The Legal Framework for Combating Hate Speech Online: Balancing Freedom of Expression and Addressing Incitement to Violence

الإطار القانوني للتصدي لخطاب الكراهية عبر الإنترنت: بين حرية التعبير ومكافحة التحريض على العنف



Imad Eddine Heded [1]

[1] University of El Oued (Algeria), heded-imadeddine@univ-eloued.dz

Received:26/02/2026 *Accepted: 07./04/2026* *Published:22./05/2026*

Abstract:

The rapid spread of hate speech online has emerged as one of the most pressing challenges at the intersection of communication law, human rights, and digital governance. This article examines the evolving legal frameworks that seek to regulate online hate speech while preserving freedom of expression. Drawing on international human rights law, regional legislation particularly the European Union's Digital Services Act and national regulatory models, it explores how different legal systems define hate speech, establish liability for online platforms, and enforce compliance. The article further considers the role of content moderation, algorithmic detection, and civic responses in countering incitement to violence.

Keywords: hate speech, freedom of expression, online regulation, incitement to violence.

ملخص:
انتشار خطاب الكراهية عبر الإنترنت بسرعة أصبح من أبرز التحديات التي تواجه تقاطع قوانين الإعلام وحقوق الإنسان وحوكمة الفضاء الرقمي. يتناول هذا المقال الأطر القانونية المتطورة التي تسعى لتنظيم خطاب الكراهية عبر الإنترنت مع الحفاظ على حرية التعبير. من خلال الاستفادة من قوانين حقوق الإنسان الدولية، والتشريعات الإقليمية خاصة قانون الخدمات الرقمية في الاتحاد الأوروبي، والنماذج التنظيمية الوطنية، يستعرض المقال كيف تحدد الأنظمة القانونية المختلفة خطاب الكراهية، وتؤسس المسؤولية على منصات الإنترنت، وتفرض الامتثال. كما يتناول دور الاعتدال في المحتوى، والكشف الخوارزمي، والاستجابات المدنية في مواجهة التحريض على العنف.
كلمات مفتاحية: خطاب الكراهية، حرية التعبير، التنظيم عبر الإنترنت، التحريض على العنف.

1- Corresponding author**: Imad Eddine Heded, **e-mail: heded-imadeddine@univ-eloued.dz

1. INTRODUCTION

The internet has transformed how human beings communicate, share ideas, and organize themselves. Yet this same transformative power has made online platforms fertile ground for the spread of hate speech messages that attack individuals or groups based on attributes such as race, religion, ethnicity, gender, sexual orientation, or disability. Incidents ranging from coordinated harassment campaigns to online propaganda that preceded real-world violence have forced governments, international bodies, and technology companies to confront a fundamental question: how can hate speech be effectively regulated online without endangering the freedom of expression that democratic societies depend upon? (Udupa et al., 2021)

Communication studies, as a discipline, sits at the centre of this debate. It examines how messages are encoded, transmitted, received, and acted upon within social and technological contexts.

When those messages are hateful or inciteful, they do not merely offend recipients; they can silence minority voices, undermine social cohesion, and in extreme cases contribute to physical violence against vulnerable groups. The 2019 Christchurch mosque attacks, whose perpetrator live-streamed the killings and whose manifesto circulated widely online, provided a devastating demonstration of the link between online hate speech and real-world violence. (Christchurch Call Foundation, 2020)

Despite widespread agreement on the seriousness of the problem, no single global legal standard governs online hate speech. Instead, a patchwork of international treaties, regional directives, national statutes, and platform community guidelines has emerged. International human rights law, embodied in instruments such as the International Covenant on Civil and Political Rights (ICCPR), obliges states both to protect freedom of expression and to prohibit advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility, or violence. (UN Human Rights Committee, 2020) Translating these obligations into workable domestic and platform-level rules has proved deeply controversial, particularly given the global, borderless nature of digital communication.

This article seeks to answer three interrelated questions. First, how do existing legal frameworks at the international, regional, and national levels define and regulate hate speech online? Second, what mechanisms do platforms employ to enforce these rules, and what are the limits of those mechanisms? Third, how can the tension between suppressing harmful speech and

protecting legitimate expression be most responsibly managed? By synthesising findings from recent scholarship in media law, communication studies, and digital governance, the article aims to provide a clear, evidence-based account of where the law currently stands and where it needs to evolve. (Molnar, 2022)

The article is structured as follows: Section 2 describes the methodology; Section 3 presents the collected information and analysis; Section 4 discusses the key results and their implications; and Section 5 concludes with a summary of findings and directions for future research.

