THE CONSEQUENCE OF HATE SPEECH ON DIGITAL PLATFORMS AND ITS REGULATION

Dr. Ved Prakash Rai[1] & Akash Yadav[2]

ABSTRACT

In an era where digital platforms have revolutionized global communication, they have also become fertile ground for hate speech, which incites discrimination, hostility, or violence against groups based on race, religion, gender, or other identities. The growth of hateful content online has been coupled with the rise of easily shareable disinformation enabled by digital tools. This raises unprecedented challenges for our societies as governments struggle to enforce national laws in the virtual world's scale and speed.

This paper explores the complex interplay between online hate speech and Consequence, examining its definitions, unique features—like anonymity, rapid spread, and persistence—and its role in catalyzing real-world hate crimes. Global studies linking social media spikes to offline violence, the analysis highlights how platforms' algorithms and echo chambers exacerbate polarization and dehumanization. It details enforcement strategies by platforms like Meta, YouTube, and X, which use AI, human moderation, and user reporting, yet face challenges in consistency and bias. The paper also outlines the profound consequences—psychological trauma, social division, economic losses, and physical violence—while reviewing legal frameworks, including UN conventions, EU directives, and India's IT Act, which balance speech freedoms with public order. Ultimately states upholding human rights in regulations, tech companies enhancing due diligence and transparency, and civil society advocating equitable monitoring to foster safer digital spaces that protect vulnerable communities without stifling democratic discourse.

Keywords: hate speech, online regulation, digital platforms, content moderation, legal frameworks.

[1] Assistant Professor, Department of Law, Deen Dayal Upadhyaya Gorakhpur University, Gorakhpur
[2] Research Scholar, Department of Law, Deen Dayal Upadhyaya Gorakhpur University, Gorakhpur

1. Introduction

The rise of digital platforms has transformed communication, enabling unprecedented access to information and expression. However, it has also facilitated the spread of hate speech, defined broadly as content that incites violence, discrimination, or hostility toward individuals or groups based on characteristics such as race, religion, or gender. Regulating hate speech online involves navigating a delicate balance: protecting individuals from harm while upholding freedom of speech, a cornerstone of democratic societies. This paper investigates the legal and ethical dimensions of this issue, analyzing how various jurisdictions and platforms address hate speech and the resulting conflicts with free expression. The Internet is also a means of communication that allows the spreading of free expression globally and for that it is supported that it promotes a democratic culture.[3] In particular, the anonymity and ability of communication of one-to-many and many-to-many has made it an ideal instrument for the wide spreading of hate speech by extremists and hate mongers.[4] Thus, the Internet has become the 'new frontier' for spreading hate.

The following table summarizes key figures:

Region	Year	Total Hate Speech Events/Incidents	Online Component	Source
India	2024	1,165 events	995 (85.4%) shared online	India Hate Lab
U.S.	2024	11,679 incidents	Not specified	FBI
UK	2018-19	1,605 online hate crimes	All online	Stop Hate UK

The India Hate Lab, a Washington-based research group, released a report in February 2025 documenting hate speech events in India for 2024. This report is particularly relevant as it captures the intersection of online platforms and hate speech, which can contribute to hate crimes. Key findings include:

[3] J. Balkin (2008).
[4] J. Banks (2010), pp. 233-239

- ***Total Hate Speech Events**:* 1,165 events targeting religious minorities, primarily Muslims, were documented in 2024, a 74% increase from 668 events in 2023. This surge was notably linked to the general election period, suggesting political influences.

- ***Online Dissemination**:* Of these, 995 (85.4%) were first shared or live-streamed on social media platforms, including Facebook (495 events), YouTube (211 events), Instagram, and X. This high online presence indicates the role of digital platforms in amplifying hate speech.

- ***Dangerous Speech Online**:* Out of 259 events classified as dangerous speech (explicit calls for violence), 219 (84.6%) were first shared or live-streamed online. As of February 6, 2025, only 3 out of reported videos were removed by Facebook, with 98.4% still accessible, highlighting enforcement challenges.

