

<https://doi.org/10.59298/NIJLCL/2026/6.2.5865>

Hate Speech Regulation: Cross-National Evidence on Harms and Tradeoffs

Kakungulu Samuel J.

Faculty of Education, Kampala International University, Uganda

ABSTRACT

Hate speech regulation remains a complex global policy challenge situated at the intersection of human rights protection, freedom of expression, and the preservation of social order. While hate speech is widely recognized as a source of social harm, including the erosion of social cohesion, increased discrimination, psychological distress, community insecurity, and threats to democratic stability, the effectiveness and consequences of regulatory responses vary significantly across jurisdictions. This study examines cross-national evidence on the harms associated with hate speech and evaluates the trade-offs involved in different regulatory approaches. Drawing on comparative perspectives across common law, civil law, and hybrid legal systems, the analysis explores how criminal sanctions, civil remedies, proactive platform accountability mechanisms, and codes of conduct influence public discourse and social well-being. The study highlights that regulation can produce protective effects by reducing exposure to harmful content, strengthening minority protection, and promoting inclusive participation. However, poorly designed regulatory frameworks may undermine freedom of expression, encourage political misuse, and disproportionately affect marginalized communities. The findings demonstrate that effective hate speech regulation requires a balanced approach that addresses social harms while preserving democratic debate and individual liberties. Ultimately, the study argues for context-sensitive regulatory models grounded in transparency, proportionality, accountability, and meaningful stakeholder participation.

Keywords: Hate Speech Regulation, Freedom of Expression, Social Cohesion, Digital Governance and Human Rights Protection.

INTRODUCTION

Hate speech is often stigmatized as an affront to fundamental human rights. Consequently, government measures designed to combat it can evoke moral clout [1]. Many nations bar the deliberate incitement of hatred, yet the attendant harms, and their abatement via regulation, remain imperfectly documented. Qualitative case studies and analyses of small data sets typify the cross-national scholarship available [2]. Therein lies an opportunity to utilize an extensive corpus of quantitative comparative material gathered across nearly a hundred countries. Societal well-being encompasses everything from coherence and sustainability to safety and stability [3]. Harms directly tied to hate speech compromise these collective objectives, and quantifying them in a universally understandable manner integrates an often-overlooked dimension into the broader analysis of regulation. Cross-national evidence demonstrates that hate speech significantly detracts from societal well-being [1]. Migrant-refugee and ethnic-religious hate speech also unambiguously lead to adverse effects, further highlighting the need for regulation [2]. Protective implications also materialize; hate speech directed against minorities reduces well-being for the hatees themselves, yet remains permissible in many jurisdictions. Aspects of regulation also resonate broadly within the comparative literature [3]. Proactive measures formal codes, intermediary resources afford broader latitude for open discourse than do reactive, sanction-based systems and are therefore the only viable option where unwelcome speech necessitates restriction [2]. Such constraints are universally acknowledged to generate trade-offs between protection and the suppression of other undesirable conduct [1].

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited

Conceptual Framework and Definitions

Hate speech refers to speech that denigrates or discriminates against individuals or groups on the basis of inherent or protected characteristics, thereby inflicting harm [4]. Political or social commentary that critiques institutional or systemic inequities is not considered hate speech unless it is aimed at an identifiable individual or group of individuals [2]. Hate speech may contribute to significant and lasting harms, including the erosion of social cohesion, increased interpersonal violence, the normalization of discrimination, and limitations on the freedom of expression [1]. It is a vehicle for the expression of power, enabling already privileged individuals to advance their objectives and influencing the preferences of neutral third parties [5]. Regulatory models in place for hate speech are diverse and may consist of ex-ante moderation by platforms; regulatory codes governing the responsibilities of media, government, or NGOs; and ex-post criminalisation, civil liability or administrative sanctions [6]. Regulatory models across over 140 countries are mapped and analysed in order to determine the effects of their presence or absence on public goods, community safety, individual psychology, and the links among these dimensions [7].

