


ARTICLE

The Fight Against Digital Hate Speech: Disentangling the EU's Regulatory Approach and Hurdles

Evangelia Psychogiopoulou* 

University of the Peloponnese, Corinth; Hellenic Foundation for European and Foreign Policy (ELIAMEP), Athens, Greece
Email: e.psychogiopoulou@uop.gr

(Received 26 October 2023; accepted 16 April 2024; first published online 18 September 2024)

Abstract

Digital platforms and social media have expanded the ways in which individuals can exercise their right to freedom of expression and obtain and diffuse information. At the same time, they have become a principal means for haters to express and spread their hate in ways that would have been unthinkable some years ago. Responsive to the challenge, the EU has progressively developed a broad range of instruments and tools to counter online hate speech. This chapter discusses the key characteristics of the EU arrangements made to fight digital hate speech, shedding light on what is a multi-faceted and daunting regulatory task.

Keywords: Hate speech; Council Framework Decision 2008/913/JHA; Audiovisual Media Services Directive; Digital Services Act; Code of Conduct on Countering Illegal Hate Speech Online

A. Introduction

Digitalization and new technologies have transformed society in several ways, affecting various aspects of our everyday life. In particular, they have profoundly changed how we communicate with, connect to and interact with others; and how we seek, how we impart, and how we access information and ideas. Digital platforms and social media have heavily impacted the information and communication spheres. The contemporary media and information ecosystem is characterized by easy access to information, the ability to have your voice heard by many without geographical limitations, unprecedented levels of user interaction and engagement, and a range of capabilities facilitating users' participation in content creation. Services provided by platforms and social media have become essential pathfinders to information and knowledge. They also offer spaces for public debate and scrutiny and for shaping and influencing public opinion, providing opportunities for democratic citizenship while complementing traditional media in this respect. The speed with which information can circulate online, the power of algorithms when it comes to moderating or structuring content for consumption, and the broader impact content made available online can have—in political, social, and other terms—and especially illegal and harmful content—are also important attributes of the current media and information environment, which is largely platform-dominated.

*Evangelia Psychogiopoulou is an Assistant Professor at the Department of Political Science and International Relations of the University of the Peloponnese, and a senior research fellow at the Hellenic Foundation for European and Foreign Policy (ELIAMEP) in Greece.

As digitalization and online platforms reshape information and media markets, regulators in Europe are grappling with the problem of hate speech online. Although hate speech is not new, it has become of growing concern in recent years.¹ Over and above socio-economic factors—economic and social crises, migration, the COVID-19 pandemic, etcetera—digital innovation has played an important role in accentuating the phenomenon. By expanding the ways in which individuals can exercise the right to freedom of expression and the right to seek, receive and impart information, platforms and social media have concurrently provided a space in which hatred can spread.²

Hate speech lacks a universally accepted legally binding definition. In what should be seen as the first European attempt at reaching a common—yet non-binding—understanding of the concept, the Council of Europe (CoE) defined hate speech in 1997 as “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.”³ States party to the CoE were invited to “establish [...] a sound legal framework,” combining civil, criminal and administrative law provisions on hate speech,⁴ and narrowly circumscribe any interference with freedom of expression.⁵

At the EU level, along with broader action taken to combat discrimination,⁶ initial efforts to curb hate speech drew on hate speech becoming *illegal* speech under EU law. They have been complemented in recent years by regulatory action aimed at combatting hatred specifically in the digital environment. Although certain EU Member States have also sought national solutions of their own,⁷ platforms and digital intermediaries too have developed policies to address hate speech online. The latter have worked closely with the EU institutions, especially the European Commission (hereinafter Commission), to improve standards and enforcement practices.

Distinct types of, and models for, regulatory intervention to cope with hate speech have thus been introduced. This Article sets out to develop a better understanding of the ways in which the EU seeks to combat hate speech online. The analysis explores the various instruments employed

¹See Judit Bayer & Petra Bárd, *Hate Speech and Hate Crime in the EU and the Evaluation of Online Content Regulation Approaches*, Study commissioned by the European Parliament, Policy Department for Citizens' Rights and Constitutional Affairs (2020), [https://www.europarl.europa.eu/RegData/etudes/STUD/2020/655135/IPOL_STU\(2020\)655135_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2020/655135/IPOL_STU(2020)655135_EN.pdf); UN Secretary-General António Guterres' Global Appeal to Address and Counter COVID-19-Related Hate Speech (2020), <https://www.un.org/press/en/2020/sgsm20076.doc.htm>. See also Eur. Comm'n against Racism and Intolerance, *ECRI Annual Reports* (May 5, 2024), <https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/annual-reports>.

²According to a 2016 Eurobarometer survey, 75% of respondents following or participating in online debates reported having witnessed or experienced threat or hate speech and/or abuse directed at people active on social media, with almost half of them stating that this discouraged them from engaging in online discussions. See Eur. Comm'n, *Special Eurobarometer 452: Media Pluralism and Democracy*, 17–18 (Nov. 2016), https://ec.europa.eu/information_society/newsroom/image/document/2016-47/sp452-summary_en_19666.pdf.

³Council of Europe, Committee of Ministers, *Recommendation No R (97) 20 to Member States on “Hate Speech”*, app. (Oct. 30, 1997), <https://rm.coe.int/1680505d5b>.

⁴*Id.* at 107, principle 2.

⁵*Id.* at 108, principle 3.

⁶See ELISE MUIR, *EU EQUALITY LAW: THE FIRST FUNDAMENTAL RIGHTS POLICY OF THE EU* (2018). See also Paul Craig & Gráinne De Búrca eds., *Equal Treatment and Non-Discrimination*, in *EU LAW: TEXT, CASES AND MATERIALS* 929 (Paul Craig & Gráinne De Búrca eds., 7th ed., 2020); ULADZISLAU BELAVUSAU & KRISTIN HENRARD, *EU ANTI-DISCRIMINATION LAW BEYOND GENDER* (1st ed., 2019); Mark Bell, *EU Anti-Discrimination Law: Navigating Sameness and Difference*, in *THE EVOLUTION OF EU LAW* 651 (Paul Craig & Gráinne De Búrca eds., 3rd ed., 2021).

⁷See e.g. *Netzwerkdurchsetzungsgesetz [NetzDG]* [German Network Enforcement Act], Sept. 1, 2017, BGBl I at 3352; *LOI 2020-766 du 24 juin 2020 visant à lutter contre les contenus haineux sur internet* [Law 2020-766 of June 24, 2020 on Combating Hateful Content on the Internet] *JOURNAL OFFICIEL DE LA RÉPUBLIQUE FRANÇAISE [O.J.]* [OFFICIAL GAZETTE OF FRANCE], June 25, 2020 (certain provisions of this law were declared unconstitutional by the French Constitutional Council); *KOMMUNIKATIONSPLATTFORMEN-GESETZ [KoPI-G]* [COMMUNICATION PLATFORM ACT] *BUNDESGESETZBLATT [BGBl]* (2020) (Austria).

by the EU for that purpose, examining their main characteristics, strengths and weaknesses whilst shedding light on what is clearly a multi-faceted and daunting regulatory task.

