CELE
Centro de Estudios en Libertad de Expresión

Research Paper

70

Legal Approaches to Hate Speech Within the Internet Ecosystem: Speed, Scope, and Scale

Joan Barata

September 2025

Joan Barata, "Legal Approaches to Hate Speech Within the Internet Ecosystem: Speed, Scope, and Scale", Artículo de investigación No. 70 (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

Legal Approaches to Hate Speech Within the Internet Ecosystem: Speed, Scope, and Scale

Joan Barata

Visiting Professor
Oporto Law School. Portuguese Catholic University
jbaratamir@gmail.com

September 2025

Introduction. The evolving definition of the notion of hate speech

It is a basic principle of the international human rights system that freedom of expression applies both online and offline. The UN Human Rights Council declared in its resolution 32/13 of 1 July 2016 that "(...) the same rights that people have offline must also be protected online, in particular freedom of expression, which is applicable regardless of frontiers and through any media of one's choice, in accordance with articles 19 of the UDHR and ICCPR." In doing so, it recalled its resolutions 20/8 of 5 July 2012 and 26/13 of 26 June 2014, on the subject of the promotion, protection and enjoyment of human rights on the Internet.

The right to freedom of expression and freedom of information is a universal human right enshrined in article 19 of the Universal Declaration of Human Rights (UDHR)[1] and the International Covenant on Civil and Political Rights (ICCPR)[2]. Article 19 ICCPR thus constitutes the basic pillar to understand how and to what extent the mentioned rights are protected within the universal human rights system. Besides the general regime established in article 19 ICCPR (introducing a general protection and restrictive exceptions), the international human rights system also incorporates a specific provision that clearly establishes obligations for States to forbid certain categories of speech. According to article 20 ICCPR, any propaganda for war shall be prohibited by law. Also, as established in its second

1 Available online at: https://www.un.org/en/universal-declaration-human-rights/

2 Available online at: https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx

paragraph, "(a)ny advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law."

Considering that article 20.2 ICCPR contains a broad definition of hate speech, it is the responsibility of the national legislator, as well as national judicial operators, to make a proper assessment of each piece of content based on principles, rules, and conditions established in international law. In particular, it is important to note the threshold test on hate speech extracted from the Rabat Plan of Action (RPA), which permits assessing if a particular statement reaches the level of actual incitement to discrimination, hostility, or violence. The Rabat framework test lays out six parameters to check if a statement may amount to a criminal offence. On a case-by-case basis, the test requires looking into the context, speaker, intent, content, extent of the speech, and likelihood of harm[3]. The Plan is not a legally binding instrument. However, it contains a series of relevant conclusions and recommendations formulated as the result of a series of expert discussions organized by the office of the UN High Commissioner for Human Rights. These standards shall, in principle, help legislators, policymakers, as well as the judiciary in the process of defining and qualifying specific hate speech cases.

In terms of definitions, and in addition to the above, the UN Strategy and Plan of Action on Hate Speech[4] defines hate speech as "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference 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 identity factor."

A slightly more detailed definition can also be found in the Council of Europe human rights system. The Additional Protocol to the Convention on Cybercrime, which seeks to prohibit "racist and xenophobic material" on the Internet, defines hate speech as "any written material, any image or any other representation of ideas or theories, which advocates, promotes or incites hatred, discrimination or violence, against any individual or group of individuals, based on race, color, descent or national or ethnic origin, as well as religion if used as a pretext for any of these factors". It is worth noting that the Additional Protocol also seeks to outlaw the "denial, gross minimization, approval or justification of genocide or crimes against humanity".

3 https://www.ohchr.org/EN/NewsEvents/Pages/Hate-speech-threshold-test.aspx

4 Available online at: https://www.un.org/en/hate-speech/un-strategy-and-plan-of-action-on-hate-speech

In any case, it is very important not to dismiss the fact that, as clearly stated in the Rabat Plan of Action, article 20 ICCPR requires a high threshold because, as a matter of fundamental principle, "limitation of speech must remain an exception" and that the three-part test (legality, proportionality and necessity) for restrictions also applies to cases involving incitement to hatred. In addition to this, the inclusion of article 20 in the agreed text of the ICCPR as a provision with additional and more specific restrictions to freedom of expression was highly controversial and was preceded by protracted and heated negotiations, particularly between liberal democratic and communist countries (Mchangama 2011).

The debate over how to define hate speech, particularly in the online world, is far from settled. Definitions from international documents are still vague and ambiguous, whereas States around the world have been adopting in their respective legal systems diverse definitions of hate speech, quite often qualified as a criminal act subject to prosecution (Alkiviadou et al. 2020). In this context, some States may have stepped beyond the thresholds established under international human rights law by using this concept as a way of categorizing ideas that the majority or those in power may find objectionable or to suppress opinions that they object to. For example, new legal provisions following the public Quran burnings in Denmark and Sweden have sparked strong criticism among freedom of expression experts and advocates who have accused these countries of sacrificing such principle when faced with threats or adverse consequences of unpopular or extremist speech[5].

The emergence of the digital public sphere and the predominant role of online platforms in this field -particularly the so-called social media platforms- has made the exploration and interpretation of hate speech norms far more complex. As further explained somewhere else (Barata 2024), online platforms have their own content governance or moderation policies, which are the consequence of an amalgam of factors including their own civility principles and values, business models, reputational constraints, investors and advertisers' pressures, as well as direct and indirect influence from relevant authorities (Keller 2019). These policies are different, in terms of formulation and enforcement, from general content restrictions established by international and national legal systems and generally applicable both offline and online. There are, however, connections between the two governance mechanisms that create inevitable and sometimes complex interactions. For instance, platforms' terms of service (ToS) and community standards include references to undesirable content in areas

5 https://time.com/6302649/denmark-swedens-quran-burnings-commitment-to-free-speech/

such as violence and incitement, dangerous organizations and individuals, inauthentic behavior, disinformation, child safety, violent content, or sale of illegal products, which might totally or partially overlap with existing statutory rules in these areas in different countries. This is also the case with hate speech.

