSTICE

Case C-81/12, *Asociaţia Accept v Consiliul Naţional pentru Combaterea Discriminării* [2013] ECLI:EU:C:2013:275.

Case C-360/10, *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV* [2012] ECLI:EU:C:2012:85.

Case C-54/07, *Centrum voor gelijkheid van kansen en voor racismebestrijding v Firma Feryn NV* [2008] ECLI:EU:C:2008:397.

Case C-18/18, *Eva Glawischnig-Piesczek v Facebook Ireland Limited* [2019] ECLI:EU:C:2019:821.

Joined Cases C-682/18 and C-683/18, *Frank Peterson v Google LLC and Others* and *Elsevier Inc v Cyando AG* [2021] ECLI:EU:C:2021:503.

Case C-152/73, *Giovanni Maria Sotgiu v Deutsche Bundespost* [1974] ECLI:EU:C:1974:13.

Joined Cases C-236/08, C-237/08 and C-238/08, *Google France SARL and Google Inc v Louis Vuitton Malletier SA*, *Google France SARL v Viaticum SA and Luteciel SARL* and *Google France SARL v Centre national de recherche en relations humaines (CNRRH) SARL and Others* [2010] ECLI:EU:C:2010:159.

Case C-376/22, *Google Ireland Limited and Others v Kommunikationsbehörde Austria* [2023] ECLI:EU:C:2023:467, Opinion of AG Szpunar.

Case C-376/22, *Google Ireland Limited and Others v Kommunikationsbehörde Austria* [2023] ECLI:EU:C:2023:835.

Case C-507/17, *Google LLC, successor in law to Google Inc, v Commission nationale de l'informatique et des libertés (CNIL)* [2019] ECLI:EU:C:2019:772.

Case C-131/12, *Google Spain SL and Google Inc v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González* [2014] ECLI:EU:C:2014:317.

Case C-324/09, *L'Oréal SA and Others v eBay International AG and Others* [2011] ECLI:EU:C:2011:474.

Case C-507/18, *NH v Associazione Avvocatura per i diritti LGBTI - Rete Lenford* [2020] ECLI:EU:C:2020:289.

Case C-406/15, *Petya Milkova v Izpalnitelen direktor na Agentsiata za privatizatsia i sledprivatizatsionen kontrol* [2017] ECLI:EU:C:2017:198.

Case C-401/19, *Republic of Poland v European Parliament, Council of the European Union* [2022] ECLI:EU:C:2022:297.

Case C-157/15, *Samira Achbita and Centrum voor gelijkheid van kansen en voor racismebestrijding v G4S Secure Solutions NV* [2017] ECLI:EU:C:2017:203.

Case C-70/10, *Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM)* [2011] ECLI:EU:C:2011:771.

EUROPEAN COURT OF HUMAN RIGHTS

Ahmet Yıldırım v Turkey [2012] ECtHR 3111/10, ECHR 2012.

Anguelova v Bulgaria [2002] ECtHR 38361/97, ECHR 2002-IV.

Association Accept and Others v Romania [2021] ECtHR 19237/16.

Ayoub and others v France [2020] ECtHR 77400/14, 34532/15, 34550/15.

Behar and Gutman v Bulgaria [2021] ECtHR 29335/13.

Beizaras and Levickas v Lithuania [2020] ECtHR 41288/15.

Bonnet v France (dec) [2022] ECtHR 35364/19.

Delfi AS v Estonia [2015] ECtHR [GC] 64569/09, ECHR 2015.

Editorial Board of Pravoye Delo and Shtekel v Ukraine [2011] ECtHR 33014/05, ECHR 2011.

Erbakan v Turkey [2006] ECtHR 59405/00.

Féret v Belgium [2009] ECtHR 15615/07.

Garaudy v France (dec) [2003] ECtHR 65831/01, ECHR 2003-IX.

Glimmerveen and Hagenbeek v the Netherlands [1979] ECommHR 8348/78, 8406/78, 18 Decisions and Reports 187.

Gunduz v Turkey [2003] ECtHR 35071/97, ECHR 2003-XI.

Handyside v the United Kingdom [1976] ECtHR 5493/72, Series A 24.

Høiness v Norway [2019] ECtHR 43624/14.

Jersild v Denmark [1994] ECtHR [GC] 15890/89, Series A 298.

Jezior v Poland [2020] ECtHR 31955/11.