2. Methodology

This article adopts a document-based qualitative research approach, combining systematic information gathering with thematic analysis. The methodology involves three main stages: source identification, data collection, and synthesis.

In the source identification stage, relevant academic literature was located through searches of major databases including Google Scholar, JSTOR, and Scopus, using search terms such as "hate speech regulation," "online content moderation," "freedom of expression digital platforms," "incitement to violence law," and "Digital Services Act." Only peer-reviewed journal articles, edited book chapters, official legal documents, and reports from authoritative international bodies published between 2020 and 2024 were considered for inclusion, in order to ensure that findings reflect the current state of the field. (Gorwa, 2020)

The data collection stage involved reading and critically evaluating each source for relevance, credibility, and methodological soundness. Sources were assessed based on the reputation of the publishing outlet, the Author expertise, and the transparency of the evidence base.

In the synthesis stage, the collected information was organized into thematic categories: definitions of hate speech in law and theory; the international human rights framework; comparative regulatory models; platform liability and content moderation practices; and the challenges of balancing free speech protection with effective harm prevention. Findings across sources were compared and contrasted to identify areas of consensus, ongoing debate, and gaps in current knowledge. Diagrams were constructed by the author to visualize key relationships between regulatory frameworks, the spectrum of regulated speech, and content moderation processes, drawing directly on the synthesized evidence. (Blanco-Herrero & Arcila-Calderón, 2021)

3. Information Collection and Analysis

3.1 Defining Hate Speech: Conceptual and Legal Foundations

Any legal framework for regulating hate speech must begin with a clear definition of what it seeks to regulate. Yet defining hate speech has proved to be one of the most contested exercises in communication law. At its broadest, hate speech refers to expression that degrades, threatens, or incites hostility against a person or a group based on protected characteristics. However, this broad understanding encompasses everything from crude insults to sophisticated ideological propaganda, and the law must

be precise enough to distinguish between genuinely harmful expression and content that is merely offensive or controversial. (Udupa et al., 2021)

Brown (2020) offers a widely cited analytical framework distinguishing four dimensions of hate speech: the speaker's intent, the content of the message, the context in which it is delivered, and its likely effect on the target group. (Brown, 2020) Legal systems have weighted these dimensions differently. Common law jurisdictions such as the United States have historically focused on intent and direct incitement, requiring a close causal link between the speech and imminent violent action before state restriction is permissible. Civil law systems particularly in continental Europe place greater emphasis on the content and likely social effect of speech, allowing prohibition of hate speech that creates a hostile environment even without an immediate threat of violence. (UN Human Rights Committee, 2020)

International human rights law attempts to bridge these traditions through the "three-part test" under Article 19(3) of the ICCPR: restrictions on expression must be provided by law, serve a legitimate aim, and be necessary and proportionate to that aim. Separately, Article 20(2) requires states to prohibit "advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence." The UN's Rabat Plan of Action, reaffirmed in subsequent OHCHR guidance published after 2020, offers a six-part threshold test to determine whether

expression crosses the line into prohibited incitement. (OHCHR, 2021)

3.2 The International Human Rights Law Framework

International human rights law (IHRL) provides the normative scaffolding for national and regional hate speech regulation. The tension between Articles 19 and 20 of the ICCPR protecting expression while mandating prohibition of hateful advocacy reflects a deliberate balancing act. How that balance should be struck in practice has been the subject of extensive commentary by the UN Human Rights Committee. (UN Human Rights Committee, 2020)

The Special Rapporteur on freedom of expression has consistently stressed that restrictions on speech must be the exception rather than the rule, and that states must demonstrate that less restrictive measures cannot adequately address the harm. (OHCHR, 2021) At the same time, the OHCHR's 2021 report on online hate speech explicitly recognised that the digital environment had created new vectors for organised hate speech campaigns that traditional criminal law was ill-equipped to address. The report called on states to update their legislative frameworks while cautioning against vague or over-broad definitions that could be weaponised to silence dissent and minority expression. (OHCHR, 2021)