B. The EU Regulatory Toolbox for Combatting (Digital) Hate Speech

All forms and manifestations of hatred and intolerance are incompatible with the values of respect for human dignity, freedom, democracy, equality, the rule of law, and respect for human rights upon which the EU is founded. Article 2 of the Treaty on European Union (TEU) enshrines these values, proudly proclaiming that they are common to Member States in a society in which pluralism, non-discrimination, tolerance, justice, solidarity and equality between women and men prevail. EU action against hate speech is a reflection of its attachment to its values and rests on an array of instruments that combine distinct regulatory approaches. Some of these instruments specifically address the digital environment; others do not.

I. Combatting Racist and Xenophobic Hate Speech Through Criminal Law

Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law is the Union's criminal law response to racism and xenophobia.⁸ As the title of the framework decision suggests, this is not an instrument focused on hate speech. Its aim is to approximate Member States' laws regarding certain offences involving racism and xenophobia. However, by prohibiting certain forms of expression and acts as "racist and xenophobic offences,"⁹ the framework decision also determines the Union's approach to *racist and xenophobic hate speech* in addition to hate crime more broadly.¹⁰

Adopted on the basis of the pre-Lisbon TEU provisions on police and judicial cooperation in criminal matters,¹¹ the framework decision defines hatred as "hatred based on race, color, religion, descent or national or ethnic origin."¹² This list of grounds is also used to define the offence of racist and xenophobic hate speech. Member States are required to criminalize public "incit[ement] to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, color, religion, descent or national or ethnic origin."¹³ Intentional conduct is required and the act can be committed by any means,¹⁴ offline and online. Member States are also required to punish the acts of publicly condoning, denying, or grossly trivializing certain crimes against humanity.¹⁵ Member States remain free to cover in their national legislation other

⁸Council Framework Decision 2008/913/JHA of Nov. 28, 2008, On combating certain forms and expressions of racism and xenophobia by means of criminal law, 2008 O.J. (L 328) 55. It was preceded by Joint Action 96/443/JHA of Jul. 15, 1996, Adopted by the Council on the Basis of Article K.3 of the Treaty on European Union, Concerning Action to Combat Racism and Xenophobia, 1996 O.J. (L 185) 5.

⁹Council Framework Decision 2008/913/JHA, art. 1, 2008 O.J. (L 328) 55.

¹⁰For more on hate crime particularly from the perspective of support victims, see Directive 2012/29, of the European Parliament and of the Council of Oct. 25, 2012, Establishing Minimum Standards on the Rights, Support, and Protection of Victims of Crime, and Replacing Council Framework Decision 2001/220/JHA, 2012 O.J. (L 315) 57. Article 22 of the directive requires Member States to identify victims' specific protection needs and determine whether and to what extent they would benefit from special measures foreseen in the directive in the course of criminal proceedings due to their particular vulnerability to secondary and repeat victimization, to intimidation and to retaliation. See Directive 2012/29 at art. 22. The individual assessment to perform shall take into account the personal characteristics of the victim, and the type, nature and circumstances of the crime, with particular attention given to victims who have suffered a crime committed with a bias or discriminatory motive which could, in particular, be related to their personal characteristics, victims of hate crime and victims with disabilities.

¹¹See Treaty on European Union, arts. 29, 31 and 34(2)(b), 2002 O.J. (C 325) 5 (consolidated version).

¹²Council Framework Decision 2008/913/JHA, recital 9, 2008 O.J. (L 328) 55.

¹³*Id.* at art. 1(1)(a).

¹⁴*Id.* at art. 1(1)(b).

¹⁵See Council Framework Decision 2008/913/JHA, art. 1(1)(c)-(d) (establishing a criminal prohibition of publicly condoning, denying or grossly trivializing crimes of genocide, crimes against humanity and war crimes as defined in Articles 6, 7, and 8 of the Statute of the International Criminal Court, as well as the crimes against peace defined in Article 6 of the

protected characteristics,¹⁶ besides race, color, etcetera, and they can also choose from certain optional qualifiers of punishable behavior.¹⁷ In addition, the framework decision sets forth rules on jurisdiction,¹⁸ provides that criminal penalties must be effective, proportionate, and dissuasive¹⁹ and enables *ex officio* investigations and prosecution²⁰—an attempt to address underreporting.

The framework decision does not set a clear benchmark for the definition of the offences at issue and affords Member States a significant degree of flexibility. This has resulted in varied understandings of hate speech as well as divergence and gaps in transposition.²¹ The Commission follows closely the implementation of the framework introduced.²² The EU High Level Group (EUHLG) on combating racism, xenophobia, and other forms of intolerance, established in 2016, also assists in implementation. It does so mostly by promoting best practices, cooperation among key stakeholders, practical guidance, and training on matters such as effective law enforcement, access to justice, victim support, and recording of hate speech and hate crime.

The Commission's European Democracy Action Plan, published in December 2020 to strengthen democratic resilience in the EU, announced further measures against hate speech. Acknowledging that digital hate speech can deter people from expressing their views and participating in online debates, the Commission envisaged an extension of the list of EU crimes under Article 83(1) of the Treaty on the Functioning of the European Union (TFEU) to cover hate speech, including online hate speech, and hate crime.²³ Article 83(1) TFEU lays down an exhaustive list of areas of "particularly serious crime with a cross-border dimension," such as terrorism, trafficking in human beings and sexual exploitation of women and children, computer crime, organized crime, etcetera, allowing the European Parliament and the Council to establish, through directives, minimum rules concerning the definition of criminal offences and sanctions applicable in all Member States. Article 83(1) TFEU adds that on the basis of "developments in crime" the Council, after obtaining the consent of the European Parliament, may unanimously adopt a decision identifying additional areas of such particularly serious crime, enabling the adoption of secondary legislation on common standards.

Charter of the International Military Tribunal appended to the London Agreement of 8 August 1945, directed against a group of persons or a member of such a group defined by reference to race, color, religion, descent or national or ethnic origin when the conduct is carried out in a manner likely to incite to violence or hatred against such a group or a member of such a group. Intentional conduct is required). See also *id.* art. 1(4) (providing that Member States enjoy the power to make the act of denying or grossly trivializing the above mentioned crimes punishable only if these have been established by a final decision of a national court and/or an international court or by a final decision of an international court only).

¹⁶For the majority of Member States, national legislation variably incorporates protected characteristics such as sexual orientation, gender identity, disability or social status. See the EU High Level Group on combating racism, xenophobia and other forms of intolerance Guidance Note on the practical application of Council Framework Decision 2008/913/JHA, On Combatting Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law (2018) 5.

¹⁷For example, they can only punish conduct that is carried out in a manner likely to disturb public order, or which is threatening, abusive or insulting. See Council Framework Decision 2008/913/JHA, art. 1(2), 2008 O.J. (L 328) 55.

¹⁸See Council Framework Decision 2008/913/JHA, 2008 O.J. (L 328) 55 (establishing the following criteria conferring jurisdiction to prosecute and investigate cases: The territoriality criterion, in other words the offence is committed in whole or in part in a Member State, and the nationality criterion, in other words the offence is committed by a national of a Member State). See also *id.* at art. 9 (providing with regard to online hate speech that when a Member State establishes jurisdiction by means of the territoriality criterion, it shall ensure that its jurisdiction extends to cases where the conduct is committed through an information system and the offender or the material hosted in that information system is in its territory).

¹⁹*Id.* at art. 3.

²⁰*Id.* at art. 8.

