LIBERTY VS. LIABILITY: LEGAL DILEMMAS IN REGULATING ONLINE HATE SPEECH ON SOCIAL MEDIA PLATFORMS

Afshan Ahmad, LL.M., Hidayatullah National Law University, Raipur

ABSTRACT

The expansion of social media platforms (SMPs) in India has intensified the dissemination of online hate speech, leading to complex legal and regulatory challenges. While India has statutory frameworks, judicial precedents, expert committee findings, and reports from the Law Commission, notable gaps persist. This paper delves into complexities surrounding online hate speech, examining the fine balance between upholding freedom of expression and protecting dignity of individual and public order. Although laws like the BNS, 2023, the IT Act, 2000 and the electoral laws address hate speech, The absence of uniform definition complicates the regulation. Indian laws have shaped jurisprudence by invalidating ambiguous laws like Section 66-A of the IT Act, 2000 for violating A.19(a), Constitution of India, 1950.

This paper also highlights international legal frameworks such as ICCPR, ICERD and Rabat Plan of Action and emphasizes that the US provides for protection of free speech while European countries enforce strict regulatory framework to curb hate speech. This paper examines the role of SMPs focusing on community standards and AI-driven content moderation. Despite massive content takedown, the social media intermediaries (SMIs) struggle with volume, sensitivity of context and inconsistencies in laws across globe. The research suggests the urgent need for a legal framework and supports a multi-stakeholder approach involving the government, tech companies, and civil society to combat online hate speech while safeguarding free speech.

However, the paper critically examines the tension between individual liberty and liability, assesses the role of intermediaries, reviews the limitation of existing laws, and proposes the need for comprehensive regulatory framework in the era of SMPs.

Keywords: Freedom of Speech & Expression, Hate Speech, Social Media Platforms, Social Media Intermediaries,

I. INTRODUCTION

Hate speech is defined as speech that is "grossly offensive, targeted at a particular group, and intended to humiliate or incite violence." In India, it is rooted in historical religious and communal tensions, continues to shape public discourse. While Article 19(1)(a) of the Indian Constitution guarantees the right to free speech, it is limited by Article 19(2). This clause is amended by the First (1951) and Sixteenth (1963) Amendments allows reasonable restrictions to be imposed by the State in interests of "security of the State, sovereignty and integrity of India, friendly relations with foreign states, public order, decency or morality, or in relation to contempt of court, defamation or incitement to an offence." Rather than defining hate speech explicitly, the legal framework in India evolved to set constitutional limits on free speech under Article 19(2), ensuring a balance between individual liberty and societal harmony. In a pluralistic society, while dissent and debate are crucial, they must not come at the cost of civil harmony and individual dignity.

The digital age has amplified hate speech's impact, leveraging the internet's anonymity, instant reach, algorithmic amplification, and ability to unite extremist groups, leading to increase in hate crimes, mob violence, communal riots, and political unrest. Online platforms are becoming serious threats to democratic institutions as it is an arena to political discourse, misinformation and diverse narratives.

Globally, both governments and technology companies are actively regulating online hate speech. In 2016, major tech firms, including Google, Microsoft, Facebook, and Twitter, signed "EU Code of Conduct on Countering Illegal Hate Speech Online", committing to review and take down content within 24 hours of receiving a legitimate complaint. This initiative later expanded into the "Code of Conduct Plus," encompassing additional platforms like Instagram, TikTok, YouTube, and LinkedIn. These initiatives seek to reduce online hate speech while maintaining a balance between free speech and public safety through mechanisms such as AI moderation, user reporting, and manual content review. Despite large scale content takedowns, for instance, Facebook's 2024 Transparency Report, reported 6.4 million hate speech takedowns, highlighting the problem persists.[1]

Cross-border regulation of hate speech presents unique legal challenges. Countries have

[1] "Actioned hate speech content items on Facebook worldwide from 4th quarter 2017 to 3rd quarter 2024", *available at:* https://www.statista.com/1004/hate-speech-contentquarter/ (Last visited on April 29, 2025).

inconsistent legal standards, complicating efforts to enforce uniform global norms.[2] For instance, *Yahoo! case,*[3] a French Court held that Yahoo! is accountable for allowing 'Nazi memorabilia' auctions, but a U.S. court refused to enforce the decision due to differences in free speech laws. The differing legal standards complicate its regulation across jurisdictions.