While it is obvious that public authorities can only order companies to eliminate illegal content, it is also clear that platforms are "free" to remove harmful content on the grounds that it violates their ToS. In reference to this matter, the European Union Agency for Fundamental Rights (FRA) has admitted that a large amount of hateful content present on social media may be considered to be "on the fringes of legality". However, legal content connected to hateful attitudes -such as certain forms of negative stereotyping, hurtful, derogatory or obscene language that offends certain individuals or groups, denigration, or dissemination of hateful political ideologies- "can also create an atmosphere of hate that may prevent people from joining online conversations". This may also have an effect on other users who "feel encouraged to express hate, including illegal hate speech". Therefore, when it comes to tackling hate speech online, this concept has become, from a normative perspective, an "umbrella term" with multiple meanings, thus including expressions of hatred that are not illegal under national or international law. (FRA 2023).

When it comes to illegal hate speech online stricto sensu, it is important to note that within the context of the European Union and as part of the legal apparatus for the enforcement of the Digital Services Act (DSA)[6], the revised Code of conduct on countering illegal hate speech online + (the 'Code of conduct+')[7] was formally endorsed by European institutions in January 2025. This Code builds on the strictly voluntary Code of Conduct originally adopted in 2016 and aims to establish the way online platforms shall deal with content deemed illegal hate speech according to EU and national law. It also facilitates compliance with and the effective enforcement of the DSA in this specific area, which, as it will be shown later, contains a series of obligations for platforms with a systemic perspective[8]. It is also important to note that co-regulatory instruments such as this Code have the effect -or at least

6 Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC.

7 https://digital-strategy.ec.europa.eu/en/library/code-conduct-countering-illegal-hate-speech-online

8 The Code of conduct+ was signed and submitted for integration under the DSA by important platforms such as Dailymotion, Facebook, Instagram, Jeuxvideo.com, LinkedIn, Microsoft-hosted consumer services, Snapchat, Rakuten Viber, TikTok, Twitch, X and YouTube.

the objective- of establishing a certain degree of consistency and integration between legal speech limitations and private community standards in this area.

Beyond legal and normative discussions, scholarship has also provided a variety of definitions of hate speech, basically around the idea of bias-motivated, hostile, and malicious language targeted at a person or group because of their actual or perceived innate characteristics. However, academics tend to pay more attention to the causes, harms, and responses to hate speech than to providing a properly defined category (Siegel 2020). In addition to this, traditional approaches to online hate from disciplines such as communication or psychology often focus on perpetrators' traits and their attitudes toward their targets. Such approaches often fail at properly tackling the social and interpersonal dynamics that social media may reinforce and through which individuals glean social approval from like-minded friends (Walther 2022).

The impact of hate speech

The Internet has made it easier for like-minded individuals on the margins to communicate and collaborate. The development of online spaces of socialization that normalize exposure and engagement with extreme content through repetition can support polarisation efforts and the spread of hate speech online (Paillé, Galley, Thue, and Wilkinson 2021). On the other hand, the architecture of social media platforms may favor content that provokes emotional engagement (outrage, amusement, or fear) regardless of its ideological content. This incentive structure would therefore facilitate the viral spread of harmful narratives, particularly when they are not directly formulated as explicit hateful expressions and use ambiguous language and formats, including certain forms of humor or satire (Gillespie, 2018).

Scholars have also indicated that, in any case, more work is needed in order to properly understand and assess intent, contextual nuances, and potential impact of hate speech (Lee and Gililland, 2024). Scholars such as Heinze (2016) have particularly argued that the freedom to express our views, including views that would qualify as hate speech, must be safeguarded not only as an individual right, but as an essential attribute of "democratic citizenship". This demand would, however, basically apply to what Heinze calls "Longstanding, Stable, and Prosperous Democracies" (LSPDs), in opposition to fragile democracies, semi-authoritarian regimes, and non-democratic states. LSPDs would therefore be equipped with sufficient legal, institutional, educational, and material resources to admit

all views into public discourse, while also being capable of protecting vulnerable groups from violence or discrimination.

It is also important to highlight that different international organizations have been generally warning about the potential impact of hate speech, particularly in the online world. The United Nations organization has insisted on the fact that historical precedents show that hate speech can be a precursor to atrocity crimes[9], whereas its agency UNESCO has established that "(h)ate speech not only causes harm at the personal level and can incite violence, it is an attack on inclusion, diversity and human rights. It undermines social cohesion and erodes shared values, setting back peace, stability, sustainable development and the fulfillment of human rights for all"[10].

In any case, properly assessing the impact of hate speech requires to consider a wide series of diverse and potentially affected areas, including issues of mental health, self-esteem, and feeling of insecurity of targeted individuals and groups, decrease in empathy and sensitization from the side of majority groups exposed to this type of content, and increased prejudice, distrust, exclusion, violence and erosion of democratic values in societies in general (Siegel 2020).

This variety of possible harms is particularly connected with the highly context-specific and evolving nature of online content, where looking into hate speech requires the adoption of a statistical approach, as explained by Yale Law School professor Robert Post:

> "The scale of the internet produces forms of harm that may best be characterized as stochastic. Previously we asked whether particular speech acts might cause particular harms. The internet has rendered this kind of question almost obsolete. Speech that is simultaneously distributed to billions of persons may produce harm in ways that cannot meaningfully be conceptualized through the lens of discreet causality. We will need instead to think in terms of statistical probability of harm. Yet at present we lack any legal framework capable of assessing such stochastic harms in ways that will not drastically over-regulate speech."[11]

---

9 https://www.un.org/en/hate-speech/understanding-hate-speech/hate-speech-and-real-harm

10 https://www.unesco.org/en/countering-hate-speech/need-know#:~:text=It%20undermines%20social%20cohesion%20and,of%20human%20rights%20for%20all.