²¹Report from the Commission to the European Parliament and the Council on the implementation of Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law, COM (2014) 027 final (Jan., 1 2014).

²²See e.g. Commission Staff Working Document, *Countering racism and xenophobia in the EU: Fostering a society where pluralism, tolerance and non-discrimination prevail*, SWD (2019) 110 final (Mar. 15, 2019).

²³See 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*, at § 2.4, COM (2020) 790 final (Dec. 3, 2020), and Consolidated Version of the Treaty on the Functioning of the European Union art. 83(1), October 26, 2012, 2012 O.J. (C326) 47 [hereinafter TFEU].

In December 2021, following an external support study that mapped Member States' laws against hate speech and hate crime,²⁴ the Commission presented an initiative based on Article 17(1) TEU²⁵ for a Council decision extending the areas of crime covered by Article 83(1) TFEU to include hate speech and hate crime.²⁶ The initiative, the Commission argued, should be seen as an effective means of comprehensively addressing the challenges posed by hate speech and hate crime, going beyond the protected grounds covered by the framework decision. However, careful attention should be paid to the fundamental rights repercussions of any action taken, with due respect for the principle of proportionality and the essence of free speech.²⁷ Given Member States' differences in and fragmented approaches to the criminalization of hate speech and hate crime thus far, the unanimity requirement in the Council may prove an unsurmountable hurdle.

In March 2022, the Commission proposed new legislation on combatting violence against women and domestic violence on the basis of Articles 82(2)²⁸ and 83(1) TFEU, which also addresses cyber incitement to violence or hatred.²⁹ The Commission observed that “the increase in internet and social media usage has led to a sharp rise in public incitement to violence and hatred, including based on sex or gender.”³⁰ It also noted that “the easy, fast and broad sharing of hate speech through the digital word is reinforced by the online disinhibition effect,” given that “the presumed anonymity on the internet and sense of impunity reduce people’s inhibition to engage in such speech.”³¹ The proposal criminalizes cyber incitement to violence or hatred based on sex or gender, namely “the intentional conduct of inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to sex or gender, by disseminating to the public material containing such incitement by means of information and communication technologies.”³² According to the legal basis of the proposal, the legislative instrument put forward is a directive adopted in accordance with the ordinary legislative procedure, which is expected to facilitate agreement in the Council and European Parliament.³³

II. Banning Hate Speech in Audiovisual Media Services and on Video-Sharing Platforms

As part of the television broadcasting policy of the then European Economic Community, the 1989 *Television Without Frontiers Directive* (TWFD) required Member States to ensure that broadcasts by operators under their jurisdiction do not contain “any incitement to hatred on

²⁴See Patricia Ypma, Célia Drevon, Chloe Fulcher, Oriana Gascon, Kevin Brown, Aleksandar Marsavelski, & Sylvie Giraudon, *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* (Eur. Comm’n, Final Report, 2021), <https://op.europa.eu/en/publication-detail/-/publication/f866de4e-57de-11ec-91ac-01aa75ed71a1/language-en>.

²⁵Consolidated Version of the Treaty on European Union art. 17(1), October 26, 2012, 2012 O.J. (C 326) 13 [hereinafter TEU] (providing inter alia that “[t]he Commission shall promote the general interest of the Union and take appropriate initiatives to that end.”).

²⁶See *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 (Dec. 9, 2021).

²⁷See *id.* at annex, recital 10 of the Council decision accompanying the communication.

²⁸See TFEU art. 82(2) (allowing for the adoption of minimum rules on the rights of victims of crime to the extent necessary to facilitate mutual recognition of judgments and judicial decisions, and police and judicial cooperation in criminal matters with a cross-border dimension).

²⁹See *Proposal for a Directive of the European Parliament and of the Council on combating violence against women and domestic violence* COM (2022) 105 final (Mar. 8, 2022).

³⁰*Id.* at recital 22.

³¹*Id.*

³²*Id.* at art. 10.

³³The Council was reported to have agreed on its position on the proposed directive on June 9, 2023. The Committee on Women’s Rights and Gender Equality and the Committee on Civil Liberties, Justice and Home Affairs of the European Parliament adopted their report on the Commission’s proposal on June 28, 2023. Interinstitutional negotiations started shortly afterwards. For more on the legislative process, see <https://www.europarl.europa.eu/legislative-train/theme-a-new-push-for-european-democracy/file-legislative-proposal-on-gender-based-violence>.

grounds of race, sex, religion or nationality.”³⁴ The TWFD indicated sex as a ground for protection against hate speech but did not address color, descent, or ethnic origin, as Framework Decision 2008/913/JHA subsequently would. The rule has been retained in all subsequent amendments of the directive, leading through to the *Audiovisual Media Services Directive* (AVMSD)³⁵ and rendered applicable to all audiovisual media services coming within the purview of the AVMSD: Traditional broadcasting, in other words linear audiovisual media services, and “television-like” services, in other words on-demand audiovisual media services, also known as non-linear audiovisual media services,³⁶ like Netflix or Hulu. In *Mesopotamia Broadcast and Roj TV*, the CJEU coined the concept of “incitement to hatred” as referring to “an action intended to direct specific behavior and [. . .] a feeling of animosity or rejection with regard to a group of persons.”³⁷

The revised AVMSD, adopted in 2018, modified its hate speech provision and also extended the list of protected grounds.³⁸ Article 6(1)(a) of the AVMSD now mandates Member States to ensure “by appropriate means” that audiovisual media services provided by media service providers under their jurisdiction do not contain “any incitement to violence or hatred against a group of persons or a member of a group” based on any of the grounds referred to in Article 21 of the Charter, namely sex, race, color, 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.³⁹ The fight against hate speech in audiovisual media services is firmly embedded within a fundamental rights context: State measures must be necessary and proportionate, respect the rights and observe the principles enshrined in the CFR.⁴⁰

With the latest revision of the AVMSD, the scope of the directive has also been extended to cover “video-sharing platforms” (VSPs),⁴¹ with specific rules introduced in their regard. Particularly as regards hate speech, the AVMSD requires Member States to ensure that VSPs take appropriate measures not only against the dissemination of programs, user-generated videos, and

³⁴Council Directive 89/552/EEC of 3 October 1989, On the Coordination of Certain Provisions Laid Down by Law, Regulation, or Administrative Action in Member States Concerning the Pursuit of Television Broadcasting Activities, 1989 O.J. (L 298) 23.

³⁵See Article 3b of Directive 89/552/EEC, as amended by Directive 2007/65/EC of the European Parliament and of the Council of 11 December 2007, 2007 O.J. (L 332) 17. See also Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010, art. 6, 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), 2010 O.J. (L 95) 1 (codified version) (“Member States shall ensure by appropriate means that audiovisual media services provided by media service providers under their jurisdiction do not contain any incitement to hatred based on race, sex, religion or nationality.”).

³⁶See Directive 2010/13/EU, art. 1(1)(g) 2010 O.J. (L 95) 1 (defining non-linear audiovisual media services as “an audiovisual media service provided by a media service provider for the viewing of programs at the moment chosen by the user and at his individual request on the basis of a catalogue of programs selected by the media service provider.”). For more details, see Peggy Valcke and Ingrid Lambrecht, *The Evolving Scope of Application of the AVMS Directive*, in RESEARCH HANDBOOK ON EU MEDIA LAW AND POLICY 282 (Pier Luigi Parcu & Elda Brogi eds, 2021).

