CELE
Centro de Estudios en Libertad de Expresión

Research Paper

69

Moderating Hate or Moderating Rights? The Paradox of the European Approach to Online Hate Speech and Platform Liability

Natalie Alkiviadou

August 2025

Natalie Alkiviadou, "Moderating Hate or Moderating Rights? The Paradox of the European Approach to Online Hate Speech and Platform Liability", Artículo de investigación No. 69 (ENG), Centro de Estudios en Libertad de Expresión (CELE), Buenos Aires (2025)

UP
Universidad de Palermo

Facultad de Derecho

Centro de Estudios en Libertad de Expresión

Moderating Hate or Moderating Rights? The Paradox of the European Approach to Online Hate Speech and Platform Liability

Natalie Alkiviadou

Senior Research Fellow, The Future of Free Speech at Vanderbilt University

natalie@futurefreespeech.com

August 2025

Abstract

This paper critically assesses the European approach to regulating online hate speech through platform liability frameworks, focusing on Germany's Network Enforcement Act (NetzDG) and the European Union's Digital Services Act. It argues that these laws, while aiming to curb online harm, risk infringing on the right to freedom of expression and non-discrimination by delegating content moderation powers to private, profit-driven companies. The paper highlights how tight removal deadlines, legal ambiguity, and reliance on Artificial Intelligence contribute to over-censorship, particularly of lawful but controversial speech. It also examines the unintended consequences of the current regulatory approach in Europe, including the creation of echo chambers and the potential contribution to social unrest. In addition to platform liability legislation, jurisprudence from the European Court of Human Rights also reveals the worrying trend of assigning speech moderation duties to private entities. Ultimately, the paper advocates for the endorsement of Article 20(2) of the International Covenant on Civil and Political Rights, and its accompanying Rabat Plan of Action, as a global benchmark for social media platforms by which to address online hate speech. This rights-respecting framework offers a high threshold for restrictions, ensuring that efforts to counter hate do not undermine democratic values or fundamental freedoms.

Keywords hate speech, platform liability, Digital Services Act, freedom of expression, non-discrimination.

Introduction

As the explosive expansion of social media has created new avenues for public discourse, it has also allowed for the distribution and enhanced visibility of phenomena such as hate speech. In Europe, Germany was a pioneer in legislating on platform liability for purposes of tackling online content including, but not limited to, insults and defamation of religions. Specifically, in 2017, it enacted the Network Enforcement Act (NetzDG). The law came into full force at the start of 2018. Provisions found therein required private companies, not bound by International Human Rights Law (IHRL), to eliminate 'manifestly illegal' speech in as little as 24 hours and 'illegal' speech within one week. The distinction between the two is not made clear in the law. The NetzDG has recently been replaced by the Digital Services Act (DSA), which came into force in 2024 across the 27 Member States of the European Union (EU). The DSA mandates platforms to remove 'illegal content,' including 'hate speech' and 'unlawful discriminatory content' without 'undue delay.'

Private profit-driven entities now have the power and, as demonstrated in this paper, the legal obligation in Europe to decide what constitutes hate speech and remove it. In light of the above, the central premise of this paper is that the new modus operandi adopted by the EU (and previously by Germany but also other countries such as the UK) to deal with online hate speech risks violating IHRL. At the same time, while relevant laws seek to curb online harm, they are dangerous for free speech and can have effects which backfire and instead cause the amplification of hate speech, including the migration of hate groups to underground platforms, the creation of echo chambers, and the fuelling of haters' self-proclaimed martyrdom. Restricting speech may also lead to enhanced social conflict and polarization. Moreover, authoritarian regimes have borrowed laws, such as the NetzDG, or used them as a justification to stifle dissent.

The paper commences its explanation by laying out the definition of hate speech, pointing out the lack of a consensus as to its meaning. It then discusses prominent platform liability legislation, the 2017 Network Enforcement Act (NetzDG) in Germany as a pioneer in the field, and the EU's DSA, assessing their impact on freedom of expression. It then discusses a landmark case of the European Court of Human Rights (EctHR), regarding platform liability when it comes to hate speech. The paper finally criticizes the deficiencies in the current approach adopted by the European Union (EU), pointing out the negative impact on the right to free expression as well as on the fight against online hate speech more broadly.

Hate Speech: What is it?

Despite its prevalent use in legal, academic and public discourse, hate speech is an ambiguous phrase, and its exact definition is disputed. [1] The 2019 United Nations (UN) Strategy and Plan of Action on Hate Speech stipulates that 'there is no international legal definition of hate speech, and the characterisation of what is 'hateful' is controversial and disputed.'[2] However, on a UN level, we do have what could be referred to as a hate speech clause. Specifically, Article 20(2) of the International Covenant on Civil and Political Rights (ICCPR), prohibits 'any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence.' It must be noted that characteristics such as gender, gender identity and sexual orientation are not incorporated into Article 20(2) or in the below-mentioned Framework Decision of the EU, nor are there any respective documents dealing with these protected characteristics. In 2021, a report by the UN's Special Rapporteur on Freedom of Opinion and Expression, noted that hate speech has a negative impact on the right of women to exercise their freedom of expression. She underlined that:

> 'although gender and sex are not mentioned in article 20(2), they can and should be considered grounds for protection in view of the gender equality clauses elsewhere in the Covenant and the broader intersectional approach to non-discrimination that international human rights law has consistently taken in recent decades.'[3]

The non-binding aforementioned Plan of Action defines hate speech as:

> 'any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identified factors.' [4]

On an EU level, the European Parliament has recognized the shortcomings arising from the absence of a standardized definition of hate speech and has encouraged the Commission to investigate the viability of creating a uniform legal definition of hate speech within the EU,

---

[1] See, amongst others, Natalie Alkiviadou, 'Regulating Hate Speech in the EU' in Stavros Assimakopoulos, Fabienne H Baider & Sharon Millar (eds), '*Online Hate Speech in the EU: A Discourse Analytical Perspective*' (1st edn. Springer Briefs in Linguistics, New York 2017); Audun Fladmoe & Marjan Nadim, 'Silenced by Hate? Hate Speech as a Social Boundary to Free Speech' in Arnfinn H. Midtbøen, Kari Steen-Johnsen and Kjersti Thorbjørnsrud (Eds.) '*Boundary Struggles: Contestations of Free Speech in the Norwegian Public Sphere'* (Cappelen Damm Akademisk, Oslo 2017); Iginio Gagliardone et al., *'Countering Online Hate Speech'* (2015 UNESCO)

[2] United Nations Strategy and Plan of Action on Hate Speech:
<https://www.un.org/en/hate-speech/un-strategy-and-plan-of-action-on-hate-speech> [Accessed 2 April 2025]

[3] Ibid. para.69

[4] Ibid.

which has not yet occurred.[5] In 2008, the EU adopted the Framework Decision. This is not a hate speech law but rather an instrument criminalizing racism and xenophobia, including racist and xenophobic speech. In this ambit, it prohibits:

> 'public incitement to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin.'

As with Article 20(2) of the ICCPR, characteristics such as sex, sexual orientation and gender identity are missing from the Framework Decision. The Framework Decision was the initial major EU-law instrument for combating hate speech, albeit restricted in only applying to racism and xenophobia. Even while it specifically targets both hate speech and hate crime, the words 'hate speech' are not used in its title or the body. It only penalizes some forms of speech on the grounds of 'race, colour, religion, descent, national or ethnic origin' and only considers racist and xenophobic motivation as aggravating elements. It leaves out other grounds such as sex, gender identity, or sexual orientation.
Recognizing the gap, the European Parliament asked the Commission to table a 'recast' Framework Decision for the inclusion of other forms of bias crime and incitement to hatred, including on the grounds of sexual orientation and gender identity.' [6]In 2014, in reaction to the inadequate response of the EU on homophobia, the Parliament issued a resolution affirming that the EU does not have an overall strategy for the protection of the fundamental rights of LGBTI individuals.[7] It asked for concerted action on the part of the Commission, the Member States and concerned agencies, to formulate an action plan for the defence of LGBTI rights for a multi-year period.

In her State of the Union speech in 2020, the President of the European Commission said that the Commission would work towards broadening the definition of EU crimes in order to encompass all types of hate crime and hate speech, whether on the grounds of race, religion, gender or sexuality.[8] Nevertheless, at an EU level, no changes have yet been made to ensure equal treatment of protected characteristics in the sphere of hate speech.