11 Quoted by Agustina del Campo in "Volume, Speed, and Accessibility as Autonomous Harms: Can Modern Legal Systems Deal With Harmful but Legal Content?", as part of a series of papers published by Carnegie Endowment for International Peace (edited by Steven Feldstein) and available online at:

The next section of this paper will precisely explore to what extent existing national and international legal instruments and human rights standards can be interpreted and adapted to tackle the harms produced by different forms of online hate speech. It will also be explored whether new approaches combining existing norms with new forms of self and co-regulation may provide possible human rights-aligned solutions to the mentioned challenges. In particular, it will be considered how the scheme introduced by the DSA establishing the need to assess the risk of dissemination of certain types of illegal content, including hate speech, from a systemic perspective and through the subsequent adoption of risk mitigation strategies has become a possible answer, yet with relevant uncertainties, to the demand for new regulatory approaches.

It must be clarified that the notion of hate speech that will be considered in this paper will, in any case, be based on the already mentioned general standards established under international human rights law. This means focusing on hateful expressions that incite discrimination, hostility, or violence against vulnerable individuals or groups on the basis of a series of protected characteristics. This paper will therefore not directly consider broader notions of hateful speech only contemplated by online platforms' community standards, or modalities of especially harmful online illegal content according to other types of legislation, such as criminal threats, propaganda for war, incitement to terrorism, or violent extremism.

Illegal hate speech in light of speed, scope, and scale

Hate speech and social climate harms in the case law of the European Court of Human Rights (ECtHR)

The ECtHR had, in the course of the recent decades, the opportunity to assess in light of article 10 of the European Convention on Human Rights (ECHR) the legitimacy of a long series of national court decisions where individuals had their freedom of expression restricted on grounds of hate speech dissemination.

Mchangama and Alkiviadou have elaborated a thorough analysis (2021) of the case law of the Court, highlighting two relevant conclusions for the purposes of the present paper.

Firstly, the Court does not appear to have clear criteria as to when to merely use the standards and limits incorporated in article 10 ECHR or to rather refer to article 17 ECHR

https://carnegieendowment.org/research/2023/11/new-digital-dilemmas-resisting-autocrats-navigating-geopolitics-confronting-platforms?lang=en#volume-speed-and-accessibility-as-autonomous-harms-can-modern-legal-systems-deal-with-harmful-but-legal-content

(prohibition "to engage in any activity or perform any act aimed at the destruction of any of the rights and freedoms set forth herein or at their limitation to a greater extent than is provided for in the Convention"). Even though the second provision should only apply to very exceptional cases of illegitimate abuse of the right to freedom of expression by causing a serious harm or danger to the rights and values of the Convention, the Court seems to have failed, so far, at establishing a clear and foreseeable threshold to differentiate between the two categories of cases.

Secondly and more importantly, at least in some cases, the Court has held that hate speech and the advocacy of totalitarian ideologies may be banned even without any demonstrable danger of violence or other criminal actions. This has taken the Court to allow for the prohibition of speech which is far from a call for violence or hatred when such expression is targeting protected characteristics such as ethnicity, religion, or sexual orientation. The mentioned authors have expressed their surprise at the fact that the Court has never required that States demonstrate that the prohibition of hate speech is the most efficient instrument towards securing tolerant societies and countering hatred and social cohesion. In their opinion, these standards may have serious negative ramifications when applied to specific frameworks, particularly the regulation of social media content.

To mention just an illustrative example, in Norwood v. The United Kingdom[12] the Court declares inadmissible an application in relation to the criminal charges imposed on an individual for publicly displaying a photograph of New York's Twin Towers in flames, with the words "Islam out of Britain – Protect the British People" and a symbol of a crescent and star in a prohibition sign. In this case, the Court fully agrees with the assessment made by local authorities by stating that the words and images on the poster amounted to "a public expression of attack on all Muslims in the United Kingdom" and that "(s)uch a general, vehement attack against a religious group, linking the group as a whole with a grave act of terrorism, is incompatible with the values proclaimed and guaranteed by the Convention, notably tolerance, social peace and non-discrimination". In a relatively similar manner, in the more recent decision of Sanchez v. France[13] the Court recognizes the right of political parties and their representatives "to defend their opinions in public, even if some may offend, shock or disturb part of the population", who can therefore propose solutions to the problems

12 Admissibility decision of 16 November 2004. Application no. 23131/03.

13 Decision of 15 May 2023. Application no. 45581/15.

linked to immigration. However, in doing so they must avoid advocating racial discrimination and "resorting to vexatious or humiliating remarks or attitudes", since they "might trigger reactions among the public that would be detrimental to a peaceful social climate and might undermine confidence in the democratic institutions".

In light of all the above, we can appreciate how the ECtHR has separated itself, in some of its judgements, from the strict standards established by international law and relevant recommendations such as the RPA in terms not only of the three-part test but also the specific requirements in terms of incitement against a target group, direct causation and likelihood of harm, and intent. The Court seems thus to open the door to a laxer interpretation of the scope of illegal hate speech by referring to vague societal risks and *ambient hate* beyond strict advocacy ("vexatious or humiliating remarks or attitudes", triggering reactions "detrimental to a peaceful social climate", etc.).

It must be noted that, in principle, the decisions of the Court have so far considered the mentioned *ambient hate* in terms of impact but are still focused on specific and identifiable pieces of content. An important step forward in terms of legal interpretation would be to analyze the risks of a social climate of hate by considering the behavior of groups or the creation of shared narratives beyond individual and specific pieces of content. Could such a perspective be incorporated into our legal systems and jurisprudence without imposing unnecessary, disproportionate, and even unforeseeable restrictions on the right to freedom of expression?