1. The Definition and features of Hate Speech

- ***Definition***

The definition of the legal term "hate speech" can be found in various international legal texts. In particular, according to the Committee of Ministers of the Council of Europe, it covers all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin. This term is also defined by the European Court of Human Rights (ECtHR), which refers to hate speech as covering all forms of expression which spread, incite, promote or justify hatred based on intolerance (including religious intolerance)[5], but also as the speech which glorifies violence. Due to the lack of a generally accepted definition of unlawful hate speech, social networking services and websites provide their own definitions. So, e.g., Facebook defines the term 'hate speech' as "direct and serious attacks on any protected category of people based on their race, ethnicity, national origin, religion, sex, gender, sexual orientation, disability or disease".

Evidently, there is no universally accepted definition of this key term, while national laws introducing penal law provisions prohibiting hate speech differ in the determination of what is

[5] ECtHR *Sürek v. Turkey (no. 1) [GC], no. 26682/95, § 62.*

being banned. In the framework of the EU-Research project "Mandola", a distinction is made between hate speech, in general, designating a "hostile verbal abuse", and illegal hate speech, as this is defined in the Recommendation No. R (97) 20 by the Ministers of the Council of Europe.[6]

In the Code of Conduct on Countering Illegal Hate Speech Online signed by Facebook, YouTube, Twitter and Microsoft, hate speech is defined as "all conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, color, religion, descent or national or ethnic origin".[7] Particularly as regards online hate speech, the Council of Europe in the Additional Protocol to the Convention of Cybercrime includes the definition of "racist and xenophobic material" 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". This definition differs from other international legal texts, such as the 12th protocol to the ECHR and the UN International Convention on the Elimination of All Forms of Racial Discrimination. The reason is that the additional protocol has a specific field of application and should be treated differently. UN Strategy and Plan of Action on Hate Speech defines hate speech as "any kind of communication in speech, writing or behavior, 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, color, descent, gender or other identity factor."

According to UNESCO's guide to online hate speech: "*The endurance of hate speech materials online is unique due to its low cost and potential for immediate revival, ensuring its continued relevance in particular spheres of discourse. As the Internet is not governed by a single entity, concerned individuals, governments and non-governmental organizations may have to address Internet Intermediaries on a case-by-case basis, although leaving the owners of a specific online space to also decide how to deal with users' actions on an ongoing basis.*

[6]See D4.1: FAQ on responding to on-line hate speech, p. 20, online available at: http://mandola-project.eu/m/filer_public/1a/af/1aaf50d3-8a38-40f4-b299-9c343f16cea1/mandola-d41.pdf

[7] Code of Conduct on Countering Illegal Hate Speech Online; see also European Commission – Press release: European Commission and IT Companies announce Code of Conduct on illegal online hate speech,http://europa.eu/rapid/press-release_IP-16-1937_en.htm

- ***Features***

Online hate speech has unique features that distinguish it from hatred Speech deployed in traditional media. In particular, illegal materials can remain in many platforms, for long time in various formats that can be done repeatedly connected. The growth of hateful content online has been coupled with the rise of easily shareable disinformation enabled by digital tools. This raises unprecedented challenges for our societies as governments struggle to enforce national laws in the virtual world's scale and speed.

Unlike in traditional media, online hate speech can be produced and shared easily, at low cost and anonymously. It has the potential to reach a global and diverse audience in real time. The relative permanence of hateful online content is also problematic, as it can resurface and (re)gain popularity over time.

2. Does social media catalyze hate crimes?

The same technology that allows social media to galvanize democracy activists can be used by hate groups seeking to organize and recruit. It also allows fringe sites, including peddlers of conspiracies, to reach audiences far broader than their core readership. Online platforms' business models depend on maximizing reading or viewing times. Since Facebook and similar platforms make their money by enabling advertisers to target audiences with extreme precision, it is in their interests to let people find the communities where they will spend the most time.