KU v Finland [2008] ECtHR 2872/02, ECHR 2008.

Kühnen v the Federal Republic of Germany [1988] ECommHR 12194/86, 56 Decisions and Reports 205.

Lehideux and Isorni v France [1998] ECtHR [GC] 24662/94, Reports 1998-VII.

Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v Hungary [2016] ECtHR 22947/13.

M'bala M'bala v France (dec) [2015] ECtHR 25239/13, ECHR 2015-VIII.

Molnar v Romania (dec) [2012] ECtHR 16637/06.

Norwood v the United Kingdom (dec) [2004] ECtHR 23131/03, ECHR 2004-XI.

Observer and Guardian v the United Kingdom [1991] ECtHR 13585/88, Series A 216.

Pavel Ivanov v Russia (dec) [2007] ECtHR 35222/04.

Perinçek v Switzerland [2015] ECtHR [GC] 27510/08, ECHR 2015.

Pihl v Sweden (dec) [2017] ECtHR 74742/14.

Remer v Germany [1995] ECommHR 25096/94.

Sanchez v France [2021] ECtHR 45581/15.

Sanchez v France [2023] ECtHR [GC] 45581/15, ECHR 2023.

Standard Verlagsgesellschaft Mbh v Austria (no 3) [2021] ECtHR 39378/15.

Stoll v Switzerland [2007] ECtHR [GC] 69698/01, ECHR 2007-V.

Taddeucci and McCall v Italy [2016] ECtHR 51362/09.

Valaitis v Lithuania [2023] ECtHR 39375/19.

Vejdeland and others v Sweden [2012] ECtHR 1813/07, ECHR 2012.

Witzsch v Germany (2) (dec) [2005] ECtHR 7485/03.

X v the Federal Republic of Germany [1982] ECommHR 9235/81, 29 Decisions and Reports 194.

Zemmour v France [2022] ECtHR 63539/19.

FRANCE

Cons Const (20 December 2018) 2018-773 DC, *Loi relative à la lutte contre la manipulation de l'information*.

Cons Const (18 June 2020) 2020-801 DC, *Loi visant à lutter contre les contenus haineux sur internet*.

TGI Paris (22 May 2000) RG 00/05308, *Ligue internationale contre le racisme et l'antisémitisme et Union des étudiants juifs de France v Yahoo!, Inc et Yahoo! France*.

TGI Paris (10 May 2019) RG 19/53935, *Vieu et Ouzoulias v Twitter France SAS*.

GERMANY

BVerfG (22 May 2019) 1 BvQ 42/19.

INDIA

Shreya Singhal v Union of India AIR 2015 SC 1523.

ITALY

Tribunal of Rome, specialized section for business, order of 12 December 2019.

Tribunal of Rome, specialized section for the rights of the person and civil immigration, order of 23 February 2020.

Tribunal of Rome, XVII civil section, order of 29 April 2020.

Tribunal of Siena, civil section, order of 19 January 2020.

HUMAN RIGHTS COMMITTEE

Malcolm Ross v Canada [2000] Human Rights Committee CCPR/C/70/D/736/1997.

META OVERSIGHT BOARD

Alleged crimes in Raya Kobo [2021] 2021-014-FB-UA.

Armenians in Azerbaijan [2021] 2020-003-FB-UA.

Dehumanizing Comments About People in Gaza [2024] 2024-026-FB-UA.

Depiction of Zwarte Piet [2021] 2021-002-FB-UA.

Fictional Assault on Gay Couple [2023] 2023-051-FB-UA.

Hateful Memes Video Montage [2024] 2024-015-FB-UA.

Heritage of Pride [2023] 2023-058-IG-UA.

Holocaust Denial [2024] 2023-022-IG-UA.

Knin cartoon [2022] 2022-001-FB-UA.

Media Conspiracy Cartoon [2023] 2023-042-FB-UA.

Myanmar post about Muslims [2021] 2020-002-FB-UA.

Planet of the Apes racism [2023] 2023-035-FB-UA.

Reclaiming Arabic words [2022] 2022-003-IG-UA.

Russian poem [2022] 2022-008-FB-UA.

South Africa slurs [2021] 2021-011-FB-UA.

Violence against women [2023] 2023-002-IG-UA, 2023-005-IG-UA.

Wampum belt [2021] 2021-012-FB-UA.

POLAND

Leszczyńska v Engelking and Grabowski [2021] Warsaw Court of Appeals, I Civil Division I ACa 300/21.