Regional human rights frameworks have taken these international standards and elaborated them in binding legal instruments. The European Convention on Human Rights (ECHR)

and the jurisprudence of the European Court of Human Rights (ECtHR) have developed a body of case law that permits restriction of hate speech under Article 10(2) where it is proportionate, but which has also struck down overbroad national laws that swept up legitimate political speech. (Council of Europe, 2022) The African Commission on Human and Peoples' Rights and the Inter-American Court of Human Rights have similarly developed regional standards, though with somewhat different emphases reflecting their distinct socio-legal contexts. (Molnar, 2022)

3.3 Comparative Regulatory Approaches

Nationally and regionally, a striking diversity of approaches to online hate speech regulation has emerged. Three models have attracted the greatest scholarly attention: the strong free-speech model of the United States, the preventive-prohibition model associated with Germany and much of continental Europe, and the emerging platform-responsibility model embodied in the EU's Digital Services Act. (Gorwa, 2020)

The United States remains the most speech-protective jurisdiction in the world, largely due to the First Amendment's near-absolute prohibition on government content regulation. Under current constitutional doctrine, hate speech however repugnant is presumptively protected unless it falls into a narrow set of categorical exceptions: true threats, incitement to imminent lawless action under the Brandenburg v. Ohio standard, or fighting words. (Citron, 2022) The practical consequence is that the federal government is largely unable to compel online

platforms to remove hate speech, and Section 230 of the Communications Decency Act immunizes platforms from civil liability for most user-generated content. Critics argue this model fails to adequately protect vulnerable communities; defenders contend it prevents the creation of a government-controlled speech apparatus. (Citron, 2022)

Germany's Network Enforcement Act (NetzDG), first enacted in 2017 and significantly amended in 2021, represents the opposite extreme within liberal democracies. Under NetzDG, large social media platforms with more than two million registered users in Germany must remove "obviously illegal" content including 22 categories of hate speech offences under the German Criminal Code within 24 hours of receiving a complaint, with a seven-day window for more complex cases. Non-compliance can result in fines of up to 50 million euros. (Blanco-Herrero & Arcila-Calderón, 2021) While the NetzDG has been praised for demonstrating that states can hold platforms accountable, it has also been criticised for incentivising over-removal: platforms, eager to avoid fines, may delete content that is lawful or merely controversial rather than engage in nuanced legal analysis. (Blanco-Herrero & Arcila-Calderón, 2021)

The European Union's Digital Services Act (DSA), which entered into full force in 2024, seeks a middle path. The DSA introduces a tiered liability regime based on the type and size of the platform, with the heaviest obligations including mandatory risk assessments for systemic harms and obligations to provide

researcher access to data falling on very large online platforms (VLOPs) with more than 45 million monthly active users in the EU. (Kettemann, 2021) The DSA does not mandate the removal of all hate speech; rather, it requires platforms to have effective mechanisms for users to flag illegal content and to act upon such flags in a timely manner. The Act also requires platforms to maintain transparency about their content moderation practices. (Kettemann, 2021)

The United Kingdom's Online Safety Act (2023) adopts a risk-based approach similar to the DSA in some respects but goes further in imposing duties of care on platforms with respect to both illegal content and for larger services legal but potentially harmful content. (Ofcom, 2023) Critics have raised concerns that the "legal but harmful" provisions could chill legitimate speech, while defenders argue that systemic risk assessment is more effective and less prone to error than reactive content removal. (Ofcom, 2023)

3.4 Platform Liability and Content Moderation

A central question in online hate speech regulation concerns the extent to which intermediaries the companies that own and operate social media platforms, search engines, and other online services should be legally responsible for content created by their users. The traditional "notice and takedown" regime, under which platforms escape liability if they remove illegal content upon notification, has been criticized as insufficient given the scale and speed at which harmful content spreads online. (Gillespie, 2020)

Contemporary content moderation practice relies on a combination of automated detection systems increasingly powered by machine learning algorithms and human review. Automated systems can process millions of posts per day and are particularly effective at identifying known illegal content. However, they struggle with contextual nuance: a video documenting atrocity for journalistic purposes may be technically indistinguishable from content glorifying those same atrocities. (Gillespie, 2020) Studies have found significant error rates in automated hate speech detection, with research suggesting that systems are more likely to flag content by users writing in African-American Vernacular English as hateful than equivalent content by other users. (Sap et al., 2020)