In India, the situation is further complicated by rising anti-minority rhetoric and politically driven misinformation, which influence public opinion, fuel real-world violence and undermine democratic values. Moreover, the absence of a clear legal definition of "hate speech" in Indian law creates significant uncertainty in enforcement.

The *India's Law Commission, in its 267th report* defined hate speech as "any statement or expression that incites hatred against a particular group, based on factors such as race, ethnicity, gender, sexual orientation, and religion." The report further stated that hate speech includes "any word spoken or written words, signs, or visible representations that intend to create fear, alarm, or provoke violence."[4] Another study by the *Observer Research Foundation* emphasized that hate speech is not only about hurtful words but also the intent behind them. It defined hate speech as "any expression that encourages violence, discrimination, or hostility towards a group protected under the Indian Constitution.[5] Broadly, hate speech includes speech i.e. derogatory, insulting, abusive, intimidating, or incites hatred against a community based on religion, caste, race, ethnicity, culture, geography, sexual orientation, or other identities." International body has also echoed this concern. A *2015 UNESCO report* described hate speech as "occurring at the intersection of multiple tensions. It is the expression of conflicts between different groups within and across societies."[6]

Although hate speech is not explicitly defined in Indian law, there are legal provisions to regulate it. The Bhartiya Nyaya Sanhita, 2023 includes two key provisions i.e. Section 196 and Section 299, which address certain categories of hate speech. Earlier, Section 66-A of the IT, 2000 criminalised the act of sending offensive messages online, but the Supreme Court struck it down, as it violated freedom of speech due to its vague and unreasonable restrictions.[7] The

[2] James Banks, "Regulating hate speech online", 24 *IRLCT* 233 (2010).
[3] *Yahoo! Inc. v. LICRA,* 433 F 3d 1199 (2006).
[4] Law Commission of India, "267th Report on Hate Speech" (March 2017).
[5] *Id.*
[6] Iginio, Gagliardone, *et. al*., *Countering Online Hate Speech* (UNESCO, 2015).
[7] *Shreya Singhal v. UOI*, AIR 2015 SC 1523.

absence of clear statutory definition and enforcement standards continues to hinder effective regulation of online hate speech.

II. CRITERIA FOR ONLINE HATE SPEECH

Freedom of speech allows individuals to freely express their thoughts and opinions; however, this right is not without limitations and may be curtailed under certain circumstances. In S*hreya Singhal judgment*[8]*,* the Supreme Court distinguished speech in three forms- 'discussion, advocacy, and incitement'. It clarified that only speech falling under the category of incitement may be lawfully restricted under Article 19(2), while the other two are safeguarded by Article 19(1)(a). To assess whether a particular expression qualifies as hate speech or warrants restricted, courts across different jurisdictions have developed certain criteria. These include:

i. Extremity of the Speech

For any expression to qualify as hate speech, it must convey intense negativity such as strong hostility or deep animosity. However, not every offensive statement qualifies as hate speech.[9] Courts have categorised speech that advocates or discusses unpopular topics or sensitive issues as "low-value speech," which may not always receive legal protection.[10]

ii. Incitement

The key factor in restricting speech is whether it incites violence or discrimination. The U.S. Supreme Court's "imminent lawless action" test aligns with principle led down in *Shreya Singhal case*.[11] Hate speech often conflicts with the values of liberty and equality. While free speech fosters equality in the "marketplace of ideas,"[12] but it can also create a discriminatory environment, especially for marginalized groups who may not have the resources to make their voices heard. However, freedom of speech should ensure that all voices, including those of weaker sections of society, are heard equally.