Users' experiences online are mediated by algorithms designed to maximize their engagement, which often inadvertently promote extreme content. Some web watchdog groups say YouTube's auto play function, in which the player, at the end of one video, tees up a related one, can be especially pernicious. The algorithm drives people to videos that promote conspiracy theories or are otherwise "divisive, misleading or false," according to a Wall Street Journal investigative report. "YouTube may be one of the most powerful radicalizing instruments of the 21st century,". While internet has made the globe a small and connected place, it has also created a space for unregulated forms of expression. In *Delphi* v. *Estonia,*[8] the applicants approached the court against the order of the Estonian court, wherein the applicants (owners of the internet news portal) had been made liable for user generated

[8] *Delfi AS* v. *Estonia*, Application no. 64569/09 (2015).

comments posted on their website. This was the first case where the court had to examine the scope of article 10 in the field of technological innovations.

- ***Amplification of Hate Speech:*** Social media platforms provide a space where hate speech can spread quickly to large audiences. Research shows that online hate speech, particularly on platforms like X, can escalate tensions and normalize prejudiced views, which may translate into real-world violence. For instance, a 2019 study found that spikes in anti-Muslim and anti-Black hate speech on X were associated with increased racially and religiously motivated hate crimes, including violence and harassment, in areas with high social media usage.[9]

- ***Echo Chambers and Polarization:*** Social media often creates "echo chambers" where users are exposed to content reinforcing their biases. A 2021 study examining anti-refugee sentiment on Facebook in Germany demonstrated that municipalities with higher social media usage saw more crimes against refugees when right-wing, anti-refugee posts were prevalent. The study used exogenous variations, like Facebook outages, to establish causality, showing that social media can act as a propagation mechanism for hate.[10]

- ***Trigger Events and Rapid Spread:*** Major events, such as terror attacks or political shifts, can lead to surges in online hate speech, which correlate with offline hate crimes. For example, a study on Brexit and UK terror attacks in 2016–2017 found that online hate speech spiked post-event, with a small group of accounts (e.g., 6% of users producing 50% of anti-Muslim content) driving much of the rhetoric. These spikes were followed by increases in offline hate crimes, suggesting social media amplifies reactive hate[11].

- ***Dehumanization and Mobilization:*** Social media can dehumanize targeted groups through inflammatory content, making violence seem more acceptable. The UCLA Initiative to Study Hate reported in 2023 that 80% of surveyed youth (ages 10–18) encountered hate speech on social media, with significant increases following events like the October 7, 2023, Hamas attack. This exposure can normalize hateful attitudes among impressionable users, potentially inciting action.[12]

[9] https://blog.oup.com/2019/10/connection-between-online-hate-speech-real-world-hate-crime/
[10] https://academic.oup.com/jeea/article-abstract/19/4/2131/5917396?login=false
[11] https://blog.oup.com/2019/10/connection-between-online-hate-speech-real-world-hate-crime/
[12] https://studyofhate.ucla.edu/smash-social-media-hate/

- ***Lack of Effective Moderation:*** Despite efforts by platforms to curb hate speech, enforcement is inconsistent. A 2020 article noted that platforms like Gab, with lax moderation, foster environments where hate thrives, potentially spilling into real-world actions. Even major platforms struggle to contain coded or subtle hate speech, which can still mobilize harmful behavior.[13]

3. How Social Media Platforms Enforce Rules to Prevent Online Hate Speech?

Social media platforms enforce their rules against hate speech through a combination of policy definitions, automated detection, human oversight, user reporting, and enforcement actions such as content removal or account suspension. These methods aim to balance free expression with the prevention of harm, but they face challenges like algorithmic biases, inconsistent application, and the sheer volume of content. Enforcement typically involves proactive and reactive approaches: proactive measures use technology to scan content before issues arise, while reactive ones respond to user reports or flagged items. Below, outline the key enforcement mechanisms, drawing from platform-specific practices and expert analyses, with examples from major platforms like Meta (Facebook/Instagram), YouTube, and X (formerly Twitter).