---

[5] Motion for a European Parliament Resolution on Establishing a Common Legal Definition of Hate Speech in the EU B8-0172/2017 (2017)

[6] European Parliament: Joint Motion for a Resolution on Strengthening the Fight against Racism, Xenophobia and Hate Crime (11 March 2013) (2013/2543(RSP)) <https://www.europarl.europa.eu/doceo/document/RC-7-2013-0121_EN.html?redirect> [Accessed 14 May 2025]

[7] European Parliament Resolution of 4 February 2014 on the EU Roadmap against Homophobia and Discrimination on Grounds of Sexual Orientation and Gender Identity (2013/2183(INI)) <https://www.europarl.europa.eu/doceo/document/TA-7-2014-0062_EN.html> [Accessed 14 May 2025]

[8] European Commission, 'The Commission Proposes to Extend the List of 'EU Crimes' to Hate Speech and Hate Crime' (2021) <https://ec.europa.eu/commission/presscorner/detail/en/ip_21_6561> [Accessed 14 May 2025]

The lack of certain protected characteristics from the UN and EU frameworks demonstrates an inherent problem with hate speech laws. Laws that protect some groups but not others can lead to what Volokh has labelled as 'censorship envy.'[9] When those groups left out start doubting how equitable the system is, asking: "If my neighbour is permitted to suppress speech they find offensive, why should I not have the same privilege?", such inconsistencies can exacerbate social gaps and undermine the legitimacy of the hate speech laws, as well as the general values of equality and freedom of expression. The regulatory bias against any point of view not only undermines the fundamental value of free speech, but it can also reinforce the very social inequities that the law should seek to improve. Furthermore, if human rights organizations approach these matters inconsistently, it undermines their credibility and weakens the general defense of freedom of expression.

In terms of a non-binding definition, the EU's Fundamental Rights Agency (FRA) has referred to hate speech as:

> 'the incitement and encouragement of hatred, discrimination, or hostility towards an individual that is motivated by prejudice against that person because of a particular characteristic.'[10]

On a Council of Europe level, some non-binding documents provide conceptualizations of hate speech. For example, General Policy Recommendation 15 on Combating Hate Speech (2015) of the European Commission against Racism and Intolerance (ECRI) notes that hate speech is:

> 'the advocacy, promotion or incitement, in any form, of the denigration, hatred or vilification of a person or group of persons, as well as any harassment, insult, negative stereotyping, stigmatization or threat in respect of such a person or group of persons and the justification of all the preceding types of expression, on the ground of race, colour, descent, national or ethnic origin, age, disability, language, religion or belief, sex, gender, gender identity, sexual orientation and other personal characteristics or status.'

In relation to the ECtHR, Gűndűz v Turkey (2003), which was a case involving speech calling for the implementation of Sharia, the Court referred to hate speech as 'all forms of expression which spread, incite, promote or justify hatred based on intolerance (including religious intolerance).'[11] This definition was based on ECRI's Recommendation No. R (97) 20 of the Committee of Ministers on hate speech.

---

[9] Eugene Volokh, 'Censorship Envy' *The Volokh Conspiracy – Reason* (13 November 2023) <https://reason.com/volokh/2023/11/13/censorship-envy-2/ [Accessed 19 May 2025]

[10] Fundamental Rights Agency, 'Hate Speech and Hate Crimes against LGBT Persons' <https://fra.europa.eu/sites/default/files/fra_uploads/1226-Factsheet-homophobia-hate-speech-crime_EN.pdf > [Accessed 10 May 2025]

[11] Gűndűz v Turkey, Application No. 35071/97 (ECHR 4 December 2003) para. 40

When it comes to social media platforms themselves, definitions of hate speech can be quite broad. For example, Meta defines 'hateful conduct' as:

> 'direct attacks against people—rather than concepts or institutions—based on what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease.'[12]

The phrase 'hate speech' was substituted with 'hateful conduct' in significant revisions made to Meta's policy on January 7, 2025. Meta's policy extends protection to individuals based on characteristics such as age and occupation, but only when these are targeted in combination with other protected traits. It also explicitly protects refugees, migrants, immigrants and asylum seekers from what it classifies as the most severe forms of attack. Prohibited content is categorized on two separate levels. Tier 1 includes the most egregious offenses as defined by Meta, such as dehumanization, incitement to violence, insults, and detrimental stereotypes historically linked to intimidation or violence. Tier 2 addresses less egregious manifestations of hate speech, including appeals for social exclusion, endorsement of segregation and general derogatory remarks. Meta seeks to encompass a wider spectrum of detrimental discourse by implementing a two-tiered approach.

X's policy on hateful conduct and incitement addresses a range of harmful content, including 'stereotypes, discrimination and harassment, slurs and tropes, dehumanisation,' as well as hateful imagery such as the swastika. Notably, incitement to violence is governed by a separate policy. A key aspect of X's approach is its recognition of contextual nuance: the rules explicitly acknowledge that 'some posts may appear to be hateful when viewed in isolation but may not be when viewed in the context of a larger conversation.' This is particularly relevant in cases where slurs are used by members of minority communities to 'reclaim terms that were historically used to demean individuals.' Such provisions highlight the policy's emphasis on both language and context. The protected characteristics under X's policy include race, ethnicity, national origin, caste, sexual orientation, gender identity, religious affiliation, age, disability, and serious disease.[13]

YouTube provides that:

[12] Meta – Hateful Conduct Policy <https://transparency.meta.com/policies/community-standards/hateful-conduct/> [Accessed 8 April 2025]

[13] X, Hateful Conduct Policy <https://help.x.com/en/rules-and-policies/hateful-conduct-policy> [Accessed 8 April 2025]

> 'hate speech is not allowed on YouTube. We don't allow content that promotes violence or hatred against individuals or groups based on any of the following attributes, which indicate a protected group status under YouTube's policy: Age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/gender, sexual orientation, victims of major violent events and their kin, veteran status.'[14]

YouTube's hate speech policy includes many protected categories, such as immigration status, veteran status, and victims of violent incidents, alongside more widely acknowledged characteristics like color and gender. The broad span is a distinguishing characteristic of YouTube's strategy, as emphasized in a 2023 analysis on the hate speech rules of eight prominent social media companies, including YouTube, X and Meta. The report saw a substantial expansion of these policies over time, encompassing both the targeted material and the spectrum of protected traits. In the mid-2000s and early 2010s, platforms primarily concentrated on forbidding the promotion of hatred and blatantly racist discourse, but subsequent advancements broadened these prohibitions to encompass, for example, harmful stereotypes, conspiracy theories and slurs aimed at protected groups.[15]

In light of the brief overview of hate speech definitions on three major platforms, it can be argued that companies have mostly embraced broad definitions of hate speech, frequently surpassing legal requirements and the principles of IHRL.[16]

On an academic level, scholars have grappled for many years with defining hate speech in a manner that would capture its multifaceted, context-specific character. Various academic definitions have been suggested, emphasizing different aspects of dangerous speech. Matsuda isolates three essential features of hate speech: it bears a message 'of racial inferiority,' is 'directed against historically oppressed groups' and is 'persecutory, hateful, and degrading.'[17] McGonagle takes the broadened approach, advancing the notion that 'virtually all racist and related declensions of noxious, identity-assailing expression could be brought into the wide compass of the term.'[18] Smolla broadens the definition beyond racism, defining hate speech as an 'umbrella term that has come to cover the use of speech attacks on the grounds of race,

---

[14] YouTube Hate Speech Policy <https://support.google.com/youtube/answer/2801939?hl=en> [Accessed 9 April 2025]

[15] Jacob Mchangama, Abby Fanlo & Natalie Alkiviadou, 'Scope Creep: An Assessment of 8 Social Media Policies' *Justitia* <https://futurefreespeech.org/scope-creep/> [Accessed 15 April 2025]

[16] For more on this, see Jacob Mchangama, Natalie Alkiviadou & Raghav Mendiratta 'A Framework of First Reference – Decoding a Human Rights Approach to content moderation on social media' (2021) *Justitia*

[17] Mari J. Matsuda, 'Public Response to Racist Speech: Considering the Victim's Story' (1989) 87 *Michigan Law Review* 8, p.2357

[18] Tarlach McGonagle, 'Wresting Racial Equality from Tolerance of Hate Speech' (2001) 23 *Dublin University Law Journal* 21, p.4

ethnicity, religion, and sexual orientation or preference.'[19] Likewise, MacKinnon makes the comparison with pornography, characterizing hate speech as an instrument of social control, affirming racial and gender domination.[20] Notwithstanding similarities among these definitions, including identity-based harm, historical persecution and structural inequality, the term is extremely controversial. Consequently, it has been noted that 'hate speech appears to be whatever individuals elect it to be,'[21] emphasizing the ambiguity of the term.