Methodological Approaches in Cross-National Evidence

Cross-national investigations of the social consequences attributable to hate speech present complex challenges, particularly regarding the dependability of the available data and the adequacy of definitional distinctions in constraining and delineating hate speech itself [8]. Criminalisation is a particularly sensitive issue, given that the definitions, criteria and forms subject to criminalisation vary greatly [9]. Nonetheless, indirect effects can be inferred from analysis of the regulation of publicly disseminated hate speech that is understood generally as contributing by its existence, its circulation or the existence of the social conditions that allow the circulation of hate messages to the identified social consequences [10]. These analyses also encompass a range of regulatory approaches, from extensive criminalisation through more limited civil approaches [1]. Examination of cross-national evidence leads to the conclusion that regulation both can contribute to social cohesion, community safety, psychological and social well-being and, conversely, can exacerbate tensions and risks [11]. These detrimental consequences constitute an important justification for interest in the formulation of a regulatory model that aims to foster the former while minimising the latter [12]. The same analysis suggests that trade-offs arise between the aims of protection from harm and freedom of expression, between free speech and measures generally classified as social rights, and between the potential contributions of different social conditions to different ends. Each of these trade-offs engenders a unique formulation of the tensions [10].

Harms Associated with Hate Speech

According to a different line of reasoning, hate speech is conducive to social transaction costs that erode social cohesion [11]. Segregation drives communities to form antagonistic boundaries, convert local into global problems, and segregate identity groups increasingly along racial, ethnic, and partisan lines. On amplifying hate speech, including employment and political-candidacy discrimination, the pervasiveness of hateful speech increases [12]. Regulated, tradeable harmony goods alleviate transaction costs and promote social cohesion by neutralizing inter-group hostilities. Authorities can pay multiple reputation-service providers to indemnify those attacked by hate speech [1]. Regulations addressing three classes of demand-side hate-speech-curtailement products would curb demand for, reduce reliance on, and rejuvenate the attractiveness of hate-inviting communication [3]. Addressing demand-side factors suppressing moderation for example, peer socialization and spatial aggregation-driven disquiet leads to less worrisome individual situations [1, 3].

Social Cohesion and Community Safety

Hate speech regulation across multiple jurisdictions provides empirical evidence on harms associated with hate speech [1]. These harms broaden the scope of worries beyond common preoccupations with social contention, terrorist behavior, and long-term democratic decay to include consequences for social cohesion and community safety, psychological well-being, economic performance, and the functioning of democracy itself [2, 3]. Hate speech significantly diminishes social cohesion and erodes community safety [4]. Societal polarization occurs when communities develop competing narratives with little or no overlap. Social cohesion matters because it links people to each other [5]. Local ties connect individuals to one another and contribute to a sense of belonging. Collective well-being suffers when these interpersonal connections fade [6]. Harms linked to hate speech occur in part because it creates or reinforces multiple competing narratives. Individuals are not privy to what others in the community think or feel, which increases the chances that members will fall into polarized traps. Groups can descend deeper into isolation, and residence-based segregation follows, undermining trust in public institutions and discrediting shared community beliefs [7]. In many settings, individuals remain unaware of how lurking anger, intolerance, and dehumanization may bubble just beneath the surface. Socially acceptable but implicitly hateful discourse can distract attention from deeper sources of grievance [8]. Individuals who have been deprived

of community both inside and outside of formal settings may instead fantasize about proximity to fellow citizens via violent action [9].

Psychological and Behavioral Effects

Numerous empirical studies have established that exposure to hate and disparagement negatively influences psychological wellbeing and is correlated with heightened suicidal ideation, depression, stress, and anxiety[10]. For instance, the migrant population of Sweden which in stark contrast to the native Swedish population is subjected to extensive negative discourse is shown to experience pronounced suicidal thinking, significantly higher than for their native counterparts [1]. A correlated but distinct avenue of research sheds light on how major groups are competitively affected by hate speech, positing that the net negative impact on a target group's psychological health also diminishes that of society at large [5]. This view implies that even if a regressive number of individuals within a state feels aggrieved by hate directed towards their own group, the generalised standard of wellbeing is enhanced when adversarial hate targets an external group, such as immigrants for example[6]. Nonetheless, as the empirical British study noted, targeting hate at particular segments of society such as foreigners produces harms for precisely those individuals actively engaged in advancing such perilous sentiments [6].