Approaches to systemic hate speech from the perspective of content moderation

As already mentioned, the dissemination of hate speech online has become systemic and networked (Benkler et al., 2018). Online platforms must therefore deal with this type of harmful content at scale, in a speedy manner, beyond borders, and assess different types of languages, formats, and contexts.

The online world may have the effect of encouraging the anonymous publication of expressions of disdain, frustration, and hate by individuals who otherwise and in different contexts would conduct themselves in a more civil manner. Independent from the actual intentions and impact of these individual attacks, it is undeniable that they may contribute to a climate of intolerance and division as part of wider online exchanges and conversations.

Some studies have, in fact, shown how generalized hate may proliferate more widely as it has the potential to appeal to a large number of users. In line with this, social media posts with generalized hate speech would be more viral than directed hate speech with specific mentions or tagging particular topics (Maarouf et al. 2023).

As it has already been explained as well, online platforms have incorporated into their ToS and community standards a series of limitations applicable to "hateful conduct" (Meta), "hate speech and behavior" (TikTok), "hate speech" (YouTube), or "hateful conduct" and "violent content" (X). These policies are to be seen as "inspired" by the legal notion of hate speech, even though they are elaborated and presented with a much higher level of detail (including exemplifications), and in general they would also cover expressions not necessarily banned by national legal systems or international standards. In fact, a 2023 report that assessed the hate speech policies of eight social media platforms found that there has been a significant expansion in the scope of hate speech policies on platforms over time, encompassing both the types of content and the protected characteristics (Mchangama et al. 2023). Moreover, a 2024 study by The Future of Free Speech, which looked at content removal practices in France, Germany, and Sweden, found that a substantial majority (87.5% to 99.7%) of deleted comments on Facebook and YouTube were legally permissible[14].

This evolution poses high pressure on platforms in terms of demanding a rethinking of content moderation strategies. Traditional approaches focused on keyword detection or isolated incidents are increasingly insufficient, and the use of AI tools for content moderation has become unavoidable. In this sense, platforms have been increasingly questioned and criticized by Governments and legislators as to their capacity to tackle harms derived from online hatred, while, at the same time, they also face the increasing risk of being considered instruments for censorship of controversial and edgy political speech. In this last sense, it is important to highlight the changes in Meta's content moderation policies announced in January 2025. Among other things, the company committed to "allowing more speech" in relation to topics in many cases potentially connected to what had been so far considered as "hateful conducts":

> "We're getting rid of a number of restrictions on topics like immigration, gender identity and gender that are the subject of frequent political discourse and debate. It's

14 https://futurefreespeech.org/preventing-torrents-of-hate-or-stifling-free-expression-online/

not right that things can be said on TV or the floor of Congress, but not on our platforms."[15]

Beyond the specific and complex dynamics of the relationship between tech companies and the new Administration in the United States, the mentioned difficulties and dilemmas were already acknowledged a few years ago by the Special Rapporteur on freedom of opinion and freedom of expression in his report to the UN Human Rights Council of 9 October 2019[16]. He assumes that entities engaged in content moderation, such as big social media platforms, can "regulate" hate speech according to the scale, complexity, and long-term challenges that such a form of speech presents on these platforms. Therefore, restrictions could thus be imposed "even if it is not clearly linked to adverse outcomes (as hateful advocacy is connected to incitement in Article 20(2) of the ICCPR)". Important to add here that, in a previous report, the Rapporteur had directly addressed platforms requesting them to recognize that "the authoritative global standard for ensuring freedom of expression on their platforms is human rights law, not the varying laws of States or their own private interests, and they should re-evaluate their content standards accordingly"[17]. This must also be connected with the UN Guiding Principles on Business and Human Rights, adopted in 2011, which focus on the role of business enterprises as specialized organs of society performing specialized functions, required to comply with all applicable laws and to respect human rights[18].

From a legal point of view, we must therefore acknowledge a very complex public/private normative framework applicable to online hate speech, particularly when it comes to moderation efforts by platforms:

a) Platforms need to adhere in principle to the legal definition of hate speech in the different jurisdictions they operate and therefore comply with orders and requests from competent authorities in relation to this type of content.

b) Platforms will establish their own complementary policies regarding hateful or violent speech and conduct, to be applied, in principle, across the different markets where they provide services. They will moderate content accordingly, in most cases through a combination of human and automated systems. Also, in many cases,

15 https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

16 A/74/486. Available online at: https://www.ohchr.org/sites/default/files/Documents/Issues/Opinion/A_74_486.pdf

17 A/HRC/38/35. Available online at: https://www.ohchr.org/en/documents/thematic-reports/ahrc3835-report-special-rapporteur-promotion-andprotection-right-freedom

18 https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf

platforms will need to deal with content that is violative of their policies and illegal at the same time. When imposing restrictions or limitations, and depending of course on their own ToS, platforms will generally count on a diversity of tools, far beyond the traditional instruments at the reach of judges and other authorities when interpreting and enforcing legal provisions. In addition to this, content moderation will increasingly include interventions not based on individual pieces of content but on the behavior of groups of accounts and the actors or associations behind them. These instruments may, in principle, help to find solutions that better accommodate the needs of necessity and proportionality (Howard et al. 2024), as demanded by international human rights standards. However, the necessary use of AI tools to properly review the myriads of potentially hateful content present on an online platform has obviously triggered major concern, especially since, as mentioned already, "human factors" such as context, intertextuality, tone and many other systemic aspects have become increasingly relevant (Duarte and Llanso 2023). Also, it is very important not to forget that a human rights-oriented approach to content moderation that includes crucial considerations, such as contextual relevance, is an inherently time-consuming process (Alkiviadou 2024).

c) When moderating content, platforms are also supposed, and almost all of them have publicly committed themselves, to remain aligned with international human rights standards, particularly in the field of the protection of the right to freedom of expression. Also, when it comes to such commitment, in the European Union the DSA (article 14.4) has established that when imposing restrictions based on their ToS platforms must act "with due regard to the rights and legitimate interests of all parties involved, including the fundamental rights of the recipients of the service, such as the freedom of expression, freedom and pluralism of the media, and other fundamental rights and freedoms" as enshrined in the Charter of Fundamental Rights of the EU.