Platform Liability Legislation: The NetzDG and the DSA

The German NetzDG

The German "Netzwerkdurchsetzungsgesetz" or Network Enforcement Act (NetzDG) is probably the most contested measure implemented by a liberal democracy to counter illicit online content. It was enacted in 2017 and entered into force on the 1st of January 2018. Due to the EU-wide changes brought about by the DSA in 2024, the NetzDG has been largely repealed, with its European counterpart taking precedence over it. Nevertheless, the German legislation occupies a central position in any academic discourse on platform liability owing to its substantial precedential value and is, therefore, examined in this paper.[22]

Following prolonged debates among legislators, civil society actors and major social media platforms, the NetzDG was enacted as a legislative response to the growing prevalence of hate speech and other unlawful content online. Its introduction was closely tied to a shifting political climate marked by mounting public dissatisfaction, much of it playing out across social media, with the federal government and then Chancellor Angela Merkel's approach to refugee reception. This discontent gave rise to xenophobic narratives targeting both incoming refugees and German officials. This constituted the central backdrop against which the NetzDG was created.

The NetzDG applied to social media platforms with at least two million users in Germany, specifically Facebook, Instagram, X (at the time of enactment, Twitter), YouTube and TikTok. Section 1(3) outlines what content is considered illegal for purposes of the NetzDG. The law did not create new crimes but, rather, transferred the criminality of certain speech from the offline to the online setting by referencing the country's Criminal Code. This included, inter alia, incitement to hatred, defamation of religions, religious and ideological

---

[19] Rodney A. Smolla, 'Academic Freedom, Hate Speech and the Idea of a University' 53 *Law and Contemporary Problems* 3, p.195

[20] Catherine A. MacKinnon, *'Only Words'* (1st edn. Harvard University Press 1996)

[21] 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* 1, p.110

[22] Nicole Goebel, 'Enforce Community Standards, Facebook Told' (2015) *DW* <https://www.dw.com/en/facebook-must-ban-abusive-content-says-german-justice-minister-maas/a-18676705> [accessed 27 April 2025]

associations, and insult. Section 3(2) therein provided that 'manifestly unlawful' content is to be removed within 24 hours of the platform receiving a complaint, and all other unlawful content must be removed within one week. The law did not elaborate on how a platform should determine between what is unlawful and what is manifestly unlawful. Importantly, the NetzDG noted that the decision regarding the one-week timeline may be surpassed in cases where the decision is dependent on 'the falsity of a factual allegation or is clearly dependent on other factual circumstances.' While this provision allows the platform to provide the user with a chance to respond to a complaint before a decision is made, it is noteworthy that no respective provision on the issue of legal assessment is provided. In brief, ascertaining the legality of content was not considered as a factor affecting the short time limit imposed on companies. Importantly, as per section 4, only the relevant administrative authority, namely, the Federal Office of Justice, can seek recourse to the courts regarding the legality of content. Specifically, in seeking to find that a regulatory offense has occurred and, thus, impose a fine on the company on the grounds that unlawful content has not been removed, it must first obtain a judicial decision establishing such unlawfulness. While this is an additional safeguard, on the one hand, and, whilst this process includes a statement by the company, there is no equivalent and direct recourse to the court provided to companies for purposes of determining the legality of the content. However, while not a judicial body, the NetzDG Law did allow for companies to refer the decision regarding the unlawfulness of content to a recognized self-regulation institution, namely, the FSM.[23] Further, the law obliged companies to publish twice-yearly transparency reports explaining how they manage complaints over illegal content. Failure to implement an adequate complaint management system or to submit complete and accurate transparency reports may result in administrative fines of up to 50 million euros. For example, in 2019, Facebook received a 2 million euro fine for violating the NetzDG's report duty through the filing of a report that was not considered transparent enough and didn't contain enough detail.[24]

The scope and nature of the NetzDG have been the subject of extensive debate, with critics arguing that the law is vague and overly broad, "privatizing" online censorship without sufficient transparency or due process, and fostering "over-implementation" by incentivizing platforms to prioritize caution over freedom of expression.[25] Relevant to this is the position of France's Constitutional Council, which, when discussing provisions of the Avia Law (a

[23] FSM <https://www.fsm.de/en/fsm/netzdg/> [Accessed 27 April 2025]

[24] CMS Law-Nom, 'Facebook Fined EUR 2 Million for Infringement of Germany's Network Enforcement Act' <https://cms-lawnow.com/en/ealerts/2019/08/facebook-fined-eur-2-m-for-infringement-of-germany-s-network-enforcement-act> [Accessed 2 September 2025]

[25] 'Germany: Flawed Social Media Law: NetzDG is Wrong Response to Online Abuse' *Human Rights Watch* <https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law> [Accessed 21 April 2025]

NetzDG equivalent), held that the law directly restricted the exercise of the right to freedom of expression through the stringent removal provisions.[26]

Furthermore, the NetzDG set a dangerous scene for freedom of expression and the handling of online hate speech within Germany, but also had consequences beyond it. By 2020, over twenty countries, such as Russia, Belarus, Venezuela, Singapore and Turkey, had implemented or introduced legislation emulating the NetzDG. So, while Germany's initiative sought to establish a regulatory instrument based upon democratic principles, it inadvertently legitimized and provided a model for more repressive speech controls within and across authoritarian and semi-authoritarian regimes. In such environments where social media may be the lone platform on which minorities and dissidents can speak, the model presented in the NetzDG is particularly risky and susceptible to abuse. This abuse can serve to silence critics further, enable the exclusion of marginalized groups and manipulate public discussion through the internet. While no one is arguing that Putin or Erdogan would wait for Germany to pass a law to think of silencing their critics, the fact that such laws exist in liberal democracies most definitely provides a form of justification and legitimization for their own silencing strategies. The above finding must also be considered when it comes to the DSA and the 'Brussels Effect',[27] which may impact its cross-fertilisation into non-democratic states.[28]

The NetzDG is not the only platform liability legislation on a European (but not EU) level. For example, the United Kingdom adopted the 2023 Online Safety Act,[29] which, amongst others, requires companies to act against illegal content which includes, for example, content related to racially or religiously aggravated public order offenses. There is no scope in this essay to go into this law in detail, but basic arguments that apply to platform liability legislation about hate speech apply to the Online Safety Act as well.

The European Union: The Digital Services Act

On an EU level, the DSA was introduced with the aim of putting an end to the so-called "Wild West"[30] of the internet, establishing instead a rules-based digital framework throughout

---

[26] Patrick Breyer, 'French law on illegal content online ruled unconstitutional: Lessons for the EU to learn' (2020) available at: <https://www.patrick-breyer.de/?p=593729&lang=en> [Accessed 1 April 2025]

[27] Anu Bradford, 'The Brussels Effect: How the European Union Rules the World' (2019) available at: <https://academic.oup.com/book/36491> [Accessed 25 April 2025]

[28] Jacob Mchangama, Natalie Alkiviadou & Raghav Mendiratta, 'Thoughts on the DSA: Challenges, Ideas and the Way Forward through International Human Rights Law' *Justitia* (2022) available at <https://futurefreespeech.com/thoughts-on-the-dsa-challenges-ideas-and-the-way-forward-through-international-human-rights-law/> [Accessed 25 April 2025]

[29] Online Safety Act 2023 <https://www.legislation.gov.uk/ukpga/2023/50> [Accessed 4 April 2025]

[30] Press Release - *Digital Services: Landmark Rules Adopted for a Safer, Open Online Environment*, (2022) *European Parliament"*

the Union.[31] The DSA adopts a notice and action mechanism for the removal of illegal content, with intermediaries acting on the receipt of notices without undue delay, considering the type of content and urgency of removal. Illegal content is defined broadly and 'should be understood to refer to information, irrespective of its form, that under the applicable law is either itself illegal, such as illegal hate speech....and unlawful discriminatory content...' However, the DSA does not define 'hate speech' or 'unlawful discriminatory content.' At the same time, no explicit reference is made in the law to the Framework Decision, which constituted a benchmark for defining hate speech in the Code of Conduct. Instead, it grants platforms the power and authority to identify and determine illegal content, the illegality of which may stem either from national law, which is compliant with EU law, or from EU law itself. The disparities in national approaches generate substantial legal uncertainty and pose significant challenges for companies operating within Europe.[32] Given the particularly contentious nature of hate speech, it is essential that major legislative instruments, such as the DSA, provide a basic definition or conceptualization of the term at the very least. Unfortunately, this was not done by EU legislators.