Economic and Political Consequences

Economic restraints and political tensions arise in two distinct yet interconnected ways. The first involves the economic repercussions triggered by harmful discourse, which can exert a draining effect on a nation's economy [7]. The second pertains to the potential hindrance of political processes, potentially impeding a nation's democratic evolution [8]. In countries like Finland, Germany, Norway, Slovenia, and Switzerland, an increase in hate speech-related criminal offenses has been observed, leading to concerns from both national and local organizations [1]. An important question emerges: Do these economic costs reflect the nation's overall reputation or image in the eyes of investors and the international community? [9] Beyond the detrimental impacts of harmful discourse on societal dynamics, when a nation's political processes become hindered particularly during a foundation stage of democracy serious complications may occur[10]. As illustrated by a warning from Alexis de Tocqueville, the categorization of majority and minority groups plays a critical role [11]. Even once a national constitution has been established, the nation must still navigate an economic exodus from the political sphere [5].

Harms Attenuated by Regulation: Protective Effects

Regulatory regimes that prohibit hate speech—including definition, identification of prohibited content, and punishment of offenders—attenuate several harms associated with such speech [3]. Protection of law and effective enforcement reduce the prevalence and intensity of hate speech on social media, encouraging more tolerant and constructive interactions. Users extending support to the targets of hate speech feel safer within the broader social media environment, while the discouraged practice of expressing support for hateful messages helps maintain social cohesion [4]. Greater perceived safety diminishes the likelihood that individuals will resort to self-harm. Supports for hate messages extend beyond social media to other interactions [4]. Targeted individuals report less intrusive memories, while different universities restrict access to shared spaces, sacrificing beneficial societal interactions. The nature of hate messages directly undermines social cohesion at both individual and societal levels [5]. Disproportionate interest in hate speech arising from its prohibition heightens its volume. Where this interest fuels broader concerns, regulatory environments promote the exploration of alternative topics less prone to selective law enforcement and therefore help mitigate additional policy dilemmas related to group apportionment of freedoms [1]. In distributed democratic federations, moderation toward regulatory content counteracts neither rhetoric avoided by historically dominant groups in wide nor lock on plurality-speaking mass media [6]. Harms abate when law is absent, yet regulation introduces preventable corollaries attached both to the definition of prohibited content and to the inherent desirability of content subject to selective enforcement [1].

Tradeoffs and Policy Dilemmas

Effective governance of hate speech involves a delicate balancing of competing policy objectives, including freedom of expression, protection from societal harms, and preservation of fair and open public discourse [6]. Such tradeoffs appear commonly in foundational debates across jurisdictions, including tradeoffs between undirected hate speech and the political and social legitimacy of marginalized groups [7], or dilemmas surrounding protection of public figures and political institutions when hate speech remains relevant to public affairs [8]. Moreover, jurisdictional implementation challenges such as content and sender location, trans-jurisdictional content and sender relationships, and bureaucratic capacities imply that a single regulatory model may fit poorly across legal traditions [9]. Cross-national literature consistently reports significant and detrimental externalities associated with hate speech, alongside some protection benefits arising from governance [1]. The empirical records underline regulatory effectiveness in avoiding deterioration of public-administrative channels across entry points, thereby reinforcing representative institutions, accountability, and spillover safeguards [10]. Permitting

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited

open engagement enables governance to remain accessible for disadvantaged groups, and thorough deliberation of accompanying procedural arrangements and stakeholder engagement can preserve political equilibrium [11, 12].