³⁷Joined Cases C-244/10 and C-245/10 *Mesopotamia Broadcast A/S METV and Roj TV A/S v. Bundesrepublik Deutschland*, ECLI:EU:C:2011:607, ¶ 41 (May 5, 2011).

³⁸Directive (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, 2018 O.J. (L 303) 69.

³⁹See Directive 2010/13/EU of the European Parliament and of the Council of March 10, 2010, Consolidated Text of the Audiovisual Media Services Directive (AVMSD), 2010 O.J. (L 303) 69, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02010L0013-20181218>.

⁴⁰*Id.* at art. 6(2).

⁴¹See *id.* art. 1(a)(aa) (defining VSP as a commercial service addressed to the public of which the principal purpose, a dissociable section or an essential functionality is devoted to the provision of programs and/or user-generated videos for which the VSP provider has no editorial responsibility to the general public, in order to inform, entertain or educate; which is made available by electronic communication networks; and whose organization is determined by the VSP provider, including by automatic means or algorithms, in particular by displaying, tagging and sequencing).

audiovisual commercial communications that infringe Framework Decision 2008/913/JHA, but also against content containing “incitement to violence or hatred against a group of persons or a member of a group” based on any of the grounds referred to in Article 21 CFR.⁴² Relevant measures must be “practicable and proportionate,” taking into account the size and nature of the VSP service concerned.⁴³ They may range from safeguards in VSPs’ terms and conditions to transparent and user-friendly content reporting and flagging mechanisms, accountability tools, effective complaint-handling and resolution procedures, and media literacy measures.⁴⁴ National regulatory authorities are entrusted with the task of assessing the measures taken,⁴⁵ and provision has to be made for judicial review⁴⁶ and out-of-court redress mechanisms for the settlement of disputes between users and VSPs.⁴⁷ The use of co-regulation is particularly promoted,⁴⁸ with the Commission assuming a key role in helping VSPs share best practice on the basis of co-regulatory codes of conduct.⁴⁹ Self-regulatory codes, broadly accepted by the main stakeholders, are also encouraged at Union level.⁵⁰ The provisions make clear that fighting hate speech on VSPs is a shared responsibility and implicates a broad range of actors, including the VSPs themselves. However, according to a 2021 study on the implementation of the revised AVMSD, VSPs have not been particularly consistent in their approach to controls on hate speech, mostly due to the fact that “definitions and guidance for users vary widely.”⁵¹

Whilst seeking to keep EU audiovisual media law up to date with technological developments, the directive’s principal rule remains freedom of reception: Member States are not allowed to restrict retransmissions of audiovisual media services from other Member States on their territory.⁵² Audiovisual media services are regulated in the country of origin, in other words the Member State from which they emanate and not the country of destination. However, Article 3(2) of the AVMSD allows Member States to provisionally derogate from the principle of the country of origin, subject to substantive and procedural conditions, enabling *inter alia* restriction of the cross-border transmission of an audiovisual media service where this “manifestly, seriously and gravely” infringes the requirements set forth in Article 6(1)(a) in relation to hate speech.⁵³ The derogation must be notified to the media service provider concerned, the Member State of origin and the Commission. The CJEU has also interpreted the AVMSD in ways that provide Member States with significant room for maneuver to restrict audiovisual media services on grounds of public order,⁵⁴ a concept that can accommodate hate speech concerns.

⁴²*Id.* at art. 28b(1).

⁴³*Id.* at art. 28b(3).

⁴⁴*Id.* at art. 28b(6) (providing that Member States are free to impose stricter or more detailed measures).

⁴⁵*Id.* at art. 28b(5).

⁴⁶*Id.* at art. 28b(8).

⁴⁷*Id.* at art. 28b(7).

⁴⁸*Id.* at art. 28b(4).

⁴⁹*Id.* at art. 28b(9).

⁵⁰*Id.* at arts. 28b(10) and 4a(2).

⁵¹Deloitte & SMIT, *Study on the Implementation of the New Provisions in the Revised Audiovisual Media Services Directive (AVMSD)*, 8 (Eur. Comm’n, Directorate-General for Communications Networks, Content, and Technology, SMART 2018/0066–Part D, 2021), <https://op.europa.eu/en/publication-detail/-/publication/6d536c6f-5c68-11eb-b487-01aa75ed71a1/language-en>.

⁵²Directive 2010/13/EU of the European Parliament and of the Council of March 10, 2010, art. 3(1), Consolidated Text of the Audiovisual Media Services Directive (AVMSD), <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02010L0013-20181218>.

⁵³*Id.* at art. 3(2).

⁵⁴See joined Cases C-244/10 and C-245/10 *Mesopotamia Broadcast A/S METV and Roj TV A/S v. Bundesrepublik Deutschland*, ECLI:EU:C:2011:607, ¶ 41 (May 5, 2011). See also Case C-622/17, *Baltic Media Alliance Ltd. v. Lietuvos radijo ir televizijos komisija*, ECLI:EU:C:2019:566 (Aug. 23, 2019).

III. Tackling ‘Illegal’ Content Online

Arrangements similar to those of the AVMSD have been made with Directive 2000/31/EC, also known as the e-Commerce Directive,⁵⁵ whose aim is to contribute to the proper functioning of the internal market by ensuring the free movement of information society services between the Member States.⁵⁶ Although Member States may not, for reasons falling within the coordinated field of the directive, restrict the freedom to provide information society services from another Member State,⁵⁷ derogations are allowed under certain conditions in respect of “a given information society service,”⁵⁸ provided that both the Commission and the Member State where the provider of the service in question is established have been notified. These include measures necessary for reasons of “public policy, in particular the prevention, investigation, detection and prosecution of criminal offences, including [...] the fight against any incitement to hatred on grounds of race, sex, religion or nationality.”⁵⁹ Interestingly, the e-Commerce Directive retains the hate speech wording of the original TWFD.

Until the enactment of the Digital Services Act,⁶⁰ Directive 2000/31/EC was the principal framework at EU level on issues pertaining to the liability of digital intermediaries and therefore a key point of reference for action taken to remove and disable access to illegal content online, including hate speech. The basic rule set forth in the directive was that digital intermediaries are exempted from liability, so long as they transmit or store information in a merely “technical, automatic and passive” manner, which implies that they have “neither knowledge of nor control over the information which is transmitted or stored,” and provided that they take expeditious action on infringements after obtaining knowledge or becoming aware of them.⁶¹ Article 15(1) of the e-Commerce Directive further precluded Member States from imposing a general obligation on digital intermediaries to monitor the information they transmit or store and/or to actively seek facts or circumstances that may indicate illegal activity.⁶² There should accordingly be no general duty imposed on intermediaries to monitor content and actively seek instances of infringement.

The Commission’s 2017 Communication, *Tackling Illegal Content Online: Towards an Enhanced Responsibility of Online Platforms*, signaled the EU institutions’ wish to revisit long-standing understandings of the obligations of digital intermediaries in relation to illegal content online.⁶³

⁵⁵Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000, On Certain Legal Aspects of Information Society Services, Particularly Electronic Commerce, in the Internal Market (‘Directive on Electronic Commerce’) 2000 O.J. (L 178) 1.

⁵⁶*Id.* at art. 1(1).

⁵⁷*Id.* at art. 3(2).