The obligation to remove illegal content, including hate speech, and to have a notice and action mechanism in place for receiving alerts (by users or by Stats) of such content, is imposed on all companies that are impacted by this legislation, namely, those offering intermediary services. In addition, the DSA imposes increased duties for very large online platforms (over 45 million users in the EU), such as audits, mitigation of risk and the need for compliance officers. Platforms are also required to justify to users why content is removed. On one hand, this is a good development and promotes freedom of expression. That said, given the Regulation's heavy duties and the sheer amount of online material, platforms could be encouraged to take content down rather than invest in providing elaborate explanations.[33] The DSA provides that Member States should fine all intermediaries (regardless of size) that do not comply with the law's requirements. The penalty depends on several factors such as the nature, gravity and recurrency of the violation, as well as the economic capacity of the company.

In terms of how quickly content should be removed, the DSA does not directly refer to a specific timeframe but, instead, refers to the EU Code of Conduct on Countering Illegal

---

<https://www.europarl.europa.eu/news/en/press-room/20220701IPR34364/digital-services-landmark-rules-adopted-for-a-safer-open-online-environment>

[31] Ibid.

[32] Ronan Ó Fathaigh, Natali Helberger & Naomi Appelman 'The Perils of Legally Defining Disinformation' (2021) 10 *Internet Policy Review* 4, p.14

[33] Jacob Mchangama, Natalie Alkiviadou & Raghav Mendiratta, 'Thoughts on the DSA: Challenges, Ideas and the Way Forward through International Human Rights Law' (2022) < https://futurefreespeech.org/thoughts-on-the-dsa-challenges-ideas-and-the-way-forward-through-international-human-rights-law/>

Hate Speech Online (2016)[34] as an example, noting that this 'sets a benchmark for the participating companies concerning the time needed to process valid notifications for removal of illegal hate speech.' An indirect reference to the 24-hour time limit set out in the Code is thus made in the text of the legislation.

It is noteworthy that the Code was introduced by the European Commission as a voluntary co-regulatory document on the removal of online hate speech. Importantly, the Code culminated in the Code of Conduct on Countering Illegal Hate Speech Online + which, in January 2025, was integrated into the DSA. For the moment, the signatories include Facebook, Instagram, X, YouTube, TikTok, Dailymotion, Jeuxvideo.com, LinkedIn, Microsoft-hosted consumer services, Snapchat, Rakuten Viber and Twitch. In its recitals, the DSA refers to voluntary codes, noting that 'while codes should be measurable and subject to public oversight, this should not impair the voluntary nature of such codes.' However, Article 45 of the DSA equips the European Commission and the European Board for Digital Services[35] with oversight powers. Specifically, it provides, amongst others, that 'in the case of systematic failure to comply with the codes of conduct, the Commission and the Board may invite the signatories... to take the necessary action.' The phrase 'necessary action' is left undefined in the regulatory framework, but the subsumption of the Code+ under the scope of the DSA might, in effect, translate its voluntary commitments into de facto obligations enforceable through regulatory supervision. This dynamic can potentially generate an implicit enforcement framework, thus obfuscating the line between voluntary compliance and regulatory compulsion. A very central challenge associated with the new Code is the issue of timing. While the DSA refers to the removal of hate speech 'without undue delay,' and refers to the previous Code's 24-hour deadline (without a set obligation to match that under the DSA), the Code+, as was the case with its 2016 version, imposes a 24-hour deadline for removal of online content. In addition to the time pressure of this not-so-seemingly voluntary code, the signatories commit to reviewing at least 50% of notices received. The signatories will also 'apply their best efforts to go beyond this target' and aim at least 67% (two-thirds) of the notices.

The DSA also incorporates "trusted flaggers," organizations which are granted enhanced authority to detect and report harmful content with prioritized attention. This was also incorporated in the 2016 Code and in the Code+. This framework is ostensibly intended to bolster efficiency and relies on the specialized knowledge of vetted entities to manage content moderation at scale. Nevertheless, critics argue that the trusted flagger system could jeopardize transparency and open the door to political manipulation while 'political

[34] The EU Code of Conduct on Countering Illegal Hate Speech Online <https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en> [Accessed 5 May 2025]

[35] European Board for Digital Services: <https://digital-strategy.ec.europa.eu/en/policies/dsa-board> [Accessed 20 April 2025]

censorship could also be effected indirectly via civil society organizations.'[36] Moreover, it is unclear how the transparency and accountability of these trusted flaggers are to be ensured in practice.[37] These concerns are not without precedent. Germany's NetzDG law, which similarly relied on trusted flaggers, faced criticism for enabling excessive censorship, particularly of opposition groups and independent media.[38]

Platform Liability Legislation — Its Impact

In a 2021 report, the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression warned that 'by compelling social media platforms to police speech, they create a risk that companies will zealously over-remove material and undermine free speech.'[39] To illustrate this point, a 2024 survey found that a large majority of the content taken down by Facebook and YouTube in France, Germany and Sweden to be legally acceptable. Depending on the collection of data, 87.5% to 99.7% of removed comments were in accordance with national laws. The most rapid rates of legal takedowns were recorded in Germany, at 99.7% on Facebook and 98.9% on YouTube. These results raise the possibility that the German Network Enforcement Act (NetzDG) can shape platform moderation behavior, and that the fear of its strict penalties might lead to over-removal.[40]An equally, perhaps even larger, risk of over-removal might arise from the DSA.

To avoid the problems associated with non-compliance with platform liability legislation, companies may, therefore, play the 'better safe than sorry' card leading to over-removal by private regulation of the public sphere.[41] To worsen the situation, scholars such as Griffin

---

[36] Rachel Griffin, "EU Platform Regulation in the Age of Neo-Illiberalism" (2024) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4777875> 14 [Accessed 20 April 2025]

[37] Ibid.

[38] Emma Llanso, "German Social Media Law Creates Strong Incentives for Censorship" (2017) *Centre for Democracy and Technology* <https://cdt.org/insights/german-social-media-law-creates-strong-incentives-for-censorship/> [Accessed 20 April 2025]; Hans-Jörg Vehlewald, 'Habecks Netz-Aufseher Rudert Zurück (2024) Bild <https://www.bild.de/politik/inland/zensur-gegen-fake-news-habecks-netz-aufseher-rudert-zurueck-670bbb5b2de6a20c12808e7a?utm_> [Accessed 20 April 2025]

[39] https://www.ohchr.org/en/press-briefing-notes/2021/07/statement-irene-khan-special-rapporteur-promotion-and-protection

[40] Preventing "Torrents of Hate" or Stifling Free Expression Online?' (2024) *The Future of Free Speech* <https://futurefreespeech.org/preventing-torrents-of-hate-or-stifling-free-expression-online/> [Accessed 15 May 2025]

[41] Marcella Atzori, "The Digital Services Act and the Freedom of Expression in the European Union: A Political Perspective" (2024) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4887181> 3; Barrie Sander, "Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation" (2020) 43 *Fordham International Law Journal,* 939 <https://ir.lawnet.fordham.edu/ilj/vol43/iss4/3/> [Accessed 2 May 2025]

argue that the DSA does not actually incorporate rigorous checks and balances.[42] This combination of factors is deeply problematic, in that social media companies constitute the central agora of the sharing and receiving of information today, given the fundamental nature of free speech. The current approach to platform liability could mean that these private, profit-driven companies, not bound by IHRL but seeking to avoid regulatory fines and other problems with national and European authorities, remove 'legal but controversial speech.'[43] In fact, the risks faced by companies as a result of the DSA 'create strong incentives to remove content alleged to be illegal, even where the case for illegality is less than convincing or where actual legal proceedings would be unlikely.'[44] As noted by Keller, assigning enforcement responsibilities to private actors undermines democracy and creates a system in which the platforms are incentivized to over-censor material in an attempt to avoid the threat of regulatory penalty.[45]

In terms of removal deadlines, the DSA, unlike the NetzDG, does not directly refer to a time frame but notes that companies must act 'without undue delay.' Nevertheless, the DSA does make reference to the 2016 Code of Conduct on Countering Illegal Hate Speech online and its obligation of removing hate speech within 24 hours. This 24-hour deadline is further enhanced by the adoption of the Code+ as mentioned above. The time deadline is a key concern, whether it is a direct or indirect reference to 24 hours or 'undue delay.' Content moderation of a contentious area of speech, such as hate speech, should include a nuanced assessment of issues such as context and others, discussed below in the framework of the RPA.

