WEAPONIZING FREEDOM OF SPEECH: A LEGAL AND ETHICAL ANALYSIS OF ABUSE OF EXPRESSION ON SOCIAL MEDIA

Karan Sati**,** Guru Gobind Singh Indraprastha University

ABSTRACT

The right to freedom of speech and expression, enshrined as a fundamental right in democratic constitutions, has undergone profound transformation in the digital age. Social media platforms have democratized expression, yet simultaneously weaponized it, facilitating the rapid dissemination of hate speech, misinformation, and defamatory content at an unprecedented scale. This paper examines the critical tension between protecting freedom of expression and regulating its abuse on social media platforms. Through analysis of landmark judicial pronouncements, existing legislative frameworks, and recent developments as of December 2025, this paper contends that the solution lies not in restricting speech through censorship, but in establishing proportionate, transparent, and rights-respecting regulatory mechanisms that account for the unique characteristics of digital communication. The paper analyzes the Indian legal landscape, particularly the Information Technology Act, 2000, the IT Rules 2021, and the emerging Karnataka Hate Speech and Hate Crimes (Prevention) Bill 2025, while drawing comparative insights from international jurisprudence. It proposes a multi-stakeholder framework combining statutory clarity, procedural safeguards, algorithmic transparency, and institutional independence to address online speech abuse while preserving democratic discourse.

Keywords: Freedom of speech, social media regulation, hate speech, content moderation, algorithmic amplification, digital rights, constitutional law, platform accountability.

I. INTRODUCTION

Freedom of speech and expression represents one of the foundational pillars of democratic societies, enabling individual autonomy, truth discovery through open debate, and meaningful democratic participation. In India, this fundamental right is protected under Article 19(1)(a) of the Constitution, which guarantees citizens the right to freedom of speech and expression.[1] However, the exercise of this right has never been absolute. Article 19(2) permits the State to impose reasonable restrictions on this freedom in the interests of national security, public order, decency, morality, contempt of court, defamation, and other enumerated grounds.[2]

The advent of social media has fundamentally altered the landscape of free expression in ways that challenge both traditional legal frameworks and democratic societies themselves. Platforms such as Facebook (3.23 billion monthly active users), X (formerly Twitter with 570 million monthly active users), Instagram (2 billion monthly active users), YouTube (2.7 billion monthly active users), and various emerging applications have transcended their role as mere conduits of information to become the primary arena of public discourse for billions globally.[3] This democratization of expression has genuinely empowered marginalized voices, facilitated grassroots movements for social change, and enabled real-time civic engagement previously impossible in traditional media ecosystems. Simultaneously, however, these same platforms have become vectors for the dissemination of hate speech, communal violence, defamation, coordinated misinformation campaigns, and systematic harassment designed to silence vulnerable populations.

The phenomenon of "weaponizing freedom of speech" refers to the deliberate, often coordinated misuse of expression rights to inflict harm whether to individuals, communities, or social cohesion itself. This represents a critical challenge distinct from traditional speech regulation: rather than State censorship, the concern involves private actors instrumentalizing free speech protections to cause demonstrable harm. As the Supreme Court of India observed in July 2025, "free speech is being weaponized particularly online to fuel communalism, defame individuals, or erode public trust in democratic institutions."[4] The Karnataka Cabinet's December 2025 approval of the Hate Speech and Hate Crimes (Prevention and Control) Bill reflects intensifying governmental concern regarding this phenomenon.[5] The challenge lies in distinguishing between legitimate, if provocative or offensive, speech and expression designed to incite violence, promote systematic discrimination, or spread dangerous falsehoods that

precipitate real-world harms.