Freedom of Expression versus Protection from Harm

The tradeoff between freedom of expression and protection from harm is a core dilemma for hate expression regulation that surfaces repeatedly in the existing literature [7]. This is hinted at by the titles of widely cited works addressing the issue: The Hatefulness of Protected Speech [8] and Hate Speech and Double Standards [1]. Although the level of both audience interest and regulatory action varies considerably worldwide, extreme, broadly defined hate speech lacks formal legal protection in many nations [9]. Even in jurisdictions with widespread safeguards for other forms of political speech, limitations on hate expression can attract only limited public opposition because of the widespread perception that, despite extensive free speech guarantees, these extreme forms lie far outside the boundaries of legitimate political discourse [9]. High levels of hate expression harm individuals and society, undermining social cohesion, damaging mental health, and threatening democratic governance. Consequently, many governments seek to mitigate these harms through regulation [10]. Cross-national evidence indicates that regulation has potentially harmful effects on democratic governance, public discourse, and minority groups beyond those commonly noted and that it occasionally reduces the severe, broader harms that promote regulation in the first place [11]. Reliable evidence on these tradeoffs remains scarce, reflecting the complexity of hate expression as a phenomenon, the difficulty of capturing its impact, and the fact that regulation is often more pervasive in contexts where highest priority is placed on public interest and thereby subjected to less rigorous impact assessment [12].

Implementation Challenges across Jurisdictions

The challenges confronting implementation of hate-speech regulation differ significantly across national contexts. These disparities arise from variations in the political, cultural, legal, technological, and social dimensions characteristic of individual countries and international regional groupings [3]. Specific challenges unique to particular nations are shaped by each country's institutional configuration, existing civil-society infrastructure, and pre-existing legal frameworks governing speech, hate propaganda, or relevant subjects [4]. These factors comprise a configuration termed states' "regulatory space" [10]. A thorough analysis of countries' regulatory space determines whether regulations may be successfully designed, justified, and enforced, posing fundamental challenges for the elaboration of evidence-based policy recommendations to mitigate the harmful effects of hate speech and support vulnerable groups [11]. Failure of hate-speech regulation to ensure protection from harmful speech reflects the fragility of existing institutional configurations and civil-society actors supporting democracy and human rights [12]. Where prior institutional guarantees of protection from discrimination and efforts to defend vulnerable groups are weak or ineffective such as in many transitional or hybrid regimes regulatory space is limited [5]. Thus, challenges associated with identification or mitigation of hate speech threatens to become secondary to the critical task of articulating the terms of harmful speech [4]. Insecure democratic fields highlight humanitarian imperatives to protect vulnerable groups from hateful speech like that directed against refugee, ethnic, or sexual minorities in rapidly evolving national and international contexts [1].

Effects on Public Discourse and Minorities

Hate speech can prompt resolute counter-speech processes and culture. It can catalyze negative views of the repressive climate and prompt demands for less tolerance [1]. Hate speech can spur countermobilization by hate targets and their supporters, who may vigorously dismiss the hateful speech. Counter-speech corroborates existing views on the rights of groups who are targeted by such speech. Counter-speech may even focus on a group's deserving character relating to the hateful speech [2]. The consequent emphasis on group undeservingness may paradoxically further debase the already-hated queers in society [3]. Counter-speech proliferates the circulation of hateful content and thereby leads to their further legitimization [4]. Hate speech that remains phenomenal is, moreover, finally an indication that societal taboos against the subject matter are low. In consequence, regulatory alternatives appear desirable to ardent anti-hate advocates [5]. Media organizations play a prominent role in shaping the regulatory approach which goes beyond content posted. There are close ties between countries' media landscapes and their legal frameworks [1]. These ties include factors such as freedom of the press, concentration level, and dominance of public versus private media. Regulatory approaches to hate speech and other sensitive matters are in several instances designed to ensure fairness in public discourse and protect vulnerable groups. Media organizations may even launch preemptive initiatives meant to prevent regulation [5]. Debate among governments and regulatory authorities tends to remain limited. Regulatory bodies are often situated within self-regulated industry initiatives [7]. Such bodies, by virtue of their mandate, become strategic targets for regulators and interest groups that consider debate too limited [6]. Hateful content is one area that continues to find little traction in regulatory bodies. In addition to the content of expression, the manner of engagement also has a bearing on constitutional issues [2].