⁵⁸*Id.* at art. 3(4). *See also* CJEU, Case C-376/22, *Google Ireland and Others*, ECLI:EU:C:2023:835, ¶ 27 (Nov. 9, 2023) (clarifying that the information society service referred to must be understood as “an individualized service provided by one or more service providers,” which entails that Member States cannot adopt “general and abstract measures aimed at a category of given information society services described in general terms and applying without distinction to any provider of that category of services.”).

⁵⁹Directive 2000/31/EC of the European Parliament and of the Council of June 8, 2000, art. 3(4), 2000 O.J. (L 178) 1.

⁶⁰Regulation (EU) 2022/2065 of the European Parliament and of the Council of October 19, 2022, On a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act), 2022 O.J. (L 277) 1 [hereinafter *Digital Services Act*].

⁶¹Directive 2000/31/EC, recital 42 and arts. 13 and 14, 2000 O.J. (L 178) 1. *See also* Case C-324/09, *L’Oréal and Others*, ECLI:EU:C:2011:474, ¶ 113, 116 (Jul. 12, 2011) (holding that operators’ liability is restricted to cases where the intermediary “plays an active role of such a kind as to give it knowledge of, or control” over the hosted content).

⁶²*See also* Directive 2010/13/EU of the European Parliament and of the Council of March 10, 2010, art. 28b(3), Consolidated Text of the Audiovisual Media Services Directive (AVMSD), <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02010L0013-20181218> (noting that the measures VSPs are required to take, pursuant to Article 28b of the AVMSD, against hate speech should not lead to *ex ante* control or upload-filtering of content, in violation of the prohibition laid down in Article 15 of the e-Commerce Directive).

⁶³*Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Tackling illegal content online: Towards an enhanced responsibility of online platforms*, COM (2017) 555 final (Sept. 28, 2017).

Stressing the “significant societal responsibility” of online platforms, which mirrored arguments revolving around the public functions of online intermediaries as enablers of speech and gatekeepers of information,⁶⁴ the Commission called upon digital intermediaries to “decisively step up their actions” aimed at detecting and removing illegal content quickly and efficiently, including by means of “voluntary, proactive measures.”⁶⁵ The Commission emphasized that proactive measures should not automatically entail losing the benefit of the “safe harbor” provisions of the e-Commerce Directive⁶⁶ and underlined the need for bolstering cooperation and investment in, and use of, automated systems,⁶⁷ acknowledging that proactive measures can rest on automation.⁶⁸

Commission Recommendation 2018/334 on measures to effectively tackle illegal content online sought to make progress in the field,⁶⁹ despite its non-binding nature. Tackling illegal content online would yet prove a key component of the unprecedented reforms introduced by the EU by means of the Digital Services Package. In particular, the Digital Services Act (DSA), which supplements the e-Commerce Directive, aspires to revolutionize the provision of online intermediary services and platform oversight in the Union, reflecting the Union’s resolve to engage in genuine platform regulation.⁷⁰ The DSA applies to providers who offer intermediary services in the Union, irrespective of their place of establishment.⁷¹ It lays down horizontal due diligence rules whose regulatory intensity is graduated, depending on the type of intermediary, its size and impact on society. It targets “providers of intermediary services,”⁷² the broadest category of operators who fall within the scope of the rules enacted, which also apply to “providers of hosting services,”⁷³ “online platforms,”⁷⁴ “very large online platforms” (VLOPs), and “very large

⁶⁴See Teresa Quintel & Carsten Ullrich, *Self-Regulation of Fundamental Rights? The EU Code of Conduct on Hate Speech, Related Initiatives and Beyond*, in FUNDAMENTAL RIGHTS PROTECTION ONLINE: THE FUTURE REGULATION OF INTERMEDIARIES 182 (Bilyana Petkova & Tuomas Ojanen eds, 2020). See also LORNA WOODS & WILLIAM PERRIN, ONLINE HARM REDUCTION—A STATUTORY DUTY OF CARE AND REGULATOR 5, 21, 28 (2019), <https://carnegieuktrust.org.uk/publications/online-harm-reduction-a-statutory-duty-of-care-and-regulator/> (arguing that an appropriate analogy for online platforms is that of a public space).

⁶⁵*Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Tackling illegal content online: Towards an enhanced responsibility of online platforms*, at 2 and 10 COM (2017) 555 final (Sept. 28, 2017).

⁶⁶*Id.* at 10. See also Case C-324/09 *L’Oreal and Others* (2011) (holding that a provider shall be “denied entitlement to the exemption from liability provided for in Article 14 of Directive 2000/31” when they are “aware of facts or circumstances on the basis of which a diligent economic operator should have identified the illegality in question and acted in accordance with Article 14(1)(b) of Directive 2000/31.”). The Court of Justice of the European Union explained that this exemption “cover[s] every situation in which the provider concerned becomes aware, in one way or another, of such facts or circumstances,” including situations in which an illegal activity or illegal information is uncovered as a result of an investigation undertaken on the provider’s own initiative. *Id.* at ¶ 120–122; Joris van Hoboken, João Pedro Quintais, Joost Poort, & Nico van Eijk, *Hosting Intermediary Services and Illegal Content Online: An Analysis of the Scope of Article 14 ECD in light of Developments in the Online Service Landscape*, at 42 (Study for the Eur. Comm’n, Final Report, 2018), <https://op.europa.eu/en/publication-detail/-/publication/7779caca-2537-11e9-8d04-01aa75ed71a1/language-en> (explaining that a proactive stance, in other words an investigation conducted by the intermediary itself, was thus considered to increase the chance of the provider acquiring knowledge of illegality and thus being exposed to liability).

⁶⁷*Id.* at 13.

⁶⁸*Id.* at 12–13.

⁶⁹See Commission Recommendation (EU) 2018/334 of March 1, 2018, On Measures to Effectively Tackle Illegal Content Online, 2018 O.J. (L 63) 50.

⁷⁰Miriam C. Buiten, *The Digital Services Act: From Intermediary Liability to Platform Regulation*, 12 JIPITEC 361 (2021).

⁷¹Digital Services Act, art. 2(1).

⁷²*Id.* at art. 3(g) (defining all providers as offering mere conduit, caching, and hosting services).

⁷³*Id.* at art. 3(g)(iii) (defining the providers as those whose services consist of storage of information provided by, and at the request of, a recipient of the service).