[8] *Id.*
[9] *Saskatchewan (HRC) v. Whatcott*, 920130 1 SCR 467.
[10] *Chaplinsky v. New Hampshire*, 315 U.S. 568 (1942).
[11] *Brandenburg v. Ohio,* 395 U.S. 44 (1969).
[12] *Police Dept. of Chicago v. Mosley,* 408 U.S. 92 (1972).

iii. Status of Speaker

The identity of the person making the speech also matters when determining whether it should be restricted. Courts have held that politicians, public figures, and influential leaders must be held to a higher standard since their words can shape public opinion and incite large groups of people.13

iv. Status of the Victims

The identity of the targeted group is also crucial. Public figures, such as politicians are bound to tolerate criticism more than that of private individuals. In *Lingens v. Austria*, [14] ECHR held that public figures knowingly subject themselves to scrutiny and must display greater tolerance.

v. Potential Impact of the Speech

The intention of speaker is supposed to be considered. For instance, in *Ramesh v. Union of India*[15], SC evaluated whether a movie had the potential to disrupt public order, highlighting the importance of assessing the likely impact of speech.

vi. Context of the Speech

A statement that may seem hateful in one context might not be considered so in another. Courts often examine the circumstances in which the speech was made to determine its permissibility.[16] Any law that restricts hate speech should at least include intent and incitement to violence as key factors. International human rights law sets a three-part test [17] to evaluate the legitimacy of restrictions on free expression: prescribed by law, legitimate purpose, necessary and proportionate.

III. HATE SPEECH v. FREEDOM OF SPEECH & EXPRESSION

Freedom of speech & expression forms the foundation of a democratic society, enabling individuals to share their views and perspectives freely. It promotes discussion and helps society grow, but it is not unlimited. To maintain public order and protect individuals,

[13] *Incal v. Turkey*, Application no. 41/1997/825/1031 (1998).
[14] *Lingens v. Austria*, (1986) 8 EHRR 407.
[15] *Ramesh v. UOI,* AIR 1988 SC 775.
[16] *Bobby Art International v. Om Pal Singh Hoon*, AIR 1996 SC 1846.
[17] UNHRC, *General Comment No. 34 in its One Hundred and Two Session, Held from July 11 to 29, 2011,* UN Doc CCPR/C/GC/34, para. 22 *(*July 21, 2011).

restrictions are necessary. History has shown that suppressing speech, as seen under regimes like Hitler's or colonial rule, leads to oppression. Learning from this, democratic nations, including India, have guaranteed this right in their constitutions. However, these rights come with reasonable limits. In the Indian context, A. 19(2) of the Constitution permits reasonable restrictions on this right in order to safeguard national security, morality, public order, and the rights of others such as in cases of defamation or incitement to violence.

During the drafting of the Indian Constitution, debates arose over whether to limit speech that could harm minority groups or promote hatred. Dr. Ambedkar referred U.S. Constitution, clarified that no right is absolute, and reasonable restrictions are necessary to prevent abuse.[18] Referring U.S. Supreme Court decisions, it has been clarified that the right to free expression does not extend to speech that is harmful or recklessly irresponsible. [19]

The idea that unrestricted speech ensures open debate is flawed.[20] While dissent and disagreement are vital for a progressive society, unchecked speech can harm public order and individual dignity. For instance, hate speech expressions that offend or incite hostility against specific groups can undermine the social harmony. Philosopher Jeremy Waldron argues that speech harming dignity does more than offend; it erodes the assurance that all citizens, especially minorities, are equal. Thus, while criticism and dissent are essential component of free discourse, expressions that infringe upon the rights and dignity of marginalized communities must be subject to regulation.

Free speech is a fundamental democratic value, acting as a safeguard against state overreach. It is central to human rights frameworks, reflecting its importance in ensuring individual liberty. However, the reluctance to define hate speech or impose restrictions stems from the fear of suppressing this freedom.

IV. LEGAL FRAMEWORK GOVERNING ONLINE HATE SPEECH IN INDIA

The discrimination against Dalits and religious tensions, rooted in the partition of India in 1947, are two major factors driving these laws. To address these issues, India has a range of legal provisions spread across various laws that restrict hate speech.