To deal with legislative obligations and avoid hefty fines, companies are faced with a huge task given the volume of online content. This is not only because of the DSA, although the situation will now become much more problematic because of it. The utilization of Artificial Intelligence (AI) by social media platforms is partly a response to increasing governmental demands to rid platforms of hate speech quickly and adequately. Companies are further subject to pressures from advertisers and users, many of whom demand timely and all-encompassing content moderation. To appease these demands and to avert hefty fines, platforms increasingly depend on AI technologies, either solely or in conjunction with human moderators, to detect and take down hate speech.

---

[42] Ibid.

[43] Rikke Frank Jørgensen & Lumi Zuleta, 'Private Governance of Freedom of Expression on Social Media Platforms' (2020) *Nordicom Review* 41, pg.53

[44] Rachel Griffin, "EU Platform Regulation in the Age of Neo-Illiberalism" (2024) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4777875> 15

[45] Daphne Keller, "*Who Do You Sue? State and Platform Hybrid Power Under the DSA"* (2019) *Hoover Institution* < https://www.hoover.org/research/who-do-you-sue> [Accessed 4 May 2025]

For example, Meta's Transparency Centre notes that its technology proactively detects and removes the vast majority of content (90%) before anyone reports it.[46] The below graphic demonstrates how removed comments are, by a large majority (99.7%), removed by automated flagging.

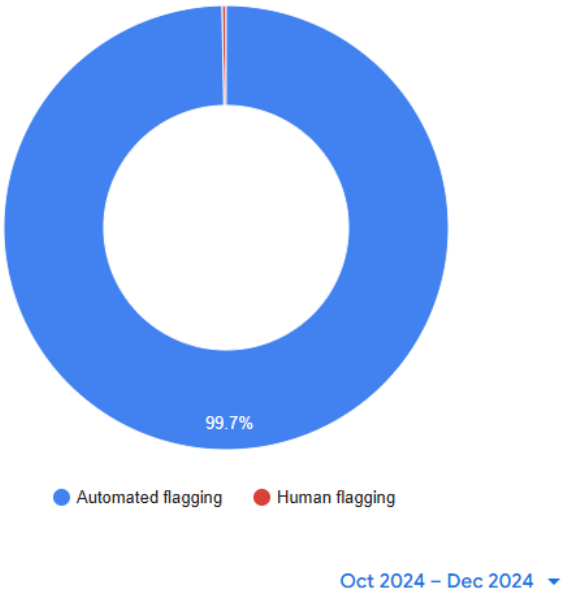

Figure 1. Automated vs. Human Flagging.

As Dias notes, these factors have caused companies to 'act proactively in order to avoid liability... in an attempt to protect their business models.' [47] The Council of Europe has previously warned that enhanced reliance on AI for moderating online content could lead to over-removal and, thereby, put freedom of expression at risk.[48] Llanso et al. argue that there is a 'strong presumption against the validity of prior censorship in international human rights law.'[49]

---

[46] Meta – Transparency Center <https://transparency.meta.com/enforcement/detecting-violations/> [Accessed 14 May 2025]

[47] Thiago Oliva Dias, 'Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression,' (2020) 20 *Human Rights Law Review* pp.607-640

[48] 'Algorithms and Human Rights - Study On the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications' (2018) *Council of Europe* <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> [Accessed 20 May 2025]

[49] Emma Llanso et al., 'Artificial Intelligence, Content Moderation and Freedom of Expression.' (2020) Transatlantic Working Group < https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf> [Accessed 4 May 2025]

Given the sheer volume of content, it would be unrealistic to believe that human moderators could do it all, and this is where AI comes in as a useful tool. While this is absolutely imperative for illegal content such as child sexual abuse material, the moderation of hate speech cannot be AI reliant. When it comes to contentious content such as hate speech, the increasing use of AI risks enhancing the capacity of content moderation but, also, 're-obscur[ing] the fundamentally political nature of speech decisions being executed at scale.'[50] Moreover, using AI for moderating content such as hate speech raises core concerns when it comes to transparency, the respect for the rule of law and accountability.[51] Contemporary technological tools employed to identify harmful textual content predominantly rely on natural language processing and sentiment analysis. While these tools have demonstrated significant advancements, their accuracy remains within the range of 70 to 80 %. While this may seem to be a high percentage, AI struggles to interpret the nuanced dimensions of human communication, particularly in discerning speaker intent or motivation which are core factors in the ambit of hate speech.[52] Consequently, these tools frequently misinterpret context, which can adversely affect users' rights to freedom of expression, access to information and equality. In addition, Dias et al. emphasize the potential for meaning to shift when normative content moderation policies are translated into machine-readable code, due to the inherent limitations of computational language compared to human linguistic expression.[53]

While the use of AI in regulating online hate speech is inherently problematic to the exercise of free speech, as such automated mechanisms cannot pick up on the nuances of language, it is also problematic for the right to non-discrimination and the broader principle of equality. Dias argues that the use of AI to implement content moderation can result in the discriminatory enforcement of social media platforms' terms of services.[54] This danger stems from several reasons, such as inadequately or even biased training sets, which can

[50] Robert Gorwa et al, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) *Big Data & Society* <https://journals.sagepub.com/doi/full/10.1177/2053951719897945> [Accessed 25 April 2025]

[51] 'Privacy and Freedom of Expression in the Age of Artificial Intelligence' (2018) *Article 19,* p.15 .<https://www.article19.org/wp-content/uploads/2018/04/Privacy-and-Freedom-of-Expression-In-the-Age-of-Artificial-Intelligence-1.pdf> [Accessed 10 May 2025]

[52] Natasha Duarte and Emma J. Llansó, 'Mixed Messages? The Limits of Automated Social Media Content Analysis.' (2018) Proceedings of the 1st Conference on Fairness, Accountability and Transparency, *PMLR* 81 pp. 106-106.

[53] Thiago Oliva Dias et al., 'Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risk to LGBTQ Voices Online' (2021)

[54] Thiago Oliva Dias, 'Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression' (2020) 20 *Human Rights Law Review* 4

disproportionately impact and potentially silence minority groups.[55] Further, technological developments such as natural language processing and sentiment analysis can detect harmful text without having to rely on specific words or phrases. However, research has demonstrated that they are 'still far from being able to grasp context or to detect the intent or motivation of the speaker.'[56] For example, research has shown that automated moderation mechanisms often fail to interpret the context specific language used by marginalized communities, such as the LGBTQ community's use of 'mock impoliteness' and reclaimed slurs like "dyke," "fag," and "tranny," which serve as tools of empowerment and resilience. They cite instances which have led to content removal, for example, a trans woman being banned from Facebook for referring to herself as a "tranny" in a post about her hairstyle.[57] Additionally, research has shown that tweets in African American English are twice as likely to be flagged as offensive, illustrating how racial bias can permeate algorithmic content moderation.[58]

The European Court of Human Rights: The case of Delfi v Estonia

Delfi v Estonia, was a seminal case on internet hate where the ECtHR ruled that Internet intermediaries, such as news portals, should remove hate speech uttered by their users. Specifically, it held that:

> '...where third-party user comments are in the form of hate speech and direct threats to the physical integrity of individuals, the member States may be entitled to impose liability on Internet news portals if they fail to take measures to remove clearly unlawful comments without delay, even without notice from the alleged victim or from third parties.'[59]

The applicant was the owner of Delfi, a portal providing internet-based news. At that time, users of the portal had been able to publish comments below each article published by Delfi. These comments were made publicly available to other users. User-generated comments appeared instantly without any prior review or editorial intervention by Delfi. On average, about 10,000 comments were made daily, most written using pseudonyms. Even though the portal was installed with an automated filter that would identify and eliminate obscenities, comments were not subject to prior screening. Delfi had implemented certain content rules barring offending, mocking or hate-inciting comments. Moreover, the portal had a

[55] Emma Llanso et al., "Artificial Intelligence, Content Moderation and Freedom of Expression' (2020) *Transatlantic Working Group* <https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf> [Accessed 15 May 2025]

[56] Thiago Oliva Dias et al., 'Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risk to LGBTQ Voices Online' (2021) 25 *Sexuality and Culture*

[57] Ibid.