UNITED KINGDOM

Regina v Osborne [1732] W Kel 230, 25 Eng Rep 584.

UNITED STATES OF AMERICA

Abrams v United States 250 US 616 (1919).

Beauharnais v Illinois 343 US 250 (1951).

Biden v Knight First Amendment Institute at Columbia 593 US __ (2021).

Blum v Yaretsky 457 US 991 (1982) 1004.

Brandenburg v Ohio 395 US 444 (1969).

Chaplinsky v New Hampshire 315 US 568 (1942).

Cohen v California 403 US 15 (1971).

Cubby, Inc v CompuServe, Inc 776 FSupp 135 (SDNY 1991).

Davison v Randall 912 F3d 666 (4th Cir 2019).

Gonzalez v Google LLC 2 F4th 871 (9th Cir 2021).

Gonzalez v Google LLC 598 US 617 (2023).

Knight First Amendment Institute at Columbia v Trump 928 F3d 226 (2nd Cir 2019).

Lindke v Freed 601 US 187 (2024).

Manhattan Community Access Corp v Halleck 587 US __ (2019).

Marsh v State of Alabama 326 US 501 (1946).

Matal v Tam 582 US __ (2017).

McIntyre v Ohio Elections Commission 514 US 334 (1995).

NetChoice, LLC v Moody 546 FSupp3d 1082 (NDFla 2021).

NetChoice, LLC v Moody 34 F4th 1196 (11th Cir 2022).

NetChoice, LLC v Paxton 573 FSupp3d 1092 (WDTex 2021).

NetChoice, LLC v Paxton 2022 WL 1537249 (5th Cir 2022).

NetChoice, LLC v Paxton 596 US __ (2022).

NetChoice, LLC v Paxton 49 F4th 439 (5th Cir 2022).

Packingham v North Carolina 582 US __ (2017).

Police Department of the City of Chicago v Mosley 508 US 92 (1972).

Prager University v Google LLC 951 F3d 991 (9th Cir 2020).

RAV v City of St Paul 505 US 377 (1992).

Reno v American Civil Liberties Union 521 US 844 (1997).

Schenck v United States 249 US 47 (1919).

Smith v California 361 US 147 (1959).

Snyder v Phelps 562 US 443 (2011).

State v Arlene's Flowers, Inc 441 P3d 1203 (Wash 2019).

Stratton Oakmont, Inc v Prodigy Services Co 23 Media L Rep 1794 (NY Sup Ct 1995).

Twitter, Inc v Taamneh 598 US 471 (2023).

Village of Skokie v Nat'l Socialist Party of America 373 NE2d 21 (Ill 1978).

Virginia v Black 538 US 343 (2003).

Volokh v James 2023 WL 1991435 (SDNY 2023).

Whitney v California 274 US 357 (1927).

Yahoo! Inc v La Ligue Contre Le Racisme Et L'Antisemitisme 169 FSupp2d 1181 (NDCal 2001).

Yahoo! Inc v La Ligue Contre Le Racisme Et L'Antisemitisme 379 F.3d 1120 (9th Cir 2004).

Yahoo! Inc v La Ligue Contre Le Racisme Et L'Antisemitisme 433 F.3d 1199 (9th Cir 2006).

Zeran v America Online, Inc 129 F3d 327 (4th Cir 1997).

Table of Legislation

Treaty on the European Union.

Treaty on the Functioning of the European Union.

B. Regulations

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), OJ L/119.

Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online, OJ L 172/79.

Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act), OJ L 265/1.

Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act), OJ L 277/1.

Regulation (EU) 2024/... of the European Parliament and of the Council of ... laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

C. Directives

Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin (Racial Equality Directive), OJ L 180/22.

Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation (General Framework for Equal Treatment Directive), OJ L 303/16.

Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services (Directive on Gender Equality in Goods and Services), OJ L 373/37.

Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce'), OJ L 178/1.

Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast) (Recast Equal Treatment Directive), OJ L 204/23.

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) (OJ L 95/1).

Directive (EU) 2015/1535 of the European Parliament and of the Council of 9 September 2015 laying down a procedure for the provision of information in the field of technical regulations and of rules on Information Society services (codification), OJ L 241/1.

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

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 (DSM Copyright Directive), OJ L 130/92.

D. Framework Decisions

Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law, OJ L 328/55.