[18]Constitutional Assembly Debates on November 4, 1948 a*vailable at:* https:/parlib.in/bitstream/126789/7996/1/cad-11-1948 (Last visited on April 30, 2025).
[19] *Gitlow v. New York*, 268 US. 652 (1925).
[20] Owen M. Fiss, *Liberalism Divided* (Routledge, 1996).

The **Bharatiya Nyaya Sanhita, 2023** contains several sections that restrict speech to prevent harm to national unity, religious sentiments, and public order:

- **S. 196 BNS, 2023** penalises "promotion of enmity between different groups on grounds of religion, race, place of birth, residence, language, etc., and doing acts prejudicial to maintenance of harmony".
- **S.197 BNS, 2023** penalises "imputations, assertions prejudicial to national-integration".
- **S.299 BNS, 2023** penalises "deliberate and malicious acts, intended to outrage religious feelings of any class by insulting its religion or religious beliefs".
- **S.302 BNS, 2023** penalises "uttering, words, etc., with deliberate intent to wound the religious feelings of any person".
- **S.353 BNS, 2023** punishes circulating statements or rumours that incite hatred, enmity, or public disorder through electronic means.

Additionally, laws like '**Protection of Civil Rights Act, 1955**'[21], and '**Scheduled Castes and Scheduled Tribes (Prevention of Atrocities) Act, 1989**'[22], specifically protect marginalized communities. These laws penalize Actions that promote humiliation, Encourage the practice of untouchability, or involve any form of caste-based discrimination.

Electoral laws also address hate speech. '**Representation of People Act, 1951'**, prevents promoting hatred or enmity based on caste, religion, race, or language during elections. S.123(3A) and S.125 classify such acts as corrupt practices and penalize them. Despite these laws, political leaders and parties often use communal rhetoric, and the ECI has been criticized for its failure to strictly enforce these provisions.

India also has specific laws to regulate hate speech in media and online platforms:

- **Cinematograph Act, 1952**, deals with certification of films and prohibits hateful content in films.[23]

- **Cable Television Networks (Regulation) Act, 1995** prohibits TV channels from broadcasting hateful or offensive content.[24]

[21] The Protection of Civil Rights Act, 1955 (Act 22 of 1955), s. 7(1)(c).
[22] The Scheduled Castes and Scheduled Tribes (Prevention of Atrocities) Act, 1989, (Act 33 of 1989), s.3(1)(x).
[23] The Cinematograph Act, 1952, (Act 37 of 1952), ss. 4, 5B and 7.
[24] The Cable Television Network Regulation Act, 1995, (Act 7 of 1995), s. 5 and 6.

- **Press Council of India Act, 1978**, ensures responsible journalism and prohibits hateful content in print media.[25]

Hate speech has become more widespread and harder to control. In 2008, India introduced S. 66A of the IT Act, 2000, to penalize online hate speech. However, in *Shreya Singhal case*[26], the SC struck down S. 66A as it was too vague and violated A.19(1)(a). Despite this, other provisions such as S. 69A provides safe harbour to social media platforms for user-generated content. Platforms must follow government regulations and remove illegal content when notified.

Due to the country's diversity and the rapid spread of harmful content online, it is difficult to regulate online hate speech in India. The internet allows it to reach a massive audience quickly, and offenders often hide behind anonymity. While laws like Section 196 and 299 of BNS also applies to online hate speech, but enforcing them in the digital space is difficult.

The UNHRC recognizes the internet's role in promoting free speech but stresses the need for limitations to prevent harm. [27] In India, the judiciary has played a key role in interpreting hate speech laws, as in *Madhu Limaye v. Ved Murti*[28], where the SC upheld S.144 CrPC despite concerns about misuse.

India's hate speech laws are crucial in addressing caste and religious discrimination, but enforcement remains a challenge, especially online. To balance free speech and social harmony, the government, judiciary, and tech companies must work together, ensuring these laws are not misused to suppress political opposition.