[58] Maarten Sap, 'The Risk of Racial Bias in Hate Speech Detection' *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019), pg. 1672

[59] Delfi AS v Estonia, Application no. 64569/09 (ECtHR 16 June 2015) para. 100

notice-and-takedown procedure in place: users would notify the portal, and Delfi would eliminate offending comments. Targets of defamatory statements would also be able to ask for them to be eliminated.

In 2006, Delfi published a news item ccalled "SLK Destroyed Planned Ice Road." The initials SLK stood for Saaremaa Shipping Company, a public limited liability. At the time in question, L occupied a seat in SLK's supervisory board and was a solo, or even majority, shareholder. Within a few days of being published, the story had garnered 185 user comments, at least 20 of them containing personal insults and threats made against L. They included statements like

> 'bloody shitheads... they bathe in money anyway thanks to that monopoly and State subsidies and have now started to fear that cars may drive to the islands for a couple of days without anything filling their purses. Burn in your own ship, sick Jew!'

The ECtHR determined that Delfi's efforts to remove comments, which the Court deemed as 'hate speech and speech inciting violence'[60] were inadequate, resulting in a six-week delay in their removal. The ECtHR emphasized that while an 'automatic word-based filter may have been useful in some instances, the facts of the present case demonstrate that it was insufficient for detecting comments whose content did not constitute protected speech under Article 10 of the Convention.'[61] Notably, the Court refrained from addressing the broader compatibility between the use of AI and the rights to freedom of expression and information, anddid not explore the specific limitations on freedom of expression as set out in Article 10. In its reasoning, the ECtHR characterized the comments at issue as being of an 'extreme nature,'[62] yet offered no further clarification as to why these statements were considered extreme or how the notion of extremism was defined. Importantly, the Court did not comment on the broader compatibility between the use of Artificial Intelligence and the Freedoms of Expression and Information, nor did it delve into the specific limits of freedom of expression as defined by Article 10.

Additionally, the ECtHR considered that Delfi was a professionally managed news portal operating commercially. Even though the Court did not directly specify how Delfi's status affected its duty to eliminate illegal content, it can be assumed that the imposition of immediate removal duties ensued from the site's dual status as a professional media outlet and a business venture. The Court's attempt to distinguish between censorship and the legal obligation to manage online hate speech was also informed by Delfi's status as a source of key information. Specifically, it was found by the Court that:

---

[60] Delfi v Estonia, App. No 65469/09 (ECtHR 16 June 2015), Para. 162
[61] Ibid.
[62] Delfi v Estonia, App. No 65469/09 (ECtHR 16 June 2015), Para. 162

> 'a large news portal's obligation to take effective measures to limit the dissemination of hate speech and speech inciting violence – the issue in the present case – can by no means be equated to "private censorship."[63]

We have thus seen the European regional human rights court delegate the responsibility of restricting speech, a fundamental right, to private companies that are not bound by IHRL. Strikingly, this crucial issue was not even addressed in the Court's reasoning. In reaching this conclusion, the Court did acknowledge the Internet's vital role in the dissemination of information.

As such, the ECtHR concluded that Delfi, a large professional, profit-making news portal, had a legal obligation promptly to remove user-generated comments that constituted hate speech and found the six-week delay in doing so unjustifiable. Its decision has been described as 'unexpected,'[64] 'controversial,'[65] and a restriction of the freedom of expression.[66] European Digital Rights found that this ruling curtails the rights of Internet users as it is 'not obvious why the Court appears to have given almost absolute priority to third party rights ahead of the free speech rights of commentators.'[67] What likely wasn't anticipated at the time, however, was that, eight years later, such responsibilities would be further extended to individual users, as confirmed in Sanchez v. France (2023). Here, the ECtHR extended moderation responsibilities to individual users.

The case of Sanchez v France (2023), decided by the Grand Chamber in 2023, is yet another concerning development in matters related to freedom of expression. Sanchez was criminally convicted, not for a statement he made, but for not promptly deleting third-party comments made on his Facebook post, which were found to incite hatred or violence against a person or a group due to religion. Sanchez was, at this time, a Front National parliamentary candidate in the Nîmes constituency. He had published a post concerning F.P.'s political party website, being a Member of the European Parliament. In reply, a third-party user, S.B., wrote that F.P. has:

---

[63] Ibid. Para.157

[64] Tatiana Synodinou, 'Intermediaries' Liability for Online Copyright Infringement in the EU: Evolutions and Confusions' (2015) 31 *Computer Law & Security Review* 1, p.63

[65] Hugh J. McCarthy, 'Is the Writing on the Wall for Online Service Providers? Liability for Hosting Defamatory User-Generated Content Under European and Irish Law' (2015) 14 *Hibernian Law Journal* 39

[66] Ibid.

[67] EDRi Paper for the Council of Europe: "Human Rights Online" (2014) *EDRi* <*https://edri.org/our-work/edri-coe-human-rights-online/*> [Accessed 20 May 2025]

> 'transformed Nîmes into Algiers, there is not a street without a kebab shop and mosque; drug dealers and prostitutes reign supreme, no surprise he chose Brussels, capital of the new world order of sharia... Thanks [F.] and kisses to Leila ([L.T]) ... Finally, a blog that changes our life ...'

Another user, L.R., added three other comments directed at Muslims, such as allegations that Muslims sell their drugs without police intervention and that they throw rocks at cars belonging to "whites." On 26 October 2011, L.T. wrote to the Nîmes public prosecutor to lodge a criminal complaint against Mr Sanchez and the users who posted the offending comments. A day later, Sanchez posted a message on the wall of his Facebook account inviting users to "monitor the content of [their] comments," but did not remove already posted comments. Sanchez appealed, but the Court of Appeal upheld the first instance verdict (but lowered the fine by 1,000 EUR).

The ECtHR held that the comments were 'clearly unlawful.'[68] The language used by users was found by the Court to have 'clearly encouraged incitement to hatred and violence against a person because of their belonging to a religion.'[69]

Although it was not Sanchez who made the impugned comments, but rather other Facebook users, the ECtHR observed that by making his Facebook wall publicly accessible, Sanchez effectively took responsibility for the comments posted there. It concurred with the findings of the French courts, which held that Sanchez had failed to remove the comments for a period of six weeks and was, therefore, liable as the producer of an online public communication platform, making him the main offender. It is noteworthy that Sanchez had published a message urging users to be mindful of their comment content. In this context, the ECtHR applied the principles established in Delfi AS v. Estonia, a case concerning the liability of a news portal for third-party comments, to an individual user. It framed this obligation within the scope of the applicant's responsibilities as a political candidate. The Grand Chamber also acknowledged the distinction between Delfi, which involved a professional news outlet, and Sanchez, where the defendant was an individual, albeit one engaged in political life, saying that the 'situation cannot be compared to that of an Internet news portal.'[70] Nevertheless, it found 'no reason to hold otherwise in the present case.'[71] It extrapolated (slightly) on this by referring to the fact that the 'duties and responsibilities' clause of Article 10 is to be 'attributed to politicians when they decide to use social media for political purposes...The applicant was not merely a private individual.' [72]

---

[68] Sanchez v France, App. No 45581/15, (ECtHR 2nd December 2021) Para.140

[69] Ibid. Para.189

[70] Sanchez v France, App. No 45581/15, (ECtHR 2nd December 2021) Para.140; Sanchez v France, App. No 45581/15, (ECHR – 15th May 2023) Para.180

[71] Sanchez v France, App. No 45581/15 (ECtHR 15th May 2023) Para.140

[72] Sanchez v France, App. No 45581/15 (ECtHR 15th May 2023) Para.80

Extending removal obligations (with time limits) to individual users (albeit political persons) for comments made by third parties constitutes a substantially worrying turn for the freedom of expression. This is, nevertheless, the path that the ECtHR has opted to follow.