Human review teams provide a necessary check on automated systems but face their own limitations. Reviewers must process vast quantities of deeply disturbing content under time pressure, leading to documented mental health harms among content moderators. (Gorwa, 2020) Cross-cultural review is particularly challenging: what constitutes hate speech in one cultural or linguistic context may carry entirely different connotations in another. Many platforms rely heavily on reviewers in lower-income countries, which can create mismatches between the social context of the content and the reviewer's cultural knowledge. (Gorwa, 2020)

Blanco-Herrero and Arcila-Calderón (2021) have argued that effective content moderation requires not just efficient detection and removal of illegal content but also affirmative

measures to promote counter-narratives and media literacy, addressing the demand side of hate speech as well as the supply side. (Blanco-Herrero & Arcila-Calderón, 2021) This perspective, echoed by the Council of Europe's work on Internet Governance, suggests that purely punitive or removal-based approaches are insufficient and must be complemented by educational and civil society interventions. (Council of Europe, 2022)

3.5 Balancing Free Expression and Harm Prevention: Key Tensions

The fundamental challenge in hate speech regulation is that both sides of the equation involve genuinely important values. Freedom of expression is not a mere procedural nicety; it is constitutive of democratic participation, individual autonomy, and the pursuit of truth through open debate. Any restriction on speech carries risks: it can be used by authoritarian governments to suppress political opposition, by majority communities to silence minority voices, or by platforms to avoid controversy rather than to prevent genuine harm. (Molnar, 2022)

At the same time, the harms caused by hate speech are real and well-documented. Research in communication studies and social psychology has found that exposure to hate speech is associated with psychological harm to members of targeted groups, including increased anxiety, depression, and sense of social exclusion. (Brown, 2020) At the societal level, sustained hate speech campaigns can normalize discriminatory attitudes, reduce the willingness of minority groups to participate

in public discourse what scholars call the "chilling effect" on targeted communities and, in extreme cases, contribute to the social conditions in which discrimination and violence become more likely. (Udupa et al., 2021)

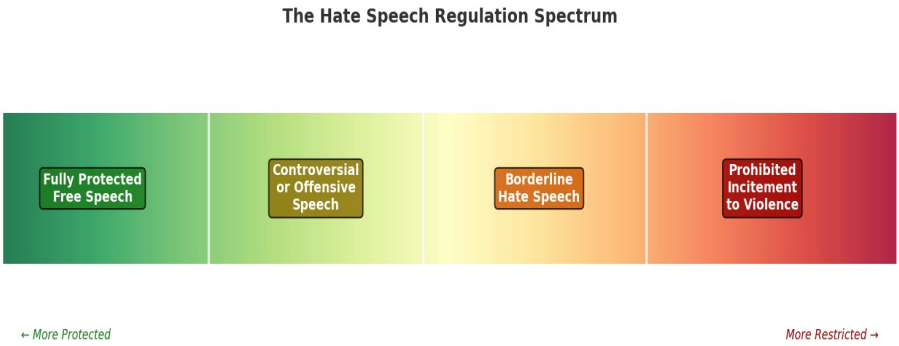
Molnar (2022) has identified what she calls the "dual chilling effect" of hate speech regulation: over-broad rules chill legitimate expression, while the failure to regulate effectively chills the participation of those targeted by hate speech. (Molnar, 2022) Any workable legal framework must grapple with both forms of chilling effect simultaneously, which requires not just clear and precise legal definitions but also robust procedural safeguards, including effective appeals mechanisms and independent oversight of platform enforcement decisions. (Kettemann, 2021)

4. Results and Discussion

4.1 The Speech Regulation Spectrum

One of the clearest findings of this research is that hate speech does not represent a monolithic category but rather sits on a spectrum of expression, with fully protected speech at one end and clearly prohibited incitement to violence at the other. Understanding this spectrum is essential for both lawmakers drafting hate speech legislation and platforms designing content moderation systems.