V. INTERNATIONAL APPROACH TO REGULATING ONLINE HATE SPEECH

Addressing hate speech involves navigating the delegate balance between upholding freedom of expression and preventing societal harm. International frameworks like 'Universal Declaration of Human Rights' (UNDHR) and 'International Covenant on Civil and Political Rights' (ICCPR) emphasise protecting individuals from discrimination and incitement to violence. A.7 of the UNDHR guarantees protection against discrimination[29], while A. 20

[25] The Press Council Act, 1978 (Act 37 of 1978), s. 12.
[26] *Supra* note 7.
[27] UNGA, '*Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression',* UN Doc A/72/350 (Aug. 18, 2017).
[28] *Madhu Limaye v. Ved Murti*, 1971 SCR (1) 145.
[29] The Universal Declaration of Human Rights,1948, art. 7.

ICCPR obliges member states to prohibit any advocacy of hatred based on religion, nationality, or race that may incite violence or hostility. [30] Additionally, A.19 (3) of ICCPR permits states to impose restrictions on free speech, provided they are lawful server legitimate aim such as safeguarding public health, security, or the rights of others and are unnecessary and proportional. [31]

Supporting these protections, the 'International Covenant on the Elimination of All Forms of Racial Discrimination' (ICERD) strengthens global efforts to combat racism. A. 4, ICERD requires state parties to implement measures to prohibit and address racial hate speech, ensuring stronger safeguards against discrimination.[32] The 'UN Committee on the Elimination of Racial Discrimination' have clarified that the racist hate speech, including online dissemination.[33] For example, in *Jewish Community, Oslo v. Norway*[34], the Committee condemned speeches praising Nazi figures and criticising the Jewish community as violations of Article 4. Similarly, the CEDAW addresses gender-based discrimination[35], with CEDAW Committee recommending measures to prevent harmful portrayals of women in media and online.[36]

The 'Rabat Plan of Action', developed under guidance of UNHCR provides for addressing hate speech.[37] The framework introduces a six-factor test to evaluate whether speech qualifies as hate speech, taking into account elements such as the speaker identity, intent, context, content, audience reach, and likelihood of inciting violence. It emphasises the need for establishing precise definition for key terms like "hatred" and "violence". [38]

At the national level, countries have adopted diverse approaches to regulate hate speech. In the US by the 1st Amendment provides protection from hate speech. However, direct threats, such as those to kill someone online, are prosecutable.[39] Courts apply a "reasonable person" test to

[30] The International Covenant on Civil and Political Rights,1966, art. 20.
[31] *Id.,* art. 19(3).
[32] The International Covenant on Elimination of All Forms of Racial Discrimination, 1965, art. 4.
[33] Conference of the Parties, UN Committee on the Elimination of Racial Discrimination, "*General recommendation No. 35: Combating racist hate speech",* CERD/C/GC/35 (Sept. 26, 2013).
[34] *Jewish Community of Oslo et.al. v. Norway,* CERD/C/67/D/30/2003.
[35] The Convention on Elimination All Forms of Discrimination Against Women, 1979, art. 2.
[36] *Supra* note 33.
[37] UN High Commissioners for Refugees, *"Report of the United Nations High Commissioner for Human Rights on the expert workshops on the prohibition of incitement to national, racial or religious hatred",* UN Doc A/HRC/22/17/Add.4 (Jan. 11, 2013).
[38] *Id.*
[39] *U.S. v. Morales*, 252 F.3d 1070 (2001).

determine if online hate speech constitutes a real threat. Communication Decency Act protect social media platforms and individuals posting illegal content will be liable for user-generated content.[40] In contrast, the United Kingdom takes a stricter approach. The Public Order Act prohibits hate speech, including online content, while the Malicious Communications Act penalizes sending offensive or false electronic communications with up to two years of imprisonment.[41]

European Union members, particularly France and Germany, have adopted stringent measures to combat hate speech. Germany has been particularly sensitive to this issue. The 'Network Enforcement Act, 2017' requires SMPs to take down clearly unlawful content within 24 hours and implement transparent systems for handling user complaints. Similarly, France imposes transparency obligations on social media platforms, requiring disclosure of sponsored content. An 1881 Law on Freedom of the Press targets fake news, allowing legal injunctions to block its dissemination.[42] Japan has also taken steps to address hate speech, enacting a national ban in 2016 following criticism from the UNCERD. The law mandates municipal governments to eliminate discriminatory actions against non-Japanese individuals.