This research paper undertakes a comprehensive analysis of the legal and ethical dimensions of speech abuse on social media, examining how existing regulatory frameworks address this challenge and identifying critical gaps requiring remediation. The paper proceeds from a foundational concern: the Internet, designed as an open architecture enabling free expression, has become simultaneously a platform for sophisticated speech weaponization. Current regulatory approaches, developed for pre-digital contexts, prove inadequate to address the unique characteristics of online communication including algorithmic amplification, pseudonymous coordination, rapid virality, and the manipulation of epistemic environments. This paper identifies five critical gaps in existing frameworks and proposes concrete solutions combining statutory precision, procedural robustness, platform transparency, algorithmic governance, and institutional independence.

II. THEORETICAL FOUNDATIONS AND CONSTITUTIONAL FRAMEWORK FOR FREE SPEECH PROTECTION

A. Philosophical Justifications for Freedom of Speech and Their Digital Implications

The protection of freedom of speech rests on multiple justifications articulated by liberal democratic theory, each offering distinct insights into why societies should protect expression even when that expression proves offensive, erroneous, or harmful. First, the autonomy rationale emphasizes that individuals possess an inherent interest in controlling the narrative of their own lives and participating in self-governance. This rationale suggests that respecting human dignity requires permitting individuals to articulate their own vision of the good life, even when others find that vision objectionable. Second, the truth-discovery rationale, exemplified by Mill's marketplace of ideas metaphor, suggests that protecting diverse viewpoints enables the identification and dissemination of truth through open debate, with falsehood eventually displaced by superior arguments supported by evidence. Third, the democratic participation rationale underscores that meaningful democratic engagement requires citizens' ability to speak freely about matters of public concern, including criticism of government and dominant institutions.[6]

However, digital contexts complicate each of these rationales substantially. The autonomy rationale assumes that individuals exercise relatively equal power over narrative-construction,

yet algorithmic curation means that platforms exercise enormous gatekeeping power, determining whose speech receives amplification and whose remains marginal. The truth-discovery rationale depends on epistemic conditions that audiences encounter diverse perspectives, evaluate them according to shared epistemic standards, and progressively identify superior arguments. Social media's algorithmic filtering creates echo chambers and filter bubbles wherein users encounter primarily confirming information, while algorithmic amplification of sensational and polarizing content actively impedes truth discovery. The democratic participation rationale presumes that speech occurs within bounded political communities with shared civic commitment; yet social media enables coordinated campaigns by malicious actors lacking any democratic commitment, seeking merely to manipulate discourse for private advantage or to sow discord.[7]

These complications do not negate the underlying justifications for speech protection; rather, they suggest that applying these rationales in digital contexts requires attention to structural conditions enabling their realization. If autonomy requires meaningful capacity for self-expression, yet algorithms silence certain voices while amplifying others, then addressing algorithmic amplification becomes essential to realizing autonomy rather than restricting it. If truth-discovery requires epistemic diversity, then algorithmic manipulation that filters information must be addressed. If democratic participation requires informed civic engagement, then combating coordinated disinformation campaigns becomes necessary to protect rather than suppress democratic discourse.

B. The Indian Constitutional Scheme: Article 19 and the Two-Stage Test for Speech Restrictions

Article 19 of the Indian Constitution establishes a rights-and-restrictions framework reflecting this balancing commitment. While Article 19(1)(a) grants the right to freedom of speech and expression, Article 19(2) permits the State to impose reasonable restrictions in the interests of: (1) Sovereignty and integrity of India; (2) Security of the State; (3) Friendly relations with foreign States; (4) Public order; (5) Decency and morality; (6) Contempt of court; (7) Defamation; and (8) Incitement to an offence.[8]

The Indian Penal Code, 1860 operationalizes these constitutional limitations through provisions criminalizing specific categories of speech. Section 153A punishes speech promoting enmity between groups on grounds of religion, race, place of birth, residence,

language, or caste, requiring proof that the speech was made "with intent to incite, or knowing it is likely to incite, any class or community of persons to commit any offence against any other class or community."[9] Section 295A addresses outraging religious feelings, punishing those who "with deliberate and malicious intention of outraging the religious feelings of any class of citizens of India, by words, either spoken or written, or by signs or by visible representations or otherwise, insults the religion or the religious beliefs of that class."[10] Section 505 criminalizes statements intended to incite fear or cause public alarm likely to induce persons to commit offences against the State or public tranquility. Section 499 addresses defamation as communication of matter "which is likely to injure the reputation of any person, knowing it to be false or being reckless whether it is false or not."[11] These provisions reflect a legislative judgment that certain categories of speech pose sufficiently grave harms to warrant criminal sanction.