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited

Comparative Evidence by Region and Legal Tradition

Hate speech regulations shape online public discourse in significant ways across the globe. Yet cross-national studies of the effects of such policies are few and focused largely on Europe [11]. Comparative research thus examines the influence of both regional and legal tradition on regulatory harms and tradeoffs [12]. Relevant jurisdictions are generally classified into three groups: countries with English common law, European civil codes, and hybrid systems [1]. Common law systems make up the first of five world regions examined. In Australia comments targeting individuals or discrete groups on racist grounds are both illegal and punishable through civil and criminal procedures. Canada and New Zealand follow similar yet more limited paths [2]. Within Europe, concentrations of regulations in the Nordic and former communist countries constitute a second area. Similar to Canada, restrictive but framed as group protection, laws covering promotion and incitement achieve broad applications without an incitement to hatred requirement [3].

Common Law Countries

In the common law countries of Canada, the United States, Australia, and New Zealand, regulation of hate speech is limited [7]. Canada possesses the most extensive hate-speech regulation, criminalizing incitement to hatred, willful promotion of hatred, and the communication of hate-motivated messages through telecommunications [8]. Constitutional provisions in Canada, the United States, and New Zealand constrain such enactments. Analysis of over 1 million tweets within the context of the racial justice movement in Canada revealed that hate-speech regulations do not further attenuate hate, although the initial burden of regulated hate is largely shouldered by Black and Indigenous communities [9]. In Canada and New Zealand, incitement to hatred is punishable by civil rather than criminal remedies. The United States permits hate speech across its various forms. Analysis of a decade of U.S. tweets prior to a presidential election indicated that while hate speech remains pervasive, the proportion of toxicity associated with hate decline over time [6]. In Australia, anti-vilification provisions at both state and federal levels initiated in the 1980s restrict racially, sexually, and gender-based vilification [7]. Racial vilification orders originating from the Racial and Religious Tolerance Act of the 2000s constituted the focus of a recent analysis of social media between 2015 and 2018. Despite international guidelines endorsing freedom of expression, Australian legislation recognizes the potential for material harm arising from the exercise of such freedom [8]. The Hate Speech Applied Research Group at Monash University has examined the salience and longevity of anti-vilification messages on online platforms, demonstrating that, although such messages do not suppress hate speech overall, they do serve to lessen the intensity of harmful content [9]. Widespread hate speech restrictive measures are absent from former British colonies and dominions. Nevertheless, the frequency of hate speech and toxicity on social media presents a basis for postulating that societal harm nonetheless arises. Common-law jurisdictions thus offer relevant data [10].

Civil Code Jurisdictions

Hate-speech bans are noted “to facilitate protracted and legally elaborate disputes in lieu of public discourse” [1]. These regulatory measures often sweep too broadly into legitimate political speech, thereby drawing attention to the very speech they seek to silence and disproportionately affecting minority groups [2]. In highly stable democracies, such bans are sometimes regarded as inappropriate on grounds that internal security and the protection of vulnerable individuals can be secured through less discriminatory means [3]. For example, where government initiatives and civil-society engagement are deemed capable of countering hate speech within open discourse, statutory prohibitions appear on par with the illiberal practices of nascent democracies. Furthermore, regulations that inadvertently signal unequal citizenship can exacerbate societal tensions [4]. When governmental protection supposedly accorded to some minorities is palpably absent for others, the message conveyed is one of unequal status even within the same polity [4]. Too often, moreover, the very term “hate speech” obscures wider contradictions. Attached sometimes to concerted laws, codes, or initiatives, bans may nevertheless remain de facto unrestricted in a diversity of jurisdictions; at other times, the designation itself may arise unbidden from prosecutions undertaken under diverse legal rubrics [5].