a) Defining Hate Speech Policies

Platforms start by establishing clear community guidelines that define hate speech. For instance, Meta categorizes hateful conduct into tiers: Tier 1 includes dehumanizing speech, slurs, and calls for violence, while Tier 2 covers insults or calls for exclusion based on protected characteristics like race, religion, or gender.[14] YouTube prohibits content that promotes violence or hatred against individuals or groups based on attributes like ethnicity or sexual orientation, under its broader harmful content policies.[15] X defines abusive behavior to include hate speech that targets protected groups, with rules against glorifying violence or using slurs.[16]

b) Detection Methods

AI and Automated Tools: Platforms heavily rely on artificial intelligence (AI) for initial

[13] https://www.ukri.org/who-we-are/how-we-are-doing/research-outcomes-and-impact/esrc/tracking-hate-on-social-media/
[14] https://transparency.meta.com/en-gb/policies/community-standards/hateful-conduct/
[15] https://support.google.com/youtube/answer/2801964
[16] https://help.x.com/en/rules-and-policies/enforcement-options

screening due to the massive scale of content (e.g., billions of posts daily). AI uses natural language processing and machine learning to flag potential violations based on keywords, patterns, and context. Meta employs AI to proactively detect a high percentage of hate speech before it's reported, with transparency reports tracking this rate.[17] YouTube's algorithms scan videos, comments, and metadata for harmful patterns, often removing content automatically if it's clearly violation.[18] To improve, platforms train AI on diverse datasets ,including linguistic markers from extremist sites (e.g., plural nouns for racial groups or racial framing of religious identities).[19]

User Reporting: Reactive detection relies on users flagging content. Platforms provide easy reporting tools; Meta routes reports to specialized teams and notifies users of outcomes via in-app support.[20]

c) Review Processes

Human Moderators: AI-flagged content often goes to human reviewers for context-aware decisions. Meta has over 15,000 global reviewers who handle appeals and complex cases, ensuring cultural and linguistic nuance.[21] AI models show promise: AI flags potential hate (e.g., 80% detection rate), while humans verify, reducing errors and enabling scale.[22]

Appeals and Transparency: Users can appeal decisions. Meta allows re-reviews and publishes quarterly reports on actions taken (e.g., content removed, restored).[23]

d) Enforcement Actions

Once violations are confirmed, platforms take graduated actions:

Content-Level: Removal, labeling (e.g., X adds context notices or reduces visibility in feeds), or down ranking. Meta uses warning screens for sensitive content. YouTube removes videos

[17] https://transparency.meta.com/en-gb/policies/community-standards/hateful-conduct/
[18] https://support.google.com/youtube/answer/2801964
[19] https://news.umich.edu/hate-speech-in-social-media-how-platforms-can-do-better/
[20] https://transparency.meta.com/en-gb/policies/community-standards/hateful-conduct/
[21]https://transparency.meta.com/en-gb/policies/community-standards/hateful-conduct/ block
[22]https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons
[23] https://transparency.meta.com/en-gb/policies/community-standards/hateful-conduct/

and may demonetize channels.[24]

Account-Level: Temporary locks, read-only modes, or permanent suspensions for repeat offenders. X suspends accounts for severe risks like inciting violence. Strikes systems (e.g., YouTube's three-strike policy) lead to channel termination.

4. Consequence of Online Hate Speech

Online hate speech, defined as derogatory, discriminatory, or threatening content targeting individuals or groups based on protected characteristics such as race, gender, religion, or ethnicity, has profound and multifaceted Consequence on victims, communities, and society at large. These consequences span psychological, social, economic, and physical dimensions, often exacerbating offline harms due to the viral nature of digital platforms.