Figure 1: The Hate Speech Regulation Spectrum



Source: Author synthesis based on Brown (2020), OHCHR (2021), and Molnar (2022)

As illustrated in Figure 1, expression ranges from fully protected free speech including political dissent, religious commentary, and social criticism, however controversial through content that is offensive but not necessarily unlawful, to borderline hate speech that may meet the legal threshold for prohibition in some jurisdictions but not others, and finally to clearly prohibited incitement to violence. The difficulty lies in the middle zones. Differences in cultural context, speaker identity, platform format, and the visibility of the targeted group all bear on whether particular speech should be considered harmful enough to remove or restrict. Legal frameworks that fail to account for this spectrum risk either removing lawful expression or allowing genuinely harmful content to remain. (Brown, 2020)

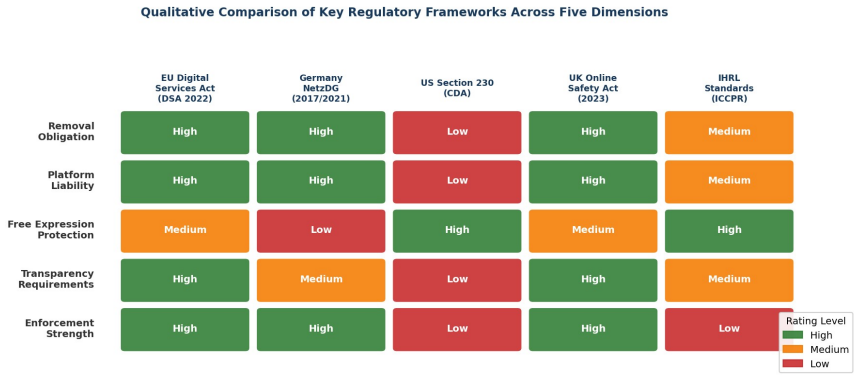
The spectrum also clarifies why purely definitional approaches attempting to resolve the hate speech debate through ever more precise definitions are ultimately insufficient. As

Brown (2020) argues, the same words can occupy different positions on the spectrum depending on who speaks them, to whom, in what context, and with what effect. (Brown, 2020) This insight points toward the importance of context-sensitive content moderation and the limitations of purely automated detection systems that analyse text without its social context. (Sap et al., 2020)

4.2 Comparing Regulatory Frameworks

A second major finding concerns the significant divergence among existing regulatory frameworks and what that divergence reveals about underlying disagreements about the nature and purpose of regulation. Figure 2 presents a qualitative comparison of five key frameworks across five dimensions. The ratings (High / Medium / Low) are derived directly from the primary legal instruments and scholarly commentary cited throughout this article.

Figure 2: Qualitative Comparison of Key Regulatory Frameworks Across Five Dimensions



Source: Author synthesis based on Gorwa (2020), Blanco-Herrero & Arcila-Calderón (2021), Citron (2022, Kettemann (2021), Ofcom (2023), and OHCHR (2021)

The comparison reveals a clear division between the US model which offers weak removal obligations and near-zero platform liability under Section 230, while providing the highest level of free expression protection and European models, which impose strong removal obligations and greater platform accountability. (Citron, 2022) NetzDG combines the strongest removal obligation (mandatory 24-hour takedown) and high enforcement strength (fines up to €50 million) with the lowest rating for free expression protection, reflecting documented concerns about systematic over-removal. (Blanco-Herrero & Arcila-Calderón, 2021)

The EU DSA achieves high ratings across removal obligation, platform liability, transparency requirements, and enforcement strength, while maintaining a medium rating for free expression protection a deliberate design choice reflecting the Act's risk-based rather than prescriptive approach to content removal. (Kettemann, 2021) The UK Online Safety Act scores similarly to the DSA across most dimensions, but its "legal but harmful" provisions even in their diluted final form introduce additional complexity that scholars have flagged as a potential threat to legitimate expression. (Ofcom, 2023)

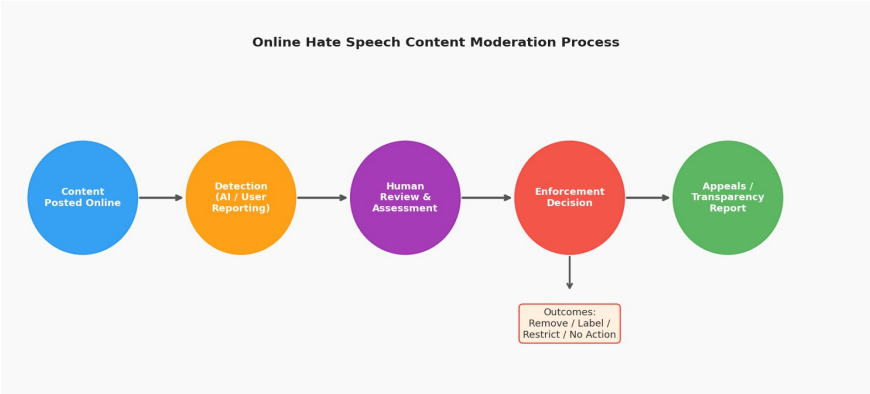
The IHRL framework presents a distinctive profile: medium ratings for removal obligation, platform liability, and

transparency, paired with a high rating for free expression protection but a low rating for enforcement strength. This reflects the structural reality that international human rights obligations bind states rather than platforms directly, and enforcement depends on state compliance with treaty bodies rather than direct regulatory action. (OHCHR, 2021)