⁷⁴*Id.* at art. 3(i) (defining the providers as the providers of a hosting service who at the request of a recipient of the service, store and disseminate information to the public, unless that activity is a minor and purely ancillary feature of another service or a minor functionality of the principal service which, for objective and technical reasons, cannot be used without that other service, where the integration of the feature or functionality into the other service is not a means to circumvent the applicability of the DSA).

online search engines” (VLOSEs).⁷⁵ The DSA retains the e-Commerce Directive’s rules on liability exemptions for intermediaries and the general prohibition on monitoring, and acknowledges that voluntary own-initiative investigations, measures aimed at detecting, identifying, and removing or disabling access to illegal content, and any measures taken to comply with EU law requirements do not entail the loss of the liability protections.⁷⁶

A set of substantive rules are then laid down to tackle illegal content,⁷⁷ which is broadly defined as “any information that in itself or in relation to an activity [. . .] 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.”⁷⁸ This includes illegal hate speech.⁷⁹ The DSA provides rules for framing transparency and due diligence obligations concerning operators’ content moderation policies and practice,⁸⁰ highlighting that relevant obligations should aim in particular to guarantee different public policy objectives such as the safety and trust of users, including users at particular risk of being subject to hate speech.⁸¹ More stringent transparency duties are imposed on online platforms, VLOPs and VLOSEs.⁸² Providers of hosting services are also mandated to put in place notice-and-action mechanisms facilitating the submission of “sufficiently precise and adequately substantiated notices”⁸³ of the presence of illegal content, and to justify any restrictions imposed, from undermining the visibility of content deemed to be illegal, removing it, disabling access to it or demoting it, to suspending or terminating the provision of the service or the user’s account, among other issues.⁸⁴ Obligations become more stringent for online platforms, VLOPs and VLOSEs. Online platforms need to take the necessary technical and organizational measures to prioritize, process and decide upon notices submitted by *trusted flaggers* without delay.⁸⁵ They also need to provide for internal complaint-handling systems,⁸⁶ with users also being allowed to resort to certified out-of-court dispute procedures to seek redress,⁸⁷ and introduce measures to protect against the misuse of their services.⁸⁸ VLOPs and VLOSEs are additionally required to, at least once a year, diligently identify, analyze and assess any “systemic risks” stemming from the design or functioning of their service, including their algorithmic systems, or from the use made of their services.⁸⁹ Systemic risks may involve the dissemination of illegal content, as well as any negative effects—actual or foreseeable—on *inter alia* the exercise of fundamental rights, civic discourse, public security, the protection of minors, and individual physical and mental well-being.⁹⁰ Relevant provisions create ample room for hate speech to come under the DSA rubric of “systemic risk.”⁹¹ Further, VLOPs and VLOSEs are mandated to consider

⁷⁵*Id.* at art. 33(1) (defining the online platforms and online search engines as those with at least 45 million monthly active users within the Union or designated as very large online platforms or very large online search engines by the Commission).

⁷⁶*Id.* at arts. 4–6, 7 and 8. For more on the interplay between these provisions, see generally Joan Barata, *The Digital Services Act and Social Media Power to Regulate Speech: Obligations, Liabilities and Safeguards*, in UNRAVELLING THE DIGITAL SERVICES ACT PACKAGE 7 (Maja Cappello, ed., IRIS Special, European Audiovisual Observatory, 2021).

⁷⁷See e.g. Alexandre de Stree & Michèle Ledger, *Regulating the Moderation of Illegal Online Content*, in UNRAVELLING THE DIGITAL SERVICES ACT PACKAGE 23 (Maja Cappello, ed., IRIS Special, European Audiovisual Observatory, 2021).

⁷⁸Digital Services Act, art. 3(h).

⁷⁹*Id.* at recital 12.

⁸⁰*Id.* at arts. 14 and 15.

⁸¹*Id.* at recital 40.

⁸²*Id.* at arts. 24 and 42.

⁸³*Id.* at art. 16.

⁸⁴*Id.* at art. 17.

⁸⁵*Id.* at art. 22.

⁸⁶*Id.* at art. 20.

⁸⁷*Id.* at art. 21.

⁸⁸*Id.* at art. 23.

⁸⁹*Id.* at art. 34.

⁹⁰*Id.*

⁹¹*Id.* at recital 40.

the ways in which their content moderation systems influence systemic risks⁹² and to put in place reasonable, proportionate, and effective mitigation measures tailored to the specific systemic risks identified.⁹³ Mitigating measures may involve actions such as adapting and applying terms and conditions as necessary or adjusting content moderation systems and/or relevant decision-making processes and resources, including the content moderation staff, their training and expertise with regard in particular to the speed and quality of the notice processing they carry out.⁹⁴ In this regard, the DSA, which recognizes the adoption of voluntary codes of conduct as a means to implement its provisions,⁹⁵ identifies risk mitigation measures against illegal content as an area that should receive consideration through self- and co-regulatory instruments,⁹⁶ and makes express reference to the *Code of Conduct on Countering Illegal Hate Speech Online*.⁹⁷

Signed in May 2016 by major digital intermediaries,⁹⁸ the *Code of Conduct on Countering Illegal Hate Speech Online* is the result of a process facilitated by the Commission in accordance with Article 16 of the e-Commerce Directive, which encourages the drawing up of codes of conduct at Union level as a contribution to the implementation of the directive.⁹⁹ The Code is based on the premise that proper enforcement of Framework Decision 2008/913/JHA must be complemented by action taken by digital intermediaries to ensure that online hate speech is dealt with expeditiously upon receipt of a valid, in other words precise and properly substantiated, notification. The Code requires operators to put in place their own rules and community standards prohibiting hate speech, as well as clear and effective procedures for reviewing notifications on the basis of such rules and community standards and “where necessary” Member States’ laws transposing Framework Decision 2008/913/JHA into their legal orders. Parties to the Code are committed to reviewing the majority of flagged content in less than 24 hours and to remove or disable access to it, if required. The Code also contains provisions on sharing information with Member States on notifications received and how they were dealt with; engaging in partnerships and training activities with civil society to establish a network of *trusted flaggers* of hate speech; providing regular staff with training; sharing best practices; and supporting independent counter-narratives and educational programs which foster positive narratives. Compliance with the Code is to be regularly reviewed and assessment must take place through a structured process of periodic monitoring involving a host of civil society organizations across the Union, which act as *trusted flaggers* along with self-reporting by the Code’s signatories to the Commission. Findings from the 7th monitoring round in 2022¹⁰⁰ show that while the average number of notifications reviewed within 24 hours has fallen from 81% in 2021 to 64.4% in 2022, the average removal rate of 63.6% was similar to the 2021 rate of 62.5%, though lower than the rate of 71% in 2020.

The Code has been praised for fostering mutual learning and synergies between digital intermediaries, civil society, and Member States’ authorities,¹⁰¹ but the truth is that concrete

⁹²*Id.*

⁹³*Id.* at art. 35.

⁹⁴*Id.* at recital 87.

⁹⁵*Id.* at art. 45(1).

⁹⁶*Id.* at recital 104.

⁹⁷*Id.*

⁹⁸The Code was originally signed by Facebook, Microsoft, Twitter and YouTube. It was subsequently joined by other operators. See EUR. COMM’N, CODE OF CONDUCT ON COUNTERING ILLEGAL HATE SPEECH ONLINE (2019), https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combating-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en. See generally Quintel & Ullrich, *supra* note 64).

⁹⁹Council 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 O.J. (L 328) 55.

¹⁰⁰See Eur. Comm’n, *Countering illegal hate speech online: 7th evaluation of the Code of Conduct* (Nov. 2022), <https://commission.europa.eu/system/files/2022-12/Factsheet%20-%207th%20monitoring%20round%20of%20the%20Code%20of%20Conduct.pdf>.