Psychological Consequence

Online hate speech inflicts severe emotional and mental health damage on victims, leading to anxiety, depression, self-doubt, and in extreme cases, suicidal ideation or actions. The anonymity and persistence of online attacks amplify feelings of isolation and helplessness, as victims may face relentless harassment without escape. For instance, cyberbullying and hate-driven content on social media can directly contribute to suicides, as seen in cases where targeted individuals are overwhelmed by abusive messages.[25]Victims often experience long-term trauma, including paranoia and social withdrawal, as hate speech dehumanizes them and erodes self-worth.[26]

Social Consequence

At a societal level, online hate speech fosters polarization, normalizes bigotry, and undermines civil discourse. It creates echo chambers where hateful ideologies spread unchecked, leading to increased intolerance and fragmentation of communities. This can desensitize bystanders, making hate more acceptable in public discourse.

Hate speech on platforms like Facebook and YouTube contributes to societal incivility by

[24] https://support.google.com/youtube/answer/2801964
[25] St. Martin's Press, 2013, Abraham H. Foxman, Christopher Wolf p. 199
[26]Harvard University Press, Danielle Keats Citron pp. 33–52

amplifying extremist views, eroding trust and cohesion.[27] It also promotes exclusion, where targeted groups face stigmatization, reinforcing stereotypes and hindering social integration.[28]

Economic Consequence

Victims of online hate speech often suffer professional repercussions, including job loss, damaged reputations, and barriers to employment. Employers may distance themselves from targeted individuals due to associated controversies, while the time and resources spent addressing harassment divert from productive activities. Cyber harassment, such as false accusations, leads to financial losses through lost opportunities and legal fees.[29] Broader societal costs include reduced innovation in digital spaces, as fear of harassment discourages participation from marginalized groups.[30]

Physical Consequence

Online hate speech can incite real-world violence, bridging digital rhetoric to offline actions. It escalates from verbal threats to physical assaults, hate crimes, or even mass violence, as seen in cases where online propaganda mobilizes extremists. Hate speech "doesn't just hurt—it kills," by fueling violent acts against vulnerable populations.[31]

5. THE EXISTING LAW RELATING TO ONLINE HATEE EXISTIHATETHE

United Nations International Convention

United Nations International Convention on the Elimination of All Forms of Racial Discrimination (1968) and the 2013 a General Recommendation on Combating Racist Hate Speech.[32]

The Recommendation states the following behaviors should be criminalize:

- The spread of hateful ideas

[27] Foxman & Wolf, 2013, p. 85, in "Hate Speech and the Gatekeepers of the Internet," highlighting how internet companies inadvertently normalize racism and bigotry
[28] Herz & Molnar, 2012, pp. 329–340
[29] Citron, 2014, pp. 93–120
[30] Foxman & Wolf, 2013, p. 173
[31] Foxman & Wolf, 2013, p. 7, emphasizing how online hate leads to deadly outcomes
[32] Recommendation No. 35, CERD 2013

- Inciting others to hate
- Threatening others in the context of hate, or inciting others to do the same
- Offensive hateful speech that is motivated by inciting others to hate
- Membership of hate-related groups that incite hatred

Article 7 addresses the causes of hate and suggests many ways of stamping out hate speech at its core in schools, workplaces, law enforcement, the judiciary, and the public sector.[33]

The International Covenant on Civil and Political Rights (ICCPR)

The International Covenant on Civil and Political Rights (ICCPR) addresses hate speech indirectly. In particular, Article 19 ICCPR provides for the right to freedom of expression, while Article 20 expressly limits this right in cases of "advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence". This can be interpreted as a regulation of unlawful hate speech.

The United Nations have also adopted the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD), which entered into force in 1969. This Convention includes the obligation of its members to eliminate racial discrimination and to promote understanding among all races. It prohibits hate speech only insofar it related to race and ethnicity. Namely, Article 4 (a) stipulates that state parties:

Shall declare as an offence punishable by law all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any race or group of persons of another color or ethnic origin, and also the provision of any assistance to racist activities, including the financing thereof;

This obligation imposed by the ICERD on state parties is also stricter than the case of Article 20 of the ICCPR covering the criminalization of racist ideas that are not necessarily inciting discrimination, hostility or violence.