In relation to what has been explained in point b), it is necessary to refer to Douek's position (2022) in the sense that we need to move towards an understanding of content moderation as *systems thinking*. This means dedicating most of the resources to the upstream choices about design and prioritization in content moderation that set the boundaries within which downstream paradigm cases can occur, instead of focusing on individual cases and individual safeguards. A position that was nuanced by Klonick (2023), who responded that considering content moderation as *systems thinking* demands to accede that content moderation contains

a combination of systemic and individual case interventions, including "individual decisions, automations, governance, governments, external influence, internal politics, constitutions, norms, legality, human judgment and biases, administration, bureaucracy, multistep processes, long legislative-like meetings, people, corporate courthouses, actual courthouses, stakeholders, economies, the media, and iterative dynamic changes".

In addition to this, the private nature of social media platforms and their lack of truly binding obligations under international human rights law place freedom of expression at a dire risk. As such, the horizontalization of content moderation responsibility is problematic as it is being carried out with limited oversight and may have severe implications for freedom of expression (Alkiviadou 2024).

In relation to point c), it is worth noting the experience of Meta's Oversight Board as the self-regulatory independent body in charge of reviewing content moderation decisions affecting users of Facebook, Instagram, and Threats, precisely, from an international human rights perspective. According to its own website, the OSB was created to help "answer some of the most difficult questions around freedom of expression online: what to take down, what to leave up and why". Particularly, and according to section 2 of article 2 of the Board's Charter, "(w)hen reviewing decisions, the board will pay particular attention to the impact of removing content in light of human rights norms protecting free expression."

Particularly in relation to hate speech and connected to the areas of interest of this paper, there are a few decisions from the Board that deserve some attention, since they show the challenges associated with content moderation of networked hateful expressions.

The first one refers to the complex situation around the conflict in Palestine. Meta decided to maintain posts made in November 2023 that included the phrase "from the river to the sea. The posts were made after the October 7 terrorist attack on Israel by Hamas and reported by users. Users appealed Meta's action on the grounds that the posts violated Meta's policies on Hate Speech, Violence and Incitement, or Dangerous Organizations and Individuals. The final decision takes into consideration the submissions made by a significant number of organizations[19].

19 See, for example, the arguments presented by the Center for Democracy and Technology (https://cdt.org/wp-content/uploads/2024/05/Comments-to-FBOB-on-River-to-the-Sea-Bundle.pdf) and The Future of Free Speech (https://futurefreespeech.org/wp-content/uploads/2024/05/FINAL-OB-Submission-May-2024-5p.pdf).

While it is acknowledged that the phrase had been used by the terrorist group Hamas with explicit violent antisemitic eliminationist intent and actions, this mere fact does not make the phrase inherently hateful or violent. In this sense, the expression under scrutiny has long been used by various actors around the globe in political debates and protests related to the Israel-Palestine conflict, including the current tensions. The majority of the Board thus establishes that factors such as context and identification of specific risks need to be properly assessed in order to analyze content posted on Meta's platforms as a whole. Removing content containing the phrase in question could have been aligned with Meta's human rights responsibilities if it had been accompanied by statements or signals calling for exclusion or violence, or legitimizing hate. However, such a removal would in any case not be based on the phrase itself, but rather on other violating elements. Ultimately, because the phrase does not have a single meaning, "a blanket ban on content that includes the phrase, a default rule towards removal of such content, or even using it as a signal to trigger enforcement or review, would hinder protected political speech in unacceptable ways"[20].

The second decision refers to the publication in the Netherlands of caricatures of Black people in the form of blackface. The Board notes that Facebook has explicitly prohibited this type of content as part of its Hate Speech Community Standard. However, while the majority argued that such caricatures "are inextricably linked to negative and racist stereotypes and are considered by parts of Dutch society to sustain systemic racism in the Netherlands", yet a minority of the Board saw insufficient evidence to directly link this piece of content to the harm supposedly being reduced by removing it. Therefore, the Board did not only consider the specific circumstances of the case or even the intention of the users when posting this kind of images and references, but the systemic and cumulative effect of this kind of speech within the context of certain societies, concluding that "allowing such posts to accumulate on Facebook would help create a discriminatory environment for Black people that would be degrading and harassing"[21] (Barata 2022).

In a third relevant and very recent decision, and regarding two posts that include videos in which a transgender woman is confronted for using a women's bathroom and a transgender athlete wins a track race, the majority of the Board upheld Meta's decisions to leave up the content. The press release announcing the decision contains a note that reads:

20 https://www.oversightboard.com/news/new-decision-highlights-why-standalone-use-of-from-the-river-to-the-sea-should-not-lead-to-content-removal/

21 https://www.oversightboard.com/decision/fb-s6nrtdaj/

> "Meta's January revisions did not change the outcome in these cases, though the Board took the rules at the time of posting and the updates into account during deliberation. On the broader policy and enforcement changes hastily announced by Meta in January, the Board is concerned that Meta has not publicly shared what, if any, prior human rights due diligence it performed, in line with its commitments under the UN Guiding Principles on Business and Human Rights. It is vital Meta ensures adverse impacts on human rights globally are identified and prevented."[22]

This decision triggered what appears to be a sensitive division between a majority and a minority of the members of this body. The majority considered that these publications neither entailed a "direct attack" nor denied the existence of people based on their gender identity. In other words, the majority finds that in this case, there are no reasons that would determine the necessity of preventing harm to transgender people or avoiding a likely or imminent risk of incitement to violence. Moreover, for these Board Members, the athlete voluntarily chose to compete in a state-level athletics championship, in front of large crowds and attracting media attention, having already been the focus of such attention for earlier athletic participation. In this case, an established exception specifically referring to "voluntary public figures" would therefore apply in relation to the implementation of new policies on Bullying and Harassment.