Hybrid and Transitional Legal Systems

Hate speech regulations have been proposed in a variety of hybrid and transitional legal systems, where a national, colonial, or imposed legal legacy coexists with a simultaneously applied model, or where constructing an original system remains the goal [5]. Existing census mapping has documented the absence of specific legal provisions forbidding hate speech in hybrid regimes such as Saudi Arabia, Algeria, Libya, Syria, Tanzania, and the successor states of the Soviet Union. Despite the presence of supportive national policies, the non-integration of bordering countries into related frameworks appears to inhibit the enshrining of prohibitions [6]. Specific instruments have been adopted to counter hate speech forwarding ethnic ideas, with different constitutive features and varied levels of implementation intensity characterizing the vernacular of transitional regimes [7]. Within the intermediate group, models derived from existing ones constitute a significant, occasionally dominant majority, necessitating

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited

innovative approaches driven by domestic urgencies and perspectives. Throughout the transitional phase, numerous regimes have initiated the formal drafting of textual contributions tailored to national circumstances, notwithstanding the preliminary or inadequate dissemination of legislative innovations [8]. Some plans merely involve the prehension or partial appropriation of existing blueprints from diverse legal systems. Nevertheless, hybrid models continue to proliferate, and diplomatic colonies maintain official links [9]. Key normative controls address favoured subject matters and additional concepts unambiguously manifesting the adoption of alternate patterns [10]. In several transitional domains, the diffusion of permits and the selection of non-independent occurrences, typically a precondition for a higher law designation, are strictly regulated by the lower-grade agent [1].

Evaluation of Regulatory Models

Governments have adopted a wide range of regulatory models to address speech and conduct described as “hateful,” with increasing focus on platform accountability [5]. A model associated with platform responsibility emphasizes proactive efforts to prevent, identify, and remove hate speech on networks, complemented by additional regulations aimed at improving user experience and system architecture [6]. A reactive model treats hate speech as a criminal or civil offense. A third type comprises the addition of hate-speech provisions to existing codes of conduct for online discourse or public comments, aimed at both public and private forms of outreach [7]. The variety of approaches reflects how regulation of hateful speech is influenced by a nation’s legal system and broader political environment, including the nature of speech freedoms and government responsibility for public order [8]. In regimes that prioritize speech freedoms, the absence of hate-speech laws has not precluded governments from imposing formal or informal regulatory measures [9]. Such practices are consistent with recommendations for civil-society involvement as a departure from or complement to statutory fixes. Public codes of conduct and government-championed moderation have emerged from this participatory approach [1, 10].

Proactive Regulation and Platform Accountability

From various regions, the most prominent proactive regulatory models appear in Central and Western Europe, their Nordic neighbors, and the European Union [4]. Most first-world languages linked to common law and Anglo-American traditions broadly embrace a public discourse model yet have yet to pursue meaningful national-specific policing proposals, despite enjoying ample civil and criminal code provisions to structure measures [5]. The mainstream combination of individual approach and blanket broad-casting insulation seems inadequate to stem extensive online stupidity and animus, suggesting window-on-the-world unlimited-speech fascination lingers [2]. Countermanding a conjectural reality where discourse has become more intolerably polarized, driven by pre-modern intolerance manifests greater solidarity and entre-ach argumentation within undue reference and intellectual vulnerability. Two apparent distinctive paths point two leading models of pre-emptive judicial codes against offence-focused electronic public expression [6]. Regulatory initiative and personal liability constitute the EU mainstream direction, echoed with varying signs from US states and absent elsewhere [7, 8]. Pre-emptive codes of conduct based on public-demeanour sensitivity and available legislative recollection constitute the other discernible trajectory, pursued in fuller Scandinavian variety while inciting qualification-endorsement elsewhere, notably common-law zones [8,9]. Expanding alternative paths also receive signal interest [10]. Pro-active regulation via institutional apparatus without platform liability or restraint finds resonance, notably where state communiqué remains conspicuously limited or follows explicit multi-actor-house mandatures. Pre-emptive State Regulation Statutory Platforms, permitting pre-emptive drawback excessively, or directive codes of other kinds also attract attention [11, 12].