¹⁰¹Commission Staff Working Document, *Countering racism and xenophobia in the EU: Fostering a society where pluralism, tolerance and non-discrimination prevail*, at 6, SWD (2019) 110 final (Mar. 15, 2019).

standards are only set with regard to the speed with which flagged content should be addressed. Clearly, there is an overemphasis on ensuring compliance with digital intermediaries' own rules and standards, though evaluation, as the Commission has revealed, rests on flagged content in the light of domestic laws transposing Framework Decision 2008/913/JHA.¹⁰² In addition, no meticulous definition of what constitutes a precise and properly substantiated notification is given, and there are no procedural safeguards in place guaranteeing the provision of systematic feedback to users. According to data from the 7th monitoring round, operators provided feedback to notifiers in 66.4% of cases,¹⁰³ compared to 60.3% in 2021, but no data has been disclosed concerning feedback to users whose content has been removed and the provision of information on the remedies available.¹⁰⁴ Notably, the monitoring process does not include an evaluation of the intermediaries' decision-making process in relation to what content is removed and what content is not.

C. Definitional and Automation Hindrances

Variation in regulatory instruments and techniques is a key feature of the EU's action against hate speech. A hard law, criminal law-based approach to racial and xenophobic hate speech through Council Framework Decision 2008/913/JHA has been combined with a special liability regime for digital intermediaries, originally set forth in the e-Commerce Directive, and regulation specifically targeting audiovisual media service providers and video-sharing platforms through the AVMSD. There are also soft law measures, such as the non-binding Commission Recommendation 2018/334 on measures to effectively tackle illegal content online, and the Code of Conduct on Countering Illegal Hate Speech, which, as it was made possible by the Commission which is also involved in the monitoring of its implementation, verges on transnational co-regulation. With the adoption of the DSA, EU regulatory efforts have stiffened substantially, resulting in the imposition of asymmetric due diligence obligations on digital intermediaries. The DSA also underlines the importance of the Code and the time benchmark it sets for the processing of valid notifications and the removal of hate speech with regard to the measures that VLOPs and VLOSEs must take to mitigate systemic risks.

Variation in EU regulatory approaches vis-à-vis hate speech goes hand in hand with varied understandings of hate speech. Whilst the e-Commerce Directive did not define "illegal information or activity" when laying out its liability protections, the DSA does not provide a substantive definition of illegal content, but rather cross-refers to EU law and Member States' laws to capture illegality. At the same time, regulatory instruments such as Council Framework Decision 2008/913/JHA and the AVMSD advance different definitions of hate speech and build flexibilities into EU law that support and intensify differences in national legislation. Indeed, research has confirmed that Member States have diverging rules on matters pertaining to hate speech: Some Member States define hate speech on the basis of a limited number of protected attributes, while others have more extensive lists of protected characteristics, and while the lists of protected grounds may be exhaustive in some Member States, they are open-ended in others.¹⁰⁵ It should therefore come as no surprise that in 2017, in a motion for a resolution, the European

¹⁰²See Eur. Comm'n, *Countering illegal hate speech online: 7th evaluation of the Code of Conduct*, *supra* note 100, at Annex (Methodology of the exercise), (Nov. 2022), <https://commission.europa.eu/system/files/2022-12/Factsheet%20-%207th%20monitoring%20round%20of%20the%20Code%20of%20Conduct.pdf>.

¹⁰³Facebook informs notifiers systematically in 84.9% of cases.

¹⁰⁴For context on the opaqueness of content moderation, see generally TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA (2018); Nicolas P. Suzor, Sarah Myers West, Andrew Quodling, & Jillian York, *What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation*, 13 INT'L J. COMM'C'N 1526 (2019).

¹⁰⁵*Bayer & Bárd*, *supra* note 1.

Parliament called on the Commission to explore the feasibility of establishing a common legal definition of hate speech in the EU.¹⁰⁶

This lack of a common legal definition means that digital intermediaries will often be required to comply with different legal approaches to hate speech which derive respectively from EU law and national rules. At the same time, they have their own say on permissible and impermissible expression on their services. Research suggests the absence of a common approach here, too.¹⁰⁷ Key players have developed detailed policies on what they consider hate speech, and they have also expanded the list of protected grounds beyond those identified in instruments such as Council Framework Decision 2008/913/JHA. Thus, characteristics such as veteran status, immigration status, socio-economic status, caste, age, or weight stand alongside protected attributes, including race, ethnicity, religion/religious affiliation, national origin, sex/gender, sexual orientation, disability, or disease, which generally feature prominently in operators' policies. Other operators have, as of yet, refrained from identifying specific protected characteristics.

The fact that digital intermediaries may be considering a wider range of content as hate speech, either because they define hate speech through more compendious lists of protected characteristics or because they proscribe hate speech in general, which can imply a wide interpretation of such speech, is problematic. That content which is not illegal as per EU law and/or Member States' rules can still be outlawed through private enforcement has important implications for freedom of expression and freedom of information. The fact that there may be a thin line between free speech and hate speech caught by the algorithm is also a source for concern. In the context of steps taken to improve the detection of hate speech—in light of the short time windows imposed for content takedowns by instruments such as the Code of Conduct on Countering Illegal Hate Speech Online—digital intermediaries are increasingly resorting to technology and automation.¹⁰⁸ According to data from the latest monitoring round of the implementation of the Code, between April and June 2022, Facebook took action on 13.5 million items of hate speech, of which 95.6% was proactively detected.¹⁰⁹ Facebook claims to have “pioneered the use of artificial intelligence technology to remove hateful content proactively, before users report it. . .”¹¹⁰ Instagram discloses similarly high levels of proactive detention at 91.2%.¹¹¹

Automated detection mechanisms may now be common, but they tend to be context blind.¹¹² As hate speech detection is most often a contextual exercise, automated tools may lead to

¹⁰⁶EUR. PARL. DOC., *Motion for a Resolution pursuant to Rule 133 of the Rules of Procedure on establishing a common legal definition of hate speech in the EU* (Feb. 17, 2017), https://www.europarl.europa.eu/doceo/document/B-8-2017-0172_EN.html.

¹⁰⁷See Natalie Alkiviadou, *Hate Speech on Social Media Networks: Towards A Regulatory Framework?* 28 INFO. & COMM'CNS TECH. L. 19 (2019). See also Deloitte & SMIT, *supra* note 51; Federica Casarosa, *The European Regulatory Approach toward Hate Speech Online: The Balance between Efficient and Effective Protection*, 55 GONZ. L. REV. 391 (2020).

¹⁰⁸For more on how automated content moderation works, see generally Robert Gorwa, Reuben Binns, & Christian Katzenbach, *Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance*, 7 BIG DATA & SOC'Y (2020).

¹⁰⁹EUR. COMM'N, *Information provided by the IT companies about measures taken to counter hate speech, including their actions to automatically detect content* (November 2022), <https://commission.europa.eu/system/files/2022-12/Information%20provided%20by%20the%20IT%20companies%20about%20measures%20taken%20to%20counter%20hate%20speech%20E2%80%93%202022.pdf>.