[33] https://www.ohchr.org/en/instruments-mechanisms/instruments/international-convention-elimination-all-forms-racial

European Commission against Racism and Intolerance (ECRI)

In Europe, the European Commission against Racism and Intolerance (ECRI) issued the General Policy Recommendation No. 6 on Combating the Dissemination of Racist, Xenophobic and Anti-Semitic Material via the Internet in 200022, and in 2015 it adopted the General Policy Recommendation No. 15 on combating hate speech.[34]

The EU has adopted the Joint Action of 15 July 1996 concerning action to combat and xenophobia and the Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law. Besides these legal acts, Articles 21, 22 and 23 of the Charter of Fundamental Rights of the EU prohibit discrimination and aim at promoting equality between genders and cultural, religious and linguistic diversity. Also, Council Directive 2000/43/EC implements the principle of equal treatment between persons irrespective of racial and ethnic origin.

The Council of Europe adopted in 2003 the Additional Protocol to the Convention on cybercrime concerning the criminalization of acts of a racist and xenophobic nature committed through computer systems. More generally, Article 14 of the European Convention on Human Rights (ECHR) prohibits discrimination when applying the other provisions of the Convention, while additional protocol no. 12 to the European Convention on Human Rights provides for a general prohibition of discrimination.

The Additional Protocol to the Convention of Cybercrime

The Council of Europe introduced the Convention on Cybercrime in 2001, which is a milestone in this area and was signed by big industrialist states such as the United States, Japan, Canada and Australia.[35] Any provisions on cyber hate were excluded from the Convention, since the United States would not accept them. Therefore, the Council of Europe adopted the Additional Protocol to the Convention on Cybercrime concerning the criminalization of acts of a racist and xenophobic nature committed through computer systems.

[34] https://www.coe.int/t/dghl/monitoring/ecri/activities/GPR/EN/Recommendation_N15/REC-15-2016-015-ENG.pdf
[35] https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/185/signatures?p_auth=61ZhEtVX

The EU legal framework on hate speech

The EU Council has adopted the Joint Action of 15 July 1996 concerning action to combat and xenophobia (96/443/JHA). This provides that EU Member States must ensure an effective judicial cooperation and, if necessary, for that purpose, take steps to punish as criminal offences:

- public incitement to discrimination, violence or racial violence or racial hatred in respect of a group of persons or a member of such a group defined by reference to color, race, religion or national or ethnic origin;

- public condoning, for a racist or xenophobic purpose, of crimes against humanity and human rights violations;

- public denial of the crimes defined in Article 6 of the Charter of the International Military Tribunal appended to the London Agreement of 8 April 1945, insofar as it includes behavior which is contemptuous of, or degrading to, a group of persons defined by reference to color, race, religion or national or ethnic origin;

- public dissemination or distribution of tracts, pictures or other material containing expressions of racism and xenophobia;

- participation in the activities of groups, organization or associations, which involve discrimination, violence, or racial, ethnic or religious hatred.

European Court of Human Rights

The European Court of Human Rights has issued a wide range of decisions covering different aspects of hate related speech and conduct, which are also relevant in the online environment and which may be summarized in the following categories:[36]

- ethnic hatred [37]
- racial hate[38]

[36] http://www.echr.coe.int/Documents/FS_Hate_speech_ENG.pdf
[37] ECtHR, *Pavel Ivanov v. Russia*, 20.2.2007.
[38] ECtHR, *Glimmerveen and Hagenbeek v. the Netherlands*, 11.10.1979

- incitement to racial discrimination or hatred[39]
- religious hate[40]
- incitement to religious intolerance [41]
- homophobic activities [42]
- Display of a flag with controversial historical connotations[43]

Law Relating to online hate speech in India h

In India, creation of content for the internet is held to be within the freedom of speech and expression guaranteed under Article 19 (1)(a) of the Constitution. This right is, however, subject to reasonable restrictions under Article 19 (2), permitting restrictions in the interests of public order, decency, morality, or the sovereignty and integrity of India. Such restrictions have to be just, fair, and reasonable. India tackles hate speech online through constitutional provisions, statutory laws, and regulations under the Information Technology Act, 2000 (IT Act), seeking to balance freedom of speech with the need to prevent harm and maintain public order.