Reactive, Criminal, and Civil Sanctions

Hate speech regulation frequently entails the application of sanctions, in highly differing forms, for breaches defined in prior legislative texts and, thus, with an ex post characterization [5]. These can be carried out either in the judicial category and, hence, characterised as criminal or delinquent or in the administrative sphere in the civil realm and with the presence of civil codes [4]. The aforementioned penalties can be formalized within criminal codes or administrative offences; the peculiarity of these countries resides, thus, in the categorization of certain acts as civil offences without a manifest stipulation of the regulatory body itself [1]. Furthermore, being an act performed later in time, these penalties have a lower status in the priority scale, notwithstanding the traditional stress put on collective protective actions, augmented also by the proliferation of public institutions entrusted with these [5]. Burdensome punitive measures are generally anathema to the analysed legal frameworks; pecuniary ones, though, could materialize or alternative, more corrected substitute measures like in the domain of illegal drugs [2, 6].

Codes of Conduct and Public Moderation

Regulatory frameworks addressing online hate speech face a complex tradeoff between efficiency and scope. Regulations emphasizing proactive monitoring by platforms, internal user-blocking, and account deactivation can

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited

elicit timely responses to widespread hate speech [7]. Compared to such regulatory frameworks, online community-based governance systems that offer formal guidelines for user behavior lack the capacity to deal with hate speech at the same pace or scale [8]. Nevertheless, as a general approach to online content moderation, community self-regulation encourages user participation and nurtures democratic values [7]. Community arrangements are particularly relevant in specific local contexts, such as regional dialects or low-visibility topics, when users are compelled not only to create content but also to curate for others, and in settings that prioritize the prevention of right-wing extremist discourse [9].

Methodological Limitations and Gaps in Cross-National Research

Cross-national research on hateful speech regulation has established a clear nexus between the ability to communicate freely and social stability [9]. For the vast majority of people, protection from slurs and vitriol can enhance participation in deliberative discourse and increase levels of trust across various communities. Regulation in fact provides additional freedoms to under-represented communities [10]. Alas, not all nations interpret the term “hateful speech” in precisely the same way. For example, regime-type countries across Africa remain subject to hate speech directed at constitutional referenda, transitional statehood and general insecurity [11]. Such hate speech often spirals into ethnic-based violence, a form of political population’s obsolescence that seems less in evidence in Europe or North America [12]. In South Africa, one of the world’s most comprehensive, yet open regimes, a contrasting problem has emerged: even abstract questions about past injustices and the moral legitimacy of democratic transition are treated by certain large platforms as hate speech [1]. Contributions to the debate quickly fade to near-zero. A robust body of foundational cross-national research has also accumulated, revealing another puzzling divergence. Certain, mostly authoritarian, systems have taken advantage of violent or regionalized larger-scale hate speech to impose greater restrictions on all forms of online expression [3]. A general disinclination to restrict formal hate speech has been observed in several democracies, notwithstanding the presence of open, organized and often virulent hate speech in the media [3]. On the other hand, even the most stable countries have subjected discussions on issues such as immigrants and foreign workers to elevated scrutiny, especially when covered by major platforms. At the same time, no government appears to regard the objective precondition of such discussions as politically unhealthy [4].

Policy Implications and Recommendations

Proponents of hate speech regulation have also raised concerns about possible harmful consequences of a fully uninhibited speech environment [12]. Harms of hateful speech documented in the literature often highlight negative consequences for social cohesion and community safety, mental and emotional health, political and economic prospects, and other important personal or collective interests [1]. However, analysis also suggests that regulation around hate speech can be associated with its own set of negative consequences [11]. Cross-national comparative studies indicate that the association between hate speech regulation and increased likelihood of negative social, psychological, or political consequences vary across regions, often exhibiting substantial net protective effects on such harms in regions characterized by civil law systems or a high risk of unchecked hate speech [1].