E. Recommendations

Commission Recommendation 2003/361/EC of 6 May 2003 concerning the definition of micro, small and medium-sized enterprises, OJ L124/36.

Commission Recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online, OJ L63/50.

F. Interinstitutional Agreements

Interinstitutional Agreement of 31 December 2003 on Better Law-Making, OJ C 321/1.

Interinstitutional Agreement between the European Parliament, the Council of the European Union and the European Commission of 13 April 2016 on Better Law-Making, OJ L 123/1.

G. Codes of Conduct/Practice

Code of Conduct on Countering Illegal Hate Speech Online 2016.

Code of Practice on Disinformation 2018.

Strengthened Code of Practice on Disinformation 2022.

France

Code Pénal.

Loi du 29 juillet 1881 sur la liberté de la presse.

Loi n 2004-575 du 21 juin 2004 pour la confiance dans l'économie numérique.

Loi organique n 2018-1201 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information.

Loi n 2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information.

Loi n 2020-766 du 24 juin 2018 visant à lutter contre les contenus haineux sur internet.

GERMANY

Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (*Netzwerkdurchsetzungsgesetz* - NetzDG) 2017 (BGBI I S 3352).

Gesetz zur Änderung des Netzwerkdurchsetzungsgesetzes 2021 (BGBI I S 1436).

Gesetz zur Durchführung der Verordnung (EU) 2021/784 des Europäischen Parlaments und des Rates vom 29. April 2021 zur Bekämpfung der Verbreitung terroristischer Online-Inhalte und zur Änderung weiterer Gesetze 2022 (BGBI I S 1182).

Strafgesetzbuch (StGB) in the version published on 13 November 1998 (BGBI I S 3322).

GHANA

Electronic Transactions Act 2008.

HUNGARY

Magyarország Alaptörvénye 2011.

T/332 számú javaslat. Magyarország Alaptörvényének hetedik módosítása 2018.

2012. évi C törvény a büntetőtörvénykönyvről 2012.

2021. évi LXXIX. törvény a pedofil bűnelkövetőkkel szembeni szigorúbb fellépésről, valamint a gyermekek védelme érdekében egyes törvények módosításáról 2021.

INDIA

Information Technology Act 2000.

Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules 2021.

INTERNATIONAL LAW

Charter of the International Military Tribunal appended to the Agreement by the government of the United Kingdom of Great Britain and Northern Ireland, the government of the United States of America, the provisional government of the French Republic and the government of the Union of Soviet Socialist Republics for the prosecution and

punishment of the major war criminals of the European Axis (UN Treaty Series No 251) 284.

International Convention on the Elimination of All Forms of Racial Discrimination 1965.

International Covenant on Civil and Political Rights 1966.

Rome Statute of the International Criminal Court 1998.

Universal Declaration of Human Rights 1948.

ITALY

AC 2936 (XVIII), *Misure per la prevenzione e il contrasto della diffusione di manifestazioni d'odio mediante la rete* internet.

AS 2688 (XVII), *Disposizioni per prevenire la manipolazione dell'informazione* online*, garantire la trasparenza sul* web *e incentivare l'alfabetizzazione mediatica*.

AS 3001 (XVII), *Norme generali in materia di* social network *e per il contrasto della diffusione su* internet *di contenuti illeciti e delle* fake news.

AS 634 (XVIII), *Modifiche al codice penale e altre disposizioni in materia di contrasto dell'istigazione all'odio e alla discriminazione* (hate speech).

AS 1455 (XVIII), *Misure per il contrasto del fenomeno dell'istigazione all'odio sul* web.

AS 2005 (XVIII), *Misure di prevenzione e contrasto della discriminazione e della violenza per motivi fondati sul sesso, sul genere, sull'orientamento sessuale, sull'identità di genere e sulla disabilità*.

Decreto legislativo 1 marzo 2018, n 21, Disposizioni di attuazione del principio di delega della riserva di codice nella materia penale a norma dell'articolo 1, comma 85, lettera q), della legge 23 giugno 2017, n 103.

Decreto legislativo 8 novembre 2021, n 208, Testo Unico dei Servizi di Media Audiovisivi (TUSMA).

Delibera n 37/23/CONS, Regolamento in materia di tutela dei diritti fondamentali della persona ai sensi dell'art. 30 del decreto legislativo 8 novembre 2021, n 208.