4.3 The Content Moderation Process and Its Limitations

A third key finding concerns the practical process by which hate speech is identified and addressed by online platforms. As Figure 3 illustrates, content moderation involves a multi-stage process from initial content posting through detection, human review, enforcement decision, and appeals.

Figure 3: Online Hate Speech Content Moderation Process



Source: Author synthesis based on Gillespie (2020) and Blanco-Herrero & Arcila-Calderón (2021)

Each stage of this process introduces potential for error. Automated detection systems can miss novel forms of hate

speech, misidentify satirical or documentary content as harmful, and reproduce biases present in their training data. Human review is inconsistent, resource-intensive, and emotionally costly for moderators. Enforcement decisions whether to remove, label, demote, or take no action are made under significant uncertainty with limited contextual information. Appeals mechanisms, where they exist, are often slow and opaque. (Gillespie, 2020)

The research also reveals a structural tension in platform incentives. Commercial platforms have financial incentives to maximize user engagement, which hate speech like other emotionally charged content can drive. (Udupa et al., 2021) At the same time, platforms face reputational and regulatory pressure to remove harmful content. This tension is reflected in inconsistent enforcement: content that is highly engaging may remain up longer than less engaging content that violates the same rules, and enforcement is often reactive rather than proactive. (Ofcom, 2023)

The findings suggest that effective content moderation requires systemic reform rather than merely improved detection algorithms. Key elements identified in the literature include: meaningful transparency about moderation policies and enforcement statistics; robust independent appeals mechanisms; binding due process requirements for users whose content is removed; and regular audits by independent researchers with access to platform data. (Gorwa, 2020) The DSA takes steps in this direction, but scholars argue that even the DSA falls short

in guaranteeing genuine procedural fairness for ordinary users. (Kettemann, 2021)

4.4 The Dual Challenge: Protecting Both Expression and Communities

The findings confirm Molnar's (2022) observation that effective hate speech regulation must simultaneously address two chilling effects: the over-restriction of legitimate expression and the failure to protect targeted communities from speech that drives them from public life. (Molnar, 2022) These are not competing goals but complementary ones: a regulatory system that over-restricts expression is both unjust in itself and counterproductive, as it generates backlash that undermines the legitimacy of hate speech regulation more broadly.

The research identifies several principles that appear across the most credible and effective regulatory frameworks. First, legal definitions of prohibited hate speech should be clear and targeted at serious harms incitement to discrimination, hostility, or violence rather than merely offensive expression. Second, enforcement should be contextual and informed by expertise in the relevant cultural and linguistic context. Third, platforms should be required to provide transparent, consistent, and independently auditable enforcement. Fourth, counter-speech and media literacy initiatives should complement prohibitory measures. (Council of Europe, 2022)

Importantly, the evidence suggests that regulatory approaches which involve civil society organizations and community representatives in the design and oversight of content

moderation systems produce better outcomes than those designed and operated exclusively by platforms or governments. (Sap et al., 2020) This finding points toward a governance model that is genuinely multi-stakeholder, drawing on the expertise and legitimacy of affected communities alongside the resources of states and platforms.

5. CONCLUSION

This article has examined the evolving legal and regulatory landscape for combating online hate speech, with a particular focus on the tension between free expression and harm prevention. The key findings can be summarised as follows. First, hate speech is not a binary category but a spectrum, and effective regulation requires legal frameworks and moderation systems that are sensitive to this complexity. Second, significant divergence exists among regulatory approaches, from the strongly speech-protective US model to the preventive-prohibition approach of continental Europe and the emerging risk-based accountability model of the EU DSA. Each approach reflects different social, historical, and constitutional assumptions, and each carries distinct advantages and disadvantages. Third, platform content moderation however sophisticated faces inherent limitations, including algorithmic bias, cultural insensitivity, and structural commercial incentives that can work against consistent enforcement.