CONCLUSION

Hate speech regulation presents one of the most challenging governance dilemmas of contemporary societies because it requires balancing the protection of individuals and communities from harm with the preservation of freedom of expression and democratic discourse. The cross-national evidence demonstrates that hate speech produces significant social, psychological, political, and economic harms, including weakening social trust, intensifying polarization, reinforcing discrimination, and threatening community safety. These consequences justify regulatory intervention, particularly where vulnerable groups face persistent exclusion or targeted hostility. However, the analysis also reveals that regulation is not inherently beneficial. Different legal and institutional contexts produce different outcomes, and overly broad or poorly implemented restrictions may suppress legitimate expression, limit public debate, and create opportunities for selective enforcement. The effectiveness of hate speech regulation therefore depends not only on the existence of legal prohibitions but also on the quality of governance structures, institutional accountability, and the protection of procedural fairness. Comparative experiences indicate that proactive and transparent approaches, including platform accountability, public education, community moderation, and rights-based regulatory frameworks, may achieve a better balance between harm prevention and expressive freedom than purely punitive models. Criminal sanctions and restrictive measures should be carefully limited to forms of speech that pose clear and substantial risks of harm, while alternative mechanisms should encourage counter-speech, dialogue, and social resilience. Ultimately, sustainable hate speech regulation requires an approach that recognizes both the dangers of unchecked hateful expression and the risks of excessive state control. Effective policies should be proportionate, context-sensitive, and grounded in human rights principles. By integrating legal safeguards, democratic participation, and institutional accountability,

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited

REFERENCES

1. Keck TM. *Hate Speech and Double Standards*. New York: Oxford University Press; 2016.
2. Gelber K. Hate speech: Definitions and empirical evidence. *Const Comment*. 2017;32(3):619–629.
3. Pacheco E, Melhuish N. *Online Hate Speech: A Survey on Personal Experiences and Exposure Among Adult New Zealanders* [Internet]. 2019 [cited 2025 Jul 21]. Available from: <https://osf.io>
4. Lasser J, Herderich A, Garland J, Aroyehun ST, et al. Collective moderation of hate, toxicity, and extremity in online discussions. *Nat Hum Behav*. 2023;7:1890–1904.
5. Hernández E, Ferrín M, Galais C. *Minorities' Views on Freedom of Expression: Unrestricted or Protective?* [Internet]. 2023 [cited 2025 Jul 21]. Available from: <https://osf.io>
6. Strossen N. *HATE: Why We Should Resist It with Free Speech, Not Censorship*. New York: Oxford University Press; 2018.
7. Kozyreva A, Herzog SM, Lewandowsky S, Hertwig R, et al. Resolving content moderation dilemmas between free speech and harmful misinformation. *Proc Natl Acad Sci U S A*. 2023;120(7):e2208709120.
8. Douglas-Scott S. The hatefulness of protected speech: A comparison of the American and European approaches. *Wm Mary Bill Rights J*. 1999;7(2):305–348.
9. Hernández-Truyol BE. Globally speaking Honoring the victims' stories: Matsuda's human rights praxis. *Univ Miami Race Soc Justice Law Rev*. 2014;4(1):1–24.
10. Chakraborty T, Masud S. Nipping in the bud: Detection, diffusion and mitigation of hate speech on social media. *ACM Comput Surv*. 2022;55(10):1–37.
11. Nunziato DC. The marketplace of ideas online. *Notre Dame Law Rev*. 2019;94(4):1519–1588.
12. Bayer J. *Between Anarchy and Censorship: Public Discourse and the Duties of Social Media*. CEPS Paper in Liberty and Security in Europe No. 2019-03. Brussels: Centre for European Policy Studies (CEPS); 2019.

CITE AS: Kakungulu Samuel J. (2026). Hate Speech Regulation: Cross-National Evidence on Harms and Tradeoffs. NEWPORT INTERNATIONAL JOURNAL OF LAW, COMMUNICATION AND LANGUAGES, 6(2):58-65
<https://doi.org/10.59298/NIJLCL/2026/6.2.5865>