¹¹⁰*Id.*

¹¹¹*Id.*

¹¹²See e.g. Emma Llansó, Joris van Hoboken, Paddy Leerssen & Jaron Harambam, *Artificial Intelligence, Content Moderation, and Freedom of Expression* (Transatlantic Working Group, 2020), <https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf>. See also Jennifer Cobbe, *Algorithmic Censorship by Social Platforms: Power and Resistance*, 34(3) PHIL. & TECH. 739 (2020); Tarleton Gillespie, *Content Moderation, AI, and the Question of Scale*, 7 BIG DATA & SOC'Y (2020); Gorwa et al., *supra* note 108; COMMITTEE OF EXPERTS ON INTERNET INTERMEDIARIES, ALGORITHMS AND HUMAN RIGHTS, *STUDY ON THE HUMAN RIGHTS DIMENSIONS OF AUTOMATED DATA PROCESSING TECHNIQUES AND POSSIBLE REGULATORY IMPLICATIONS* (2017), <https://edoc.coe.int/en/internet/7589-algorithms-and-human-rights-study-on-the-human-rights-dimensions-of-automated-data-processing-techniques-and-possible-regulatory-implications.html>; Kate Klonick,

mislabeling, over-detection, or no detection. Reliance on a combination of machines and human moderators is essential to assuage concerns over both over-inclusive and under-inclusive approaches to hate speech. The DSA has taken some steps towards improving transparency in the use of automation by digital intermediaries.¹¹³ It also requires providers of hosting services to state when automated means are employed in the processing of notices of illegal content on their service and in their relevant decision-making,¹¹⁴ and to include “information on the use made of automated means” in their “statement of reasons” regarding any restrictive measures taken against illegal content, including information on whether decisions were “taken in respect of content detected or identified using automated means”.¹¹⁵ Finally, it precludes online platforms from relying solely on automation when handling complaints.¹¹⁶

D. Conclusion

Hate speech undermines the very foundations of a democratic and pluralistic society and the common values enshrined in Article 2 TEU. It is not unique to the online world, but digitalization has brought with it unprecedented challenges with regard to its volume, dissemination and reach. In recognition of the societal impact hate speech has in terms of eroding social cohesion, solidarity and trust between members of a society, the EU, as a *Union of values*,¹¹⁷ has strengthened its efforts to combat digital hatred in recent years. The EU regulatory toolbox against hate speech presently includes instruments with varying degrees of regulatory breadth and intensity, while the EU is also a staunch supporter of voluntary codes of conduct by the industry and co-regulation. However, the mechanisms that ensue lack a common definition of hate speech. Definitions vary at both the EU and national level, and digital intermediaries are advancing their own policies and standards on what is treated as hate speech on their services. This entails a complex multi-level approach to hate speech and the interplay of different regimes in terms of the rules set forth, their nature and scope, and models of regulation. Further complexity stems from the fact that the detection, removal and disabling of hate speech is increasingly reliant on automated means, which may not be a good fit for hate speech detection, which tends to be contextual. This lack of contextual understanding can lead to “false positives” and “false negatives.” While the former has a negative bearing on free speech, the latter may impact individual rights, human dignity, equality and non-discrimination.

In such a context, it becomes essential to ensure that fundamental rights protection is not fully outsourced or automated. According to the *OSCE 2021 Policy Manual on Artificial Intelligence and Freedom of Expression*, strong automation transparency frameworks should be combined with “human rights due diligence” as part of human rights-compliant content governance policies.¹¹⁸ The Policy Manual advocates against requiring intermediaries to implement proactive measures based on automated tools. Instead, it recommends introducing an obligation on digital intermediaries to perform robust human rights impact assessments vis-à-vis their algorithmic content moderation, whilst noting the importance of inclusive and participatory processes for designing and implementing automated systems.

What I Learned in Twitter Purgatory, THE ATLANTIC (Sept. 8, 2020), <https://www.theatlantic.com/ideas/archive/2020/09/what-i-learned-twitter-purgatory/616144/>; ELISKA PIRKOVA, MATTHIAS KETTEMANN, MARLENA WISNIAK, MARTIN SCHEININ, EMMI BEVENSEE, KATIE PENTNEY, LORNA WOODS, LUCIEN HEITZ, BOJANA KOSTIC, KRISZTINA ROZGONYI, HOLLI SARGEANT, JULIA HAAS, & VLADAN JOLER, SPOTLIGHT ON ARTIFICIAL INTELLIGENCE AND FREEDOM OF EXPRESSION: A POLICY MANUAL 29 (Deniz Wagner & Julia Haas eds., 2021).

¹¹³See Digital Services Act at art. 15.

¹¹⁴*Id.* at art. 16(6).

¹¹⁵*Id.* at art. 17(3)(c).

¹¹⁶*Id.* at art. 20(6).

¹¹⁷For more on the EU as a “Union of values,” see MARK DAWSON & FLORIS DE WITTE, EU LAW AND GOVERNANCE (2022).

¹¹⁸PIRKOVA ET AL., *supra* note 112.

Still, the EU may need to mobilize a broader range of policies and instruments to cope with hate speech effectively. No society is immune to hate, but whether such hatred is tamed or diffused, bolstered, and consolidated also depends on measures taken to address the underlying tensions which provide the fertile ground in which hate speech can flourish. There is much to be gained from EU policy in the spheres of education, culture, cohesion, research, or immigration including actions aimed at decoding and mitigating the hate narrative, fostering social inclusion and resilience, and supporting the integration and empowerment of vulnerable and marginalized groups. Support measures such as the Citizens, Equality, Rights and Values Programme,¹¹⁹ Horizon Europe,¹²⁰ Creative Europe,¹²¹ the European Social Fund Plus,¹²² the European Regional Development Fund,¹²³ and the Asylum, Migration and Integration Fund¹²⁴ could make an important contribution in this respect.

Acknowledgements. An earlier version of this article was presented at the e-Conference “EU Values, Diversity and Intercultural Dialogue: Enhancing the Debate”, co-organized by the Jean Monnet Project EU VaDis (<https://jmpeuvalidis.uom.gr/>) with the Hellenic Association for European Law (HAEL), in collaboration with the Center for Research on Democracy and Law, and with the support of the European Parliament Liaison Office in Greece, on 21-23 April 2021. I would like to thank the organizers for their kind invitation and the anonymous reviewers at GLJ for their helpful comments.

Competing Interest. The author declares none.

Funding Statement. No specific funding has been declared for this article.

¹¹⁹Regulation (EU) 2021/692 of the European Parliament and of the Council of April 28, 2021, Establishing the Citizens, Equality, Rights and Values Program and Repealing Regulation (EU) No 1381/2013 of the European Parliament and of the Council and Council Regulation (EU) No 390/2014, 2021 O.J. (L 156) 1.

¹²⁰Regulation (EU) 2021/695 of the European Parliament and of the Council of April 28, 2021, Establishing Horizon Europe—the Framework Program for Research and Innovation, Laying Down Its Rules for Participation and Dissemination, and Repealing Regulations (EU) No 1290/2013 and (EU) No 1291/2013, 2021 O.J. (L 170) 1.

¹²¹Regulation (EU) 2021/818 of the European Parliament and of the Council of May 20, 2021, Establishing the Creative Europe Program (2021 to 2027) and Repealing Regulation (EU) No 1295/2013, 2021 O.J. (L 189) 34.

¹²²Regulation (EU) 2021/1057 of the European Parliament and of the Council of June 24, 2021, Establishing the European Social Fund Plus (ESF+) and Repealing Regulation (EU) No 1296/2013, 2021 O.J. (L 231) 21.

¹²³Regulation (EU) 2021/1058 of the European Parliament and of the Council of June 24, 2021, On the European Regional Development Fund and On the Cohesion Fund, 2021 O.J. (L 231) 60.

¹²⁴Regulation (EU) 2021/1147 of the European Parliament and of the Council of July 7, 2021, Establishing the Asylum, Migration, and Integration Fund, 2021 O.J. (L 251) 1.