SMPs significantly influence the spread of hate speech, and legislation such as Germany's *Network Enforcement Act* demonstrate proactive measures to hold platforms accountable. However, global consistency in addressing hate speech is lacking, with countries like the US prioritizing free speech and others like Germany and Japan adopting stricter measures. Harmonizing these approaches through international cooperation and shared standards could enhance global efforts to combat hate speech.

VI. ROLE OF SOCIAL MEDIA PLATFORMS IN CURBING ONLINE HATE SPEECH

Regulating hate speech on SMPs Presence a major challenge for governments, tech companies, and the broader society. Given the vast reach and impact of platforms like Meta, X, YouTube, Instagram, and TikTok, these companies have developed community standards and policies to regulate online speech. However, enforcing these policies consistently, balancing free speech and censorship, and using AI-based moderation tools come with several challenges. This

[40] The Communication Decency Act, 1996, s. 230.
[41] The Malicious Communications Act 1988, s.1.
[42] "Initiatives to Counter Fake News: France", *available at:* https://maintgov/law/fakenews/france.php (Last visited on May 01, 2025).

chapter explores the methods used by major social media platforms, the role of AI and algorithmic moderation and the difficulties in content regulation.

Facebook's Initiative

Facebook has established Community Standards that ban hateful content. Under these standards, hate speech refers to "direct attack on people based on protected characteristics such as race, ethnicity, religion, gender, sexual orientation, or disability." Facebook divides hate speech into three tiers:

- **Tier 1:** Includes violent or dehumanizing speech, mocking hate crime victims, or promoting harmful stereotypes.
- **Tier 2:** Involves expressions of contempt, disgust, or inferiority toward a group or individual.
- **Tier 3:** Covers attempts to exclude or segregate people based on their identity.

Facebook allows exceptions for humor, social commentary, or content shared to raise awareness or educate others. For example, a post discussing racism to educate others would be allowed, but a post mocking a racial group would be removed. Recently, in 2023, Facebook (now Meta) has faced criticism for not doing enough to curb hate speech in regions like Myanmar and Ethiopia, where inflammatory posts have fueled violence. [43] Meta has been accused by Human Rights Watch for restricting pro-Palestinian content on Facebook and Instagram, decrying "systemic online censorship" during Israel-Palestine conflict.[44] However, the company has introduced AI tools to detect and remove hate speech more efficiently.[45]

X's initiative

Twitter also has strict rules against hate speech. Its policy states that "users cannot promote violence, threaten, or harass others based on race, ethnicity, religion, gender, or other protected characteristics."[46] Twitter also bans hateful symbols or images in profile pictures or headers. In

[43] "Myanmar: Facebook's systems promoted violence against Rohingya; Meta owes reparations – new report", *available at:* https:///www.amnesty.org//news/2022/10/myanmarfacebook-promoted-violencerohingya-meta-owes-reparations/ (Last visited on May 01, 2025).
[44]"Meta Criticised for Restricting Pro-Palestine Content on Facebook, Instagram", *available at:* https://www.ndtv.com/world-news/meta-criticised-for-restricting-pro-palestine-content-on-facebook-instagram-4716492 (Last visited on May 03, 2025).
[45]"How Facebook uses AI to moderate content", *available at:* https://www.facebook.com/help/1584908458516247/ (Last visited on May 03, 2025).
[46]The X Rules*, available at:* https://x.com/policies/x-rules (Last visited on May 03, 2025).