A 2021 empirical study analyzed the time taken for national legal processes in hate speech cases in Denmark, Germany, the United Kingdom, France and Austria. [73]It then compared these time frames with the significantly shorter time frames demanded of social media companies by legislation like Germany's NetzDG, requiring the evaluation and elimination of potentially illegal material within hours or a few days. The limitations of available data preclude the possibility of exact country-to-country comparisons, but the results nevertheless reveal an apparent imbalance: national courts take substantially longer to decide the illegality of hate speech than the tight time frames demanded of platforms. Where platforms are expected to take action in a few hours to a week, judicial processes in these nations run far longer than that. The typical time frames for state authorities are as follows:

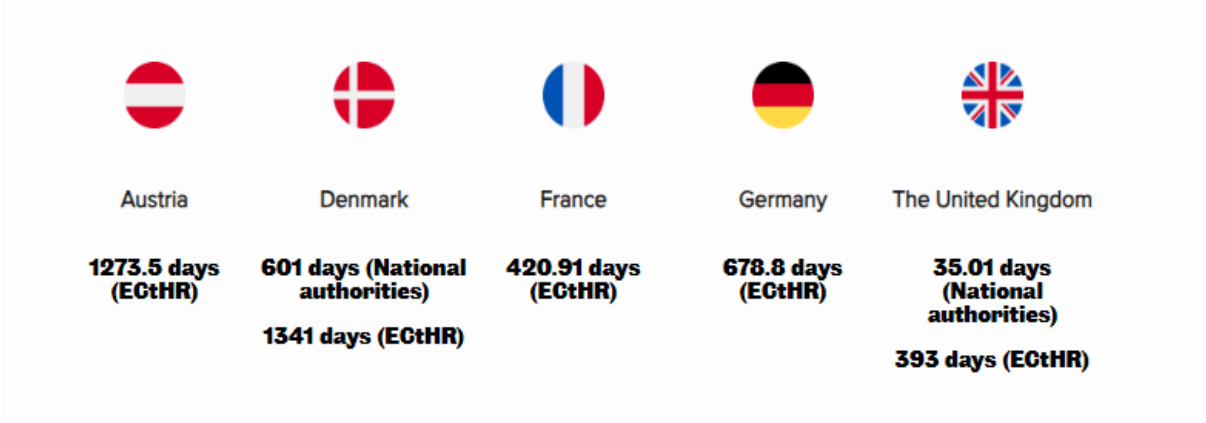

Figure 2. Time frames for hate speech removal (by country).

A compilation of ECtHR case- law in relation to the three member states of this report, illustrates an average period of 778.47 days from the date of offense to the conclusion of the first-instance proceedings for the prosecution of offenses related to speech. A period as lengthy as this differs significantly from expedited procedures mandated by EU regulatory instruments. For example, Germany's NetzDG requires the removal of manifestly unlawful content within 24 hours and other unlawful content within seven days. Similarly, the EU's Digital Services Act (DSA) includes provisions aimed at preventing undue delays in content moderation and removal processes. This disparity highlights a significant tension between the rapid response mechanisms imposed on digital platforms and the inherently slower pace of traditional judicial proceedings. It raises important questions about the feasibility of aligning

[73] Jacob Mchangama, Natalie Alkiviadou and Raghav Mendiratta, 'Rushing to Judgment: Are Short Mandatory Takedown Limits for Online Hate Speech Compatible with The Freedom of Expression?' (Justitia 2021) <https://futurefreespeech.org/wp-content/uploads/2021/01/FFS_Rushing-to-Judgment-3.pdf> [Accessed 2 September 2025]

these two regulatory spheres while ensuring procedural fairness and safeguarding against overreach.

Grappling with the Harms of Hate Moderation

Hate Speech and Harm: No Consensus

This paper argues that hate speech which meets the criteria under Article 20(2) of the ICCPR and its accompanying Rabat Plan of Action, both of which are discussed further down, should be moderated. The rest should be left online. To this end, particular types of hate speech, especially those involving calls to violence, should be subject to removal. However, there is a body of studies that suggests exposure to hate speech online can lead to fear, psychological trauma and self-censorship, especially among minority communities.[74] For example, Matsuda argues that hate speech erodes the victim's sense of identity and self-worth, while also diminishing their perception of personal safety.[75] In fact, a 2024 study found that victims of online hate speech had a heightened concern regarding their sense of security.[76] Lawrence notes the 'injurious impact' of direct racial insults, equating them to a 'slap in the face' that prevents opportunities for response, often resulting in silence or retreat due to the fear, anger, and shock they provoke.[77] Some research has found that hate speech may lead to psychological effects on victims,[78] which vary according to exposure and endurance[79] but also increased hate crimes.[80]

---

[74] See, for example, Richard Delgado & Jean Stefancic, *'Must We Defend Nazis?'* (1st edn. New York University Press, New York 2018); Eric Barendt, *'Hate Speech: Lecture given at Hull'* (November 21, 2013): <www2.hull.ac.uk/fass/pdf/Eric%20Barendt-HATE%20SPEECH.pdf> [Accessed 1 May 2025]
As she discusses in Mari J Matsuda et al., '*Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment'* (1st edn, Westview Press, Boulder CO 1993) and Mari J Matsuda, 'Public Response to Racist Speech: Considering the Victim's Story' (1989) 87 *Michigan Law Review* 8

[75] As she discusses in Mari J Matsuda et al., '*Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment'* (1st edn, Westview Press, Boulder CO 1993) and Mari J Matsuda, 'Public Response to Racist Speech: Considering the Victim's Story' (1989) 87 *Michigan Law Review* 8

[76] Arne Dreißigacker et al., 'Online Hate Speech Victimization: Consequences for Victims' Feelings of Insecurity' (2024) 13 *Crime Science* 4

[77] Charles R Lawrence III, 'If He Hollers Let Him Go: Regulating Racist Speech on Campus' (190) 40 *Duke Law Journal*, p. 452

[78] Arne Dreißigacker et al., 'Online Hate Speech Victimization: Consequences for Victims' Feelings of Insecurity' (2024) 13 *Crime Science* 4

[79] Koustuv Saha, Eshwar Chandrasekharan and Munmun De Choudhury, 'Prevalence and Psychological Effects of Hateful Speech in Online College Communities' (2019) *Proceedings of the 10th ACM Conference on Web Science*

[80] Karsten Müller & Carlo Schwarz, 'Making America Hate Again? Twitter and Hate Crime under Trump' (2021) *American Economic Review Papers and Proceedings*, Karsten Müller, Karsten & Carlo Schwarz, 'Fanning the Flames of Hate: Social Media and Hate Crime' (2020) <https://ssrn.com/abstract=3082972> [Accessed 3 May 2025]

It is important to underline that 'the exact individual impact of speech varies depending on content.'[81] This paper does not argue that non-violent hate speech has no consequences on its victims, bystanders, perpetrators, and society more broadly. It does, however, take the position that restricting freedom of expression in such cases is not necessarily the remedy. While any form of hate speech is deplorable for victims and society more broadly, it does not necessarily follow that restrictions on free speech are an effective way to tackle such speech.[82] As such, a response to hate speech must carefully balance the need to address alleged harm, considering the context in which the harm is taking place and the type of harm in question. At the same time, such a response may be dangerous. Empirical research shows that deplatforming those who spread hate discourse could lead them to move to unregulated platforms, thus promoting even more extreme discourse.[83]

For instance, one study examined the network of far-right actors on Telegram, a platform characterized by a high degree of decentralisation. The analysis revealed that Telegram's 'explosive growth coincides in time with the mass bans of the far-right actors on mainstream social media platforms.' These findings imply that deplatforming may be ineffective, as the migration to Telegram facilitated the rapid reestablishment of these actors' networks and influence. Furthermore, the presence of less moderated spaces may foster more insular and radical echo chambers for exchange and discussion. By driving harmful ideologies underground, such restrictions can hinder law enforcement and counter-narrative initiatives in their efforts to monitor and address extremism. However, it must be underlined that the exact individual impact of speech varies depending on content.'[84] Studies have also shown that repression of freedoms, such as that of freedom of expression, may contribute to increased physical violence[85] and cause social unrest.[86]

It is important to underline that 'the exact individual impact of speech varies depending on content.'[87]Although research has demonstrated substantial correlations between hate speech and real-world harm, the evidence is not without its limitations and has been subject to