The Constitution of India acts as the base framework for hate speech regulation in internet content. While Article 19 (1)(a) enshrines free speech, Article 19 (2) permits restrictions to prevent public disorder, protect decency, or uphold national security.[44] This constitutional base is further complemented by statutory provisions such as the Bharatiya Nyaya Sanhita, 2023, which criminalize various forms of hate speech. Section 196 penalizes actions, including electronic communication, that promote enmity or hatred among religious, racial, or linguistic groups and disturb public tranquility. Section 197 criminalizes assertions that undermine the allegiance of a community to the Constitution or promote disharmony. Similarly, Section 298 punishes acts intended to insult the religious sentiments of a group, while Section 302 penalizes deliberate actions that wound religious feelings. Provisions such as Section 356 (3) and (4) deal

[39] ECtHR, *Le Pen v. France*, 20.4.2010
[40] ECtHR, *Noorwood v. the U.K.*, 16.11.2004
[41] ECtHR, *I.A.v. Turkey*, 13.9.2005
[42] ECtHR, *Vejdeland and Others v. Sweden*, 9.2.2012
[43] ECtHR, *Faber v. Hungary*, 24.7.2012
[44] Constitution of India: Article 19(1)(a) and Article 19(2).

with defamation content, holding liable those persons who print, engrave, or sell the material.[45]

The Information Technology Act, 2000, is important in regulating hate speech online. Section 69A gives the government authority to block access to online content in the interest of sovereignty, security, or public order. The IT Rules, 2021 require social media platforms to remove unlawful or hateful content upon direction from courts or authorities and ensure grievance mechanisms for harmful content. Other laws pertain to specific contexts in which hate speech occurs.[46]

The Representation of the People Act, 1951, forbids hate speech during elections through Sections 123 (3A) and 125.[47] The Cable Television Networks (Regulation) Act, 1995 forbids hate speech in broadcasts, and the Indecent Representation of Women (Prohibition) Act, 1986 tackles gender-based hate speech.[48]

6. Conclusion

The widespread dissemination of hate speech on digital platforms carries far-reaching implications, from intensifying societal divisions and inflicting emotional trauma on marginalized communities to fueling tangible acts of violence in the real world. In India, for instance, the India Hate Lab's 2024 report revealed 1,165 documented hate speech incidents, with 995 of them—equivalent to 85.4%—originating or being amplified online, often directed at religious minorities and including rhetoric that risks escalating into hate crimes.[49] On a broader international level, the FBI documented 11,679 hate-related incidents in the United States during 2024, and emerging research points to online hate speech as a key driver in amplifying offline offenses, even as precise data on digital-specific crimes for 2025 remains elusive.[50] Striking a balance between combating hate speech and safeguarding free expression continues to pose a formidable dilemma, especially as platforms such as X grapple with criticism over uneven moderation practices, exemplified by a 50% uptick in hate speech after its ownership change.

[45]Ministry of Electronics and Information Technology: Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021.
[46] Ministry of Electronics and Information Technology: Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021
[47] Representation of the People Act, 1951: Sections 123(3A) and 125 (No. 43 of 1951).
[48] Cable Television Networks (Regulation) Act, 1995: No. 7 of 1995
[49] https://indiahatelab.com/2025/02/10/hate-speech-events-in-india-2024
[50] AAI Statement on the FBI's 2024 Hate Crime Data Release

7. Suggestions

Ensure Respect for Human Rights and the Rule of Law when Countering Online Hate Speech, and Apply these Standards to Content Moderation, and Regulation.