What is particularly interesting is that while the majority of the Board focuses on the specific circumstances of the case, the minority considers that the posts meet the threshold of imminent risk of "discrimination, hostility or violence" against transgender people as a group, under international human rights law, which would therefore justify the removal of this content. These members of the Board point at the fact that the videos were posted against a backdrop of worsening violence and discrimination against LGBTQIA+ people, including in the United States, and that they do not only deliberately attack and misgender specific transgender individuals (and in one case affects the safety of a child) but also target transgender people as a group.

Beyond some allegations that "the Board is more than willing to bend under political pressure" (Kayyali 2025), what is obvious is that this case might be showing a "subtle shift" towards a more permissive stance on harmful or discriminatory expression (Tuovinen 2025). In other words, it seems that in the last case the Board would have established a much higher threshold when it comes to the potential harm deriving from attacks on individuals based on protected characteristics, in an apparent inconsistency with the approach used in other cases

22 https://www.oversightboard.com/decision/bun-1ynnk264/

where the societal impact of such expressions was given a higher degree of attention and consideration.

Hate speech as a systemic risk according to the DSA

The DSA establishes a series of fundamental rules and principles regarding, essentially, the way intermediaries participate in the publication and distribution of online content. In particular, it contains a series of provisions regarding the obligations and responsibilities of online platforms concerning illegal content.

The DSA maintains the general liability regime applicable to intermediary service providers at the EU level, already in place since 2000. This means that, as a basic principle, service providers are not liable for the information stored at the request of a user. In order to retain liability exemptions, platforms must not have actual knowledge of illegal activity or information, and/or not be aware of facts or circumstances from which the illegal activity or information is apparent, and upon obtaining such knowledge or awareness, act expeditiously to remove or to disable access to the illegal content. In connection with this, the DSA does not define or identify specific categories of illegal content online. According to article 3.h) illegal content means 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".

Legal liability exemptions do not preclude the possibility for third parties to request and for judicial and administrative authorities to issue valid orders and injunctions against service providers. Article 9 DSA contemplates orders "to act against one or more specific items of illegal content, issued by the relevant national judicial or administrative authorities, on the basis of the applicable Union law or national law in compliance with Union law". Article 16 DSA establishes the obligation for service providers to "put mechanisms in place to allow any individual or entity to notify them of the presence on their service of specific items of information that the individual or entity considers to be illegal content". Important also to note that according to article 7 DSA, intermediaries may not lose their liability protections "solely because they, in good faith and in a diligent manner, carry out voluntary own-initiative investigations into, or take other measures aimed at detecting, identifying and removing, or disabling access to, illegal content, or take the necessary measures to comply with the requirements of Union law and national law in compliance with Union law, including the requirements set out in this Regulation".

On the basis of all the above, we can therefore conclude that, regarding illegal hate speech as a modality of illegal content, platforms have a series of specific obligations on how to tackle this type of publication. However, it is also important to note that even though the Council Framework Decision 2008/913/JHA of 28 November 2008 on "combating certain forms and expressions of racism and xenophobia by means of criminal law" establishes a general legal framework regarding certain serious forms of hate speech, it will be the legislation of each of the members States that will provide the definition of this conduct, to be applied by online platforms in the respective jurisdiction.

A very important new element included in the DSA is the fact that online intermediaries are subject to a series of new diligence and due process obligations. These obligations vary depending on the nature and size of the service provider. Providers with more than 45 million users in the EU are considered very large online platforms (VLOPs) and very large search engines (VOSE), and are subject to the highest level of regulatory obligations and intervention.

In particular, VLOPs and VLOSEs are obliged to "diligently identify, analyse and assess any systemic risks in the Union stemming from the design or functioning of their service" (article 34) and subsequently "put in place reasonable, proportionate and effective mitigation measures" (article 35). Systemic risks would include, in terms of legality, "the dissemination of illegal content through their services" and "any actual or foreseeable negative effects for the exercise of fundamental rights". It shall be noted that at this point, there are no available sources as to the evaluation by the Commission of the first round of systemic risk assessment and mitigation reports or the results of independent audits conducted on the basis of article 37. In addition to this, there are still no guidelines or best practices provided by the Commission regarding the "reasonable, proportionate and effective mitigation measures, tailored to the specific systemic risks" that need to be put in place according to the DSA (article 35.3). This means that it still remains unknown how platforms will end up assessing matters such as the actual impact of the dissemination of illegal content and its "systemic" nature, the complex interplay between illegal content and the violation of ToS, cross-platform proliferation, or the differences in the ways that various types of illegal content are disseminated.

Another issue is whether platforms can be put in the position of making such complex analyses and deciding the best tools to deal with this very wide and diverse range of negative

effects, considering the very strong human rights implications that these tasks entail (Barata 2021). Since this is a unique and unprecedented exercise, there is still a lot of uncertainty regarding the implications, in practice, deriving from the enforcement of the mentioned provisions. In particular, there is no clarity regarding the consideration of specific safeguards to avoid excessive and inadequate restrictions to users' right to freedom of expression, besides vague and declaratory provisions such as article 14.4 DSA. As already mentioned, this norm indicates the need for platforms to pay due regard to "to the rights and legitimate interests of all parties involved, including the fundamental rights of the recipients of the service" when applying and enforcing their terms of service, such as freedom of expression, freedom and pluralism of the media, and other fundamental rights and freedoms as enshrined in the Charter.