Legge 13 ottobre 1975, n 654, Ratifica ed esecuzione della convenzione internazionale sull'eliminazione di tutte le forme di discriminazione razziale, aperta alla firma a New York il 7 marzo 1966.

Legge 25 giugno 1993, n 205, Conversione in legge, con modificazioni, del decreto-legge 26 aprile 1993, n 122, recante misure urgenti in materia di discriminazione razziale, etnica e religiosa.

Regio decreto 19 ottobre 1930, n 1398, Codice penale.

JAPAN

Act on the Limitation of Liability for Damages of Specified Telecommunications Service Providers and the Right to Demand Disclosure of Identification Information of the Sender 2001.

Act on the Promotion of Efforts to Eliminate Unfair Discriminatory Speech and Behavior against Persons with Countries of Origin other than Japan 2016.

KENYA

Computer Misuse and Cybercrimes Act 2018.

MALAWI

Electronic Transactions and Cyber Security Act 2016.

ORGANIZATION OF AMERICAN STATES

American Convention on Human Rights (Pact of San José) 1969.

POLAND

Ustawa z dnia 18 grudnia 1998 r. o Instytucie Pamięci Narodowej - Komisji Ścigania Zbrodni przeciwko Narodowi Polskiemu, Dz.U. 1998 Nr 155 poz. 1016.

Ustawa z dnia 26 stycznia 2018 r. o zmianie ustawy o Instytucie Pamięci Narodowej - Komisji Ścigania Zbrodni przeciwko Narodowi Polskiemu, ustawy o grobach i cmentarzach wojennych, ustawy o muzeach oraz ustawy o odpowiedzialności podmiotów zbiorowych za czyny zabronione pod groźbą kary, Dz.U. 2018 poz. 369.

Ustawa z dnia 27 czerwca 2018 r. o zmianie ustawy o Instytucie Pamięci Narodowej - Komisji Ścigania Zbrodni przeciwko Narodowi Polskiemu oraz ustawy o odpowiedzialności podmiotów zbiorowych za czyny zabronione pod groźbą kary, Dz.U. 2018 poz. 1277.

SOUTH AFRICA

Cybercrimes Act 2020.

Electronic Communications and Transactions Act 2002.

SPAIN

Acuerdo suscrito entre el Consejo General del Poder Judicial, la Fiscalía General del Estado, el Ministerio de Justicia, el Ministerio de Interior, el Ministerio de Educación y Formación Profesional, el Ministerio de Trabajo, Migraciones y Seguridad Social, el Ministerio de la Presidencia, Relaciones con las Cortes e Igualdad, el Ministerio de Cultura y Deporte y el Centro de Estudios Jurídicos para cooperar

institucionalmente en la lucha contra el racismo, la xenofobia, la LGBTIfobia y otras formas de intolerancia 2018.

Ley Orgánica 10/1995, de 23 de noviembre, del Código Penal.

Ley 34/2002, de 11 de julio, de Servicios de la Sociedad de Información y Comercio Electrónico.

Ley 19/2007, de 11 de julio, contra la violencia, el racismo, la xenofobia y la intolerancia en el deporte.

Protocolo para combatir el discurso de odio ilegal en línea 2021.

UGANDA

Electronic Transactions Act 2011.

UNITED KINGDOM

Public Order Act 1986.

Malicious Communications Act 1988.

Crime and Disorder Act 1998.

Football (Offences) Act 1999.

Communications Act 2003.

Criminal Justice Act 2003.

Racial and Religious Hatred Act 2006.

Criminal Justice and Immigration Act 2008.

Online Safety Act 2023.

UNITED STATES OF AMERICA

Anti-Terrorism Act 1990.

Communications Decency Act 1996.

Digital Millennium Copyright Act 1998.

DISCOURSE Act, S. 2228, 117th Cong. (2021).

Fla SB 7072 on Social Media Platforms 2021.

Justice Against Sponsors of Terrorism Act 2016.

NY State Assembly Bill 2021-A7865A 2021.

Tex HB 20 relating to censorship of or certain other interference with digital expression, including expression on social media platforms or through electronic mail messages 2021.

21st Century FREE Speech Act, S. 1384, 117th Cong. (2021).

VENEZUELA

Ley Constitucional contra el Odio, por la Convivencia Pacífica y la Tolerancia 2017.

ZAMBIA

Electronic Communications and Transactions Act 2009.

Electronic Communications and Transactions Act 2021.