To States:

Regulation of digital communications must always comply with their [States] obligations under international human rights law, in particular article 19 of the International Covenant on Civil and Political Rights (ICCPR) on the right to hold opinions without interference and the right to freedom of expression. The right to freedom of expression under international human rights law, includes "the freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers". Any restriction to the right to freedom of expression shall only be such as are provided by law and are necessary, as established under article 19.3. It should also take into consideration their obligations related to non-discrimination and equality, including article 2.1 of the ICCPR, as well as the freedom to seek, receive and impart information and ideas of all kinds. For any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence, States have an obligation to prohibit this by law (ICCPR, article 20.2). Also, States must ensure that technology and social media companies comply with their responsibility to conduct human rights due diligence in accordance with the UN Guiding Principles on Business and Human Rights (UNGPs). In this regard, guidance has been provided, in particular on restrictions to the right to freedom of expression, by the UN Human Rights Committee in General Comment no. 34 on article 19, 'Freedoms of Opinion and Expression', the UN Committee on the Elimination of Racial Discrimination in General Recommendation no. 35 on 'Combating Racist Speech', and in the Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence, including the latter's six-part threshold test.[51]

Legislative efforts to prohibit hate speech that reaches the threshold identified in article 20.2 of the ICCPR must be developed through participatory efforts, in particular with the participation of groups who are subject to such incitement to hatred.

- Policies must also be formulated to counter and address hate speech that does not reach

[51] UN Doc. A/HRC/22/17/Add.4, appendix, para. 29.

this threshold through similar participatory efforts.

- Measures taken to address hate speech must aim at building systems that effectively address the problems holistically, offline and online. Such measures should not encourage mass surveillance[52] , criminalization of the exercise of freedom of expression as guaranteed under international law, undermine trust or attempt to regulate each and every piece of content.

- Ensure that official requests for takedowns and removal of online content follow existing 7 A/HRC/51/17 8 A/HRC/50/55 guidelines and are compliant with human rights norms and standards and enhance transparency of such requests, in line with the Rabat Plan of Action.

- States institutions and public authorities should not weaponized social media to spread hate speech.

To Technology and Social Media Companies:

- Conduct human rights due diligence to identify the risks the use of their services may pose to people, and take all reasonable steps to prevent or mitigate such risks, in accordance with the UN Guiding Principles on Business and Human Rights. Effective due diligence has to be a continuous, ongoing and iterative process; supported by efforts to embed human rights into policies and management systems; and aimed at enabling companies to remediate adverse impact that they cause or contribute to.[53]

- Adopt community standards and frameworks for content moderation that are in line with international human rights norms and standards, in particular the guarantees of freedom of thought, opinion and expression, and the rights to equality, non-discrimination and privacy.

- Ensure that frameworks and policies are transparent and are applied consistently. In addition, ensure accessible and consistent notice and review procedures, and effective

[52] A/HRC/51/17

[53] For further information about Human Rights Due Diligence in the technology sector, please see Key Characteristics of Business Respect for Human Rights: A B-Tech Foundational Paper at https://www.ohchr.org/sites/default/files/Documents/Issues/Busines s/B-Tech/key-characteristics-business-respect.pdf

remedies.

- Demonstrate that policies and decision-making processes draw on international human rights norms and standards and associated guidance, including the UN Strategy and Plan of Action on Hate Speech, and improve communication regarding key concepts and definitions, including those related to hate speech.

- Invest in improving the capacity and quality of content moderation in all languages in which their platforms can be used.

- Continue to strengthen efforts to detect and address non-verbal hate speech that appears including through videos, music pictures, memes and other media, as well as through coded language, that might be harder to detect.

To Civil Society:

- Root hate speech analysis and assessment in international human rights norms and standards to monitor online hate speech trends, in collaboration with affected communities; and promote regulations that are in line with international human rights norms and standards.

- Continue advocacy for increased transparency from technology and social media companies and Member States on content regulation and moderation; continue to advocate for enhanced transparency by technology and social media companies, in their products, operations as well as in the data that they make available.