It is also important to indicate that the DSA gives special significance to co-regulatory approaches to guide these processes (article 45). A previous section of this paper has already referred to the revised Code of conduct on countering illegal hate speech online + (the Code of conduct+), which was formally endorsed in January 2025.

Even though one of the main objectives of the Code is to facilitate a proper enforcement of the DSA, it shall be said that its nature, design, and provision raise significant concerns in relation to its impact on the right to freedom of expression. Among other matters, the Code of conduct + has the effect of "privatizing" hate speech definitions and enforcement via a co-regulatory instrument that in many ways is closer to a private contract between tech companies and the Commission than to an actual and foreseeable norm. The definition of hate speech is left to the "margin of appreciation" of companies' ToS on the basis of the wrong assumption that, from a legal perspective, there is a common and agreed-upon notion of hate speech applicable across the Union. In addition to this, broadly defined enforcement powers in the hands of online platforms have the effect of blurring the line between illegal and harmful (as in merely violative of community standards) hate speech thus incentivizing over removals while at the same time depriving users from defenses and safeguards generally applicable to cases where measures are adopted at the request of competent authorities. This situation is aggravated by the fact that there is a complete lack of properly articulated and meaningful protections to the right to freedom of expression beyond merely rhetorical references without a clear significance in terms of specific due diligence on the platforms' side.

Even though these caveats had been put forward by organizations such as Article 19[23] or relevant experts (Keller 2019) since the adoption of the very first version of the Code, these significant issues remain unaddressed within the context of the relevant relationship between the Code and the DSA.

It shall also be noted that hate speech may not only constitute a form of systemic risk as illegal content. Certain forms of legally hateful speech might also be considered as systemic risks under other and much broader risk categories, such as negative effects on civic discourse, electoral processes, public security, gender-based violence, public health, minors, or users' physical and mental well-being. This means that risk assessment and mitigation measures to be adopted by platforms according to the DSA will not only affect illegal forms of speech but also legal but harmful content, which is, of course, more problematic in terms of impact on freedom of expression.

The first round of risk assessments from 19 of the EU's largest internet platforms and search engines became available in late November 2024. Thousands of pages not so easy to digest and systematize which would still show the necessity to further define what constitutes a risk and when it becomes systemic, to further work on a consistent approach to these matters across different platforms as well as to clarify the scope of oversight role of the European Commission as the main regulatory body in this area (Sullivan 2025). These uncertainties also affect, in a particular manner, the dissemination of illegal hate speech, as well as other forms of hateful speech, susceptible to also be considered a systemic risk.

The example of the dissemination of online hate speech as a modality of systemic risk under the DSA has therefore shown what are probably the most relevant weaknesses of this important European norm. Such frailties include the poor definition of what constitutes a risk of systemic nature, the lack of guidance for platforms on how to understand and mitigate such risks on the basis of a proper consideration of the human rights implications of such tasks, allow for real transparency and accountability in these processes as well as consider the potential negative role of States in these areas (for instance, by establishing illegitimate restrictions to freedom of expression through overbroad definitions of what constitutes hate speech) (Del Campo et al. 2025). Meta's Oversight Board has also recently insisted on the importance of some of these matters via a paper on "Why freedom of expression must be the centerpiece of systemic risk assessments":

23 https://www.article19.org/data/files/medialibrary/38430/EU-Code-of-conduct-analysis-FINAL.pdf

> "The Board emphasizes that systemic risk assessments must include greater focus on respect for human rights, including freedom of expression, if they are to enhance meaningful platform accountability to users and improve content governance in line with the DSA's objectives."[24]

Conclusion

The regulation of hate speech in the conditions of speed, scope, and scale that characterise the online ecosystem represents one of the most complex challenges for platform legal and policy frameworks. As this paper has shown, the definition and scope of hate speech remain inconsistent across jurisdictions and institutional contexts, particularly when it comes to the interplay between international human rights standards, national legislation, and platform-level content policies. This uncertainty, coupled with the sheer scale and dynamics of online discourse, complicates both enforcement and the protection of fundamental rights, particularly freedom of expression.

While Article 20(2) of the ICCPR sets a high threshold for criminalizing hate speech, the practice of national authorities and courts, notably the European Court of Human Rights, has at times broadened this scope by incorporating notions such as "ambient hate" and social climate harms. These interpretations, although aimed at protecting vulnerable groups, risk diluting the rigorous incitement-based criteria established in the Rabat Plan of Action and may lead to disproportionate restrictions on lawful expression.

At the same time, online platforms have become central actors in the governance of hate speech, operating within a hybrid normative framework that blends legal obligations with private standards. Their moderation practices increasingly rely on AI-driven, systemic approaches, yet raise critical concerns regarding transparency, due process, and alignment with international human rights norms. The DSA represents an ambitious attempt to structure this regulatory landscape by introducing systemic risk assessments and co-regulatory mechanisms such as the revised Code of Conduct+. However, significant uncertainties persist around the implementation, interpretive standards, and accountability mechanisms required to ensure that these measures do not become vectors for over-removal or state-driven censorship.

---

24 https://www.oversightboard.com/news/why-freedom-of-expression-must-be-the-centerpiece-of-systemic-risk-assessments/

The notion of hate speech as a systemic risls, as introduced by the DSA, offers a potentially transformative perspective. It recognizes the need to move beyond the individual content level and assess broader patterns and harms within the digital public sphere. Nonetheless, this conceptual shift must be matched by robust safeguards that prioritize necessity, proportionality, and the protection of free expression. Without such safeguards, efforts to mitigate online hate may inadvertently erode the democratic values they seek to uphold.