The article also identified a number of gaps and directions for future research. Despite growing awareness of the harms caused by hate speech, empirical evidence on the effectiveness of different regulatory interventions remains limited. Longitudinal

studies tracking the impact of legislative changes such as the implementation of the DSA or the UK Online Safety Act on the prevalence of hate speech and on the participation of targeted communities in online public discourse are urgently needed. Similarly, comparative research on how different cultural and linguistic contexts shape both the production of hate speech and the effectiveness of counter-measures remains underdeveloped. (Brown, 2020)

A further gap concerns the position of users in content moderation systems. Most research has focused on platforms and states as the primary actors; less attention has been paid to the perspectives and experiences of users who are either exposed to hate speech or whose content is removed. Research centering the voices of affected communities particularly from the Global South, where both hate speech and its regulation have distinctive features is essential for developing genuinely inclusive governance frameworks. (OHCHR, 2021)

Finally, the rapid pace of technological change continues to outstrip regulatory responses. The emergence of generative artificial intelligence systems capable of producing targeted, personalized hate speech at scale poses challenges that existing frameworks were not designed to address. How international human rights law, platform governance, and national legislation should adapt to this new threat is perhaps the most pressing question facing scholars and policymakers in this field. (Kettemann, 2021)

In conclusion, combating hate speech online is not a problem that can be solved by any single legal instrument, technical system, or institutional actor. It requires sustained, coordinated effort across legal, technological, educational, and civil society domains, guided by clear principles and held accountable by transparent, independent oversight. The frameworks reviewed in this article represent important steps forward, but significant challenges remain. Continued interdisciplinary research at the intersection of communication studies, law, and technology is essential to meeting them.

References

1. Blanco-Herrero, D., & Arcila-Calderón, C. (2021). Deontology and Self-Regulation of Hate Speech in Online Media: Analysis of Codes of Ethics and Comments Moderation Rules in Spain and Other European Countries. Journalism Practice, 15(4), 478–498.
2. Brown, A. (2020). The "Who?" of Hate Speech: Dominant Groups and the Limits of Hate Speech Law. Journal of Law and Society, 47(1), 1–29. https://doi.org/10.1111/jols.12201
3. Christchurch Call Foundation. (2020). Christchurch Call Progress Report 2020: Eliminating Terrorist and Violent Extremist Content Online. Christchurch Call Foundation / OECD. https://www.christchurchcall.com
4. Citron, D. K. (2022). The Internet in Action: Restoring the Promise of American Communications Law. Yale University Press.
5. Council of Europe. (2022). Countering Online Hate Speech: Updated Guidelines for Internet Service Providers. Council of Europe Publishing.
6. Gillespie, T. (2020). Content Moderation, AI, and the Question of Scale. Big Data & Society, 7(2), 1–11. https://doi.org/10.1177/205395172094323

7. Gorwa, R. (2020). The Platform Governance Triangle: Conceptualizing the Informal Regulation of Online Content. Internet Policy Review, 8(2), 1–22. https://doi.org/10.14763/2019.1.1387

8. Kettemann, M. C. (2021). The Normative Order of the Internet: A Theory of Rule and Regulation Online. Oxford University Press. https://doi.org/10.1093/oso/9780198865995.001.0001

9. Molnar, P. (2022). The Content Moderation Dilemma: Platform Governance and Human Rights. Cambridge University Press.

10. Ofcom. (2023). Online Safety Act: Implementation Framework. Office of Communications. https://www.ofcom.org.uk/online-safety

11. Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2020). The Risk of Racial Bias in Hate Speech Detection. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 1668–1678. https://doi.org/10.18653/v1/P19-1163Udupa, S., Gagliardone, I., & Hervik, P. (Eds.). (2021). Digital Hate: The Global Politics of Hate Speech in the Digital Age. University of Michigan Press.

13. UN Human Rights Committee. (2020). General Comment No. 37 on Article 21 of the ICCPR (Freedom of Assembly). United Nations. CCPR/C/GC/37. [Cross-referenced with General Comment No. 34 on freedom of expression, as cited in post-2020 scholarship.]

14. UN Office of the High Commissioner for Human Rights (OHCHR). (2021). Online Hate Speech: Report of the United Nations High Commissioner for Human Rights. A/HRC/47/42. United Nations.