[81] Lyn K. L. Tjon Soei Len & Anniek de Ruijter, 'Conceptualising the Tortuous Harms of Sexist and Racist Hate Speech' (2023) 2 *European Law Open,* p.23

[82] Jacob Mchangama & Natalie Alkiviadou, 'Hate Speech and the European Court of Human Rights: Whatever happened to the Right to Offend, Shock or Disturb?' (2021) 21 *Humam Rights Law Review* 4

[83] Aleksandra Urman & Stefan Katz, 'What They Do in the Shadows: Examining the Far-Right Networks on Telegram' (2020) 7 *Information, Communication & Society*

[84] Lyn K. L. Tjon Soei Len & Anniek de Ruijter, 'Conceptualising the Tortuous Harms of Sexist and Racist Hate Speech' (2023) 2 *European Law Open,* p.23

[85] Jacob Aasland Ravndal, 'Explaining Right-Wing Terrorism and Violence in Western Europe: Grievances, Opportunities and Polarisation' (2017) 57 *European Journal of Political Research* 4

[86] Christian Bjørnskov & Jacob Mchangama, *'Freedom of Expression and Social Conflict'* (2023) IFN Working Paper No.1473

[87] Lyn K. L. Tjon Soei Len & Anniek de Ruijter, 'Conceptualising the Tortuous Harms of Sexist and Racist Hate Speech' (2023) 2 *European Law Open,* p.23

counterarguments. Dworkin underlines the 'malign, chilling force'[88] of hate speech, but argues that claims about its harms are often 'inflated and some are absurd.'[89] In Heinze's Long Standing and Prosperous Democracies (LSPDs), Heinze argues that 'despite decades of pro-ban law and policy... no empirical evidence has, in any statistically standard way, traced hatred expressed within general public discourse to specifically harmful effects.'[90]

International Human Rights Law as a Safe Framework

The above analysis has demonstrated that private, profit-making companies, not governed by IHRL, minus some non-binding obligations in the form of guidelines for business,[91] are today's judges of the limits of the fundamental right of free speech. National legislation such as the German NetzDG and, even more worryingly, pan-EU legislation in the form of the DSA, have placed platforms in a position of immense power to decide on what speech should be allowed in the public sphere. As well as granting them this power, even if the DSA comes, at least for VLOPs, enhanced audit and transparency requirements, the DSA and previously the NetzDG (as a national example) has also put them in a position of fear of fines and disruption of their profit and business plan.

This combination of factors is destructive, not only for the freedom of expression, but also for the right to non-discrimination. Automated content moderation mechanisms are just not apt for dealing with the nuances and context-specific analysis that the analysis of hate speech requires and could even be affected by biased datasets. This has led to the silencing of already marginalized groups. Additionally, the analysis demonstrates that there is no consensus on how to deal with hate speech. Some studies link hate to psychological harm, fear and self-censorship, particularly among minority groups, whereas other work finds no link between hate speech and actual harm in LSPDs (Heinze). At the same time, research highlights that responses like deplatforming can drive extremists to less regulated spaces, potentially increasing radicalization and complicating efforts to monitor harmful content.

In light of the above, Article 20(2) of the ICCPR and its accompanying RPA become particularly important. Together, they offer a carefully calibrated legal and policy framework for addressing hate speech. This framework establishes a high threshold for limiting speech, focusing on speech that intentionally incites discrimination, hostility or violence, thereby ensuring that any restrictions are legitimate, necessary and proportionate. It emphasizes the protection of vulnerable groups while safeguarding freedom of expression by avoiding overly

---

[88] Ronald Dworkin, 'A New Map of Censorship' (1994) *Index on Censorship*, p.10 <https://journals.sagepub.com/doi/pdf/10.1080/03064229408535633> [Accessed 1 May 2025]

[89] Ibid. 12

[90] Eric Heinze, 'Hate Speech and Democratic Citizenship' (1st edn. Oxford University Press 2016) pp.126-127

[91] UN Guiding Principles on Business and Human Rights <https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf> [Accessed 20 April 2025]

broad or arbitrary censorship. Following this framework helps ensure that responses to hate speech maintain democratic legitimacy, respect international human rights standards and effectively target only the most harmful forms of expression, without driving dangerous ideas underground or exacerbating societal divisions.

Article 20(2) of the ICCPR provides that any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law. The Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred (RPA), which was adopted in 2012 by a group of experts appointed by the UN, underlines that, in order for Article 20(2) to be applicable, there must be a high threshold.[92] The RPA clarifies that to assess the severity of hatred and determine whether the high threshold is met, factors to be considered include 'the cruelty of what is said or the harm advocated and the frequency, amount, and extent of the communications.'[93] This is further supported by the 2012 Report of the UN's Special Rapporteur on Freedom of Opinion and Expression, which asserts that 'the threshold of the types of expression that would fall under the provisions of Article 20(2) should be high and solid.'[94] Article 20's dual character, which encompasses the ban of war propaganda in part , and the prohibition of advocating for national, racial, or religious hatred in part 2, indicates that it pertains to especially egregious forms of expression. In addition, the RPA sets out a six-part threshold test that could guide the moderation of online hate speech covering the social and political context, status of the speaker, intent to incite the audience against a target group, content and form of the speech, extent of its dissemination and likelihood of harm, including imminence. A 2021 report by the Future of Free Speech sets out a framework through which IHRL can be used for the moderation of online hate speech.[95] Importantly, IHRL is also the framework through which the Oversight Board bases its decisions.[96] The use of IHRL in the sphere of hate speech is significant as it would allow for a global set of rules that protect the right to freedom of expression and the principle of non-discrimination to infiltrate into the current situation where private companies are having to decipher, without any clear benchmark, what speech

[92] Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred that constitutes Incitement to Discrimination, Hostility or Violence (2002) <https://www.ohchr.org/sites/default/files/Rabat_draft_outcome.pdf> para.22 [Accessed 30 April 2025]

[93] Ibid. para. 22

[94] Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2012) A/67/357, para.45 <https://documents.un.org/doc/undoc/gen/n12/501/25/pdf/n1250125.pdf> [Accessed 10 May 2025]

[95] Jacob Mchangama, Natalie Alkiviadou & Raghav Mendiratta, 'A Framework of First Reference: Decoding a Human Rights Approach to Content Moderation in the Era of "Platformization"' (2021) *Justitia* <https://futurefreespeech.org/wp-content/uploads/2021/11/Report_A-framework-of-first-reference.pdf> [Accessed 15 May 2025]

[96] Oversight Board – Case Decisions and Policy Advisory Opinions <https://www.oversightboard.com/decision/> [Accessed 20 May 2025]

should and should not be removed. This could be the first step in overcoming the risk that companies are removing 'legal but controversial speech.'[97]

Conclusion

This paper has critically examined the European approach to moderating online hate speech, arguing that the current model, anchored in legislation such as Germany's NetzDG and the EU's Digital Services Act, risks undermining fundamental rights, particularly the freedom of expression and equality. By outsourcing speech regulation to private companies that are neither democratically accountable nor bound by IHRL, Europe has created a system where profit-driven entities wield unprecedented power over what speech remains online. The pressure for rapid removals, coupled with the threat of punitive fines, fosters over-censorship, reliance on flawed AI moderation tools, and a chilling effect on lawful speech, especially for historically marginalized groups. The lack of definitional clarity surrounding hate speech further compounds these issues. While efforts to curtail genuinely harmful expression are essential, the inconsistent and often vague legal and platform-specific standards render moderation efforts arbitrary, opaque, and susceptible to abuse. The risk of unintended consequences, such as the migration of extremist actors to unregulated platforms, the formation of echo chambers, and the legitimization of censorship practices in authoritarian regimes, illustrates the dangerous ripple effects of well-meaning but poorly calibrated policy. In light of these challenges, the paper advocates for anchoring hate speech regulation within the robust, rights-respecting framework of IHRL. Article 20(2) of the ICCPR and the RPA provide a carefully balanced, high-threshold test for identifying and addressing the most dangerous forms of hate speech, without stifling democratic discourse. Adopting this framework as the global benchmark for online speech moderation would help ensure that responses to hate speech are legitimate, proportionate and effective, while safeguarding freedom of expression, non-discrimination and the democratic integrity of the digital public sphere.

[97] Rikke Frank Jørgensen & Lumi Zuleta, 'Private Governance of Freedom of Expression on Social Media Platforms' (2020) *Nordicom Review* 41, pg.53