Future work should aim to clarify the operationalization of systemic risk frameworks, promote greater transparency and oversight of platform practices, and reaffirm international human rights law vis-à-vis not national legislative vagaries or opaque corporate policies as the foundational standard for content governance. Only through such a balanced and rights-oriented approach can societies meaningfully address the harms of online hate speech while safeguarding the digital space as a forum for pluralism, dissent, and democratic participation.

Furthermore, this is a context where the limits to restrictive approaches based on the application of normative frameworks need to be acknowledged. Social media content is often a mere reflection of reality. Without prejudice to the capacity of tech platforms to amplify and spread certain types of publications and information campaigns, we shall not neglect the fact that social media content is only a part of a complex communication and political ecosystem. Problems of polarisation, social cohesion, youth disengagement, dangerous behaviours, or information vulnerability can only be properly addressed through comprehensive social policies, literacy improvement efforts, and overall democratic quality. Adopting expeditious and restrictive measures affecting online content is often the consequence of short-sighted, yet sometimes well-intended, political willingness to show resolution and provide oversimplified answers to complex problems.

References

Natalie Alkiviadou (2024): "Platform liability, hate speech and the fundamental right to free speech". *Information & Communications Technology Law.* *https://doi.org/10.1080/13600834.2024.2411799*

Natalie Alkiviadou, Jacob Mchangama, Raghav Mendiratta (2020): *Global Handbook on Hate Speech Laws*. Justitia. https://futurefreespeech.org//wp-content/uploads/2020/11/Report_Global-Handbook-on-Hate-Speech-Laws.pdf

Joan Barata (2021): “The Digital Services Act and Its Impact on the Right to Freedom of Expression: Special Focus on Risk Mitigation Obligations”. DSA Observatory.

https://dsa-observatory.eu/2021/07/27/the-digital-services-act-and-its-impact-on-the-right-to-freedom-of-expression-special-focus-on-risk-mitigation-obligations/

Joan Barata (2022): “The Decisions of the Oversight Board from the Perspective of International Human Rights Law”. Special Collection of the Case Law on Freedom of Expression. Global Freedom of Expression Columbia University.

Joan Barata (2024): “The Future of Free Speech: Old Threats and New Challenges”. In: Eric Heinze, Natalie Alkiviadou, Tom Herrenberg, Sejal Parmar and Ioanna Tourkochoriti (eds) *The Oxford Handbook of Hate Speech*. Oxford University Press.

Yoshai Benkler, Robert Faris and Hal Roberts (2018): *Network Propaganda*. Oxford University Press.

Agustina del Campo, Nicolás Zara, Ramiro Álvarez Ugarte (2025): "Reclaiming Human Rights for Platform Governance: Proposals for Restoring Their Centrality in the Era of Risks". In: *Proceedings of the Fourth European Workshop on Algorithmic Fairness (EWAF'25). Proceedings of Machine Learning Research*. Centro de Estudios de Libertad de Expresión. Research Paper 66.

Natasha Duarte and Emma Llanso (2107): ‘Mixed Messages? The Limits of Automated Social Media Content Analysis”. Centre for Democracy & Technology

https://cdt.org/insights/mixed-messages-the-limits-of-automated-socialmedia-content-analysis/

Evelyn Douek (2022): “Content Moderation as Systems Thinking”. 136 Harvard Law Review 526.

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

Eric Heinze (2016): *Hate Speech and Democratic Citizenship*. Oxford University Press.

European Union Agency for Fundamental Rights (2023): *Online content moderation. Current challenges in detecting hate speech*. Publications Office of the European Union.

Jeffrey W. Howard, Beatriz Kira, Louisa Bartolo (2024): "Remove or Reduce: Demotion, Content Moderation, and Human Rights". SSRN: https://ssrn.com/abstract=4891835

Dia Kayyali (2025): "Meta's Oversight Board Gives Hate a Pass". Tech Policy Press. https://www.techpolicy.press/metas-oversight-board-gives-hate-a-pass/

Daphne Keller (2019): "Who Do You Sue? State and Platform Hybrid Power over Online Speech". Hoover Institution, Aegis Series Paper No. 1902.

Kate Klonick (2023), "On Systems Thinking and Straw Men". 136 Harvard Law Review Forum 339.

Seul Lee and Anne Gilliland (2024): "Evolving Definitions of Hate Speech: The Impact of a Lack of Standardized Definitions'" in Isaac Sserwanga, Hideo Joho, Jie Ma, Preben Hansen, Dan Wu, Masanori Koizumi, Anne J. Gilliland. Wisdom. *Well-Being, Win-Win*. iConference 2024.

Abdurahman Maarouf, Nicolas Pröllochs, Stefan Feuerriegel (2024): "The Virality of Hate Speech on Social Media". Proceedings of the ACM on Human-Computer Interaction, Volume 8, Issue CSCW1. https://doi.org/10.1145/3641025

Jacob Mchangama (2011): "The Sordid Origin of Hate Speech Laws". Hoover Institution. https://www.hoover.org/research/sordid-origin-hate-speech-laws

Jacob Mchangama, Natalie Alkiviadou, and Raghav Mendiratta (2023): "Scope Creep: An Assessment of 8 Social Media Platforms' Hate Speech Policies". Justitia. https://futurefreespeech.org/scope-creep/

Pauline Paillé, Catherine Galley, Kristin Thue, and Dr. Benedict Wilkinson (2021): *Gamification and online hate speech*. European Commission – Radicalisation Awareness Network.

Alexandra A. Siegel (2020). "Online Hate Speech": In: Persily N., Tucker J.A. (eds) *Social Media and Democracy (SSRC Anxieties of Democracy)*. Cambridge University Press.

Juha Tuovinen (2025): "The Meta Oversight Board in the Trump Era". Tech Policy Press. https://verfassungsblog.de/the-meta-oversight-board-in-the-trump-era/

Joseph B. Walther (2022): "Social Media and Online Hate". Current Opinion in Psychology, Volume 45.