2022, Twitter permanently banned several accounts linked to white supremacist groups for spreading hate speech.[47]

Users can report hateful content, and Twitter reviews these reports to take action, such as suspending or deleting accounts. In 2023, Twitter (now X under Elon Musk's ownership) has faced challenges in enforcing these rules consistently, with critics arguing that hate speech has increased on the platform.[48] Despite this, Twitter continues to use automated systems and human moderators to tackle harmful content.[49]

You Tube' s initiative

YouTube uses a flagging system, where users can report harmful videos. If a video violates the rules, YouTube removes it or demonetizes it. Categories like "hateful content, violent and graphic content, harmful or dangerous content, nudity or sexual content, copyright violations and threats" are strictly monitored.[50]

Despite efforts by social media platforms to control hate speech, several challenges persist. The sheer volume of content uploaded every minute makes it nearly impossible to monitor everything effectively. While AI tools help detect hateful content, they often struggle to understand context, satire, or nuanced speech, leading to errors. Meanwhile, human moderators cannot review every flagged post. The free speech vs. censorship debate remains controversial, as some argue that removing posts limits freedom of expression, while others demand stricter enforcement. Additionally, concerns about political bias arise, with critics accusing platforms of favouring certain ideologies while censoring opposing views.

VII. CONCLUSION

Regulating online hate speech in India presents both legal and societal complexities. With the Internet surfing is primary space for individual expression efforts to uphold free speech often

[47]"The mass unbanning of suspended Twitter users is underway", *available at:* https:/edition.cnn.com/2024/08/twitter-unbanned-users/index.html (Last visited on May 03, 2025).
[48] "Elon Musk has made a complete mess of X: He is being accused of bias, promoting hate speech", *available at:* https://www.indiattoday.in/technolllogy/news/story/musk-has-made-mess-of-x-accused-of-promoting-hate-speech-2579097-2024-08-08 (Last visited on May 03, 2025).
[49] "*Twitter leans on automation to moderate content as harmful speech surges", available at:* https://www.reuters.com/technology/moving-fast-moderation-harmful-content-surges-2022-21-04/ (Last visited on May 03, 2025).
[50]Community Guidelines, *available at:* https://www.youtube.com/intl/ALL_in/howyoutubeworks/policies/community-guidelines/ (Last visited on May 03, 2025).

clash with the need to shield individuals and communities from harmful content. Although the Indian Constitution guarantees freedom of speech as a fundamental right, it is not absolute. This freedom must be weighed against concerns such as public order, dignity and protection of marginalized groups. Any limitation placed on speech must satisfy three-fold criteria: prescribed by law, necessary and proportionate in addressing the harm intended to be prevented and legitimate purpose.

Despite existing legal provisions under the BNS, the IT Act, and electoral laws, its enforcement remains challenging due to the internet's anonymity, political misuse, the need to balance free speech with social harmony and technological loopholes. Judicial decisions have helped shape the boundaries of free expression, notably through the invalidation of S. 66A of IT Act, which was deemed overly broad and ambiguous. Addressing hate speech effectively in India demands a collaborative effort between the government, judiciary, and tech companies to balance free speech with social harmony, ensuring laws are not misused to suppress dissent.

Countries like the U.S., U.K., Germany, and Japan have adopted varying approaches to regulate hate speech, reflecting their historical and cultural contexts, with some prioritizing free speech and others imposing stricter measures. Hence, India must draw a thin line while ensuring that free speech does not become a shield for hate speech and preventing misuse of laws on hate speech to silence dissent and criticism.

To effectively address the online hate speech in India while preserving democratic values, a multi prolonged and balanced approach is necessary. The first step involves introducing new section in the IT Act or amending BNS provisions to specifically address online hate speech with clear and well-defined terms to prevent vague or arbitrary enforcement. This is to be accompanied by the development of alternative dispute resolution mechanisms that handle online hate speech cases efficiently, reduce the burden on judiciary and provide faster, more user-friendly approach to resolving disputes. As the rapid spread of online content, it is essential to impose stricter penalties for online hate speech not just on the originators of the content but also on individuals who forward or share offensive content. The recommendations from the Vishwanathan and Bezbaruah Committees to draft a comprehensive hate speech law, including Section 153C IPC (penalising acts which are prejudicial to human dignity) and Section 509A IPC (penalising racial insults and discrimination) should be consolidated. The role of social media intermediaries must be strengthened to ensure transparency, improve content moderation and establish grievance redressal mechanisms. At the end promoting digital

literacy and public awareness is crucial. A well informed and responsible citizen accompanied by strong legislative framework plays a crucial role in building inclusive digital environment that uphold the constitutional value of dignity, equality and freedom of expression.