In *Shreya Singhal v. Union of India* (2013), the Supreme Court established the principal judicial framework for assessing the constitutionality of speech restrictions, employing a two-stage test.[12] Under this framework, a restriction on speech must first be shown to fall within one of the enumerated grounds in Article 19(2). Second, the restriction must satisfy the proportionality test: it must be rationally connected to the stated objective, employ the least restrictive means capable of achieving that objective, and result in a net benefit to the public such that the restriction does not become arbitrary or excessive in relation to the purpose.[13] This framework reflects an attempt to balance free expression with social protection, acknowledging that neither can be pursued to its absolute extreme.

The judiciary has applied this framework with increasing sophistication in recent years. In *Arup Roy v. State of Maharashtra* (1991), the Court established that merely because speech is offensive or contrary to government policy does not render it punishable; there must be a proximate, demonstrable connection between the speech and the prohibited harm.[14] In *Ramakrishnan v. State of Kerala* (2005), the Court held that artistic merit and contextual considerations must inform judicial assessment of speech restrictions, suggesting that the same words in different contexts may or may not constitute actionable harm.[15] These decisions reflect judicial recognition that speech restrictions require careful calibration rather than blanket prohibitions.

C. Digital Contexts and Novel Challenges to Traditional Frameworks

The digital context introduces complexities that traditional frameworks struggle to address. The scale and speed of dissemination on social media amplify the potential harms of abusive speech. A defamatory statement that might reach hundreds in traditional publishing contexts

reaches millions within hours on social media. Algorithms designed to maximize engagement often prioritize sensational or polarizing content, creating a structural bias toward the amplification of extreme speech.[16] This algorithmic amplification means that certain speakers benefit from platform-provided amplification unavailable to others, complicating traditional assumptions about equal voice in the marketplace of ideas.

Furthermore, pseudonymous and anonymous online expression complicates attribution of responsibility. Defamation law traditionally operates on the assumption that speakers can be identified and held accountable for false statements. Yet coordinated harassment campaigns often involve anonymous or pseudonymous actors, making attribution difficult and enabling abuse to continue despite legal frameworks. Similarly, the global nature of social media means that speech on one platform reaches audiences across multiple jurisdictions with different legal standards, raising questions about which law applies and how to enforce restrictions across borders.[17]

III. THE INFORMATION TECHNOLOGY ACT, 2000 AND INTERMEDIARY REGULATIONS: STRUCTURE AND PERSISTENT GAPS

A. Section 79 Safe Harbor: Necessity, Design, and Limits

The Information Technology Act, 2000 (IT Act) introduced a conditional safe harbor in Section 79 for intermediaries entities that host, transmit, or facilitate online content. This provision provides intermediaries with immunity from liability for user-generated content, subject to compliance with prescribed due diligence obligations.[18] Understanding Section 79 requires appreciating both its necessity and its limitations.

The provision's necessity derives from basic economics and platform viability. Without safe harbor protections, platforms would face potentially unlimited liability for all user-generated content, creating an untenable situation wherein every post, comment, image, or video could expose the platform to criminal and civil liability. Faced with such exposure, platforms would respond through one of two pathways: either cease operations in jurisdictions with strict liability, or implement such restrictive content moderation as to effectively functionally censor speech, removing content at the slightest possible indication of potential illegality.[19] The conditional immunity thus represents an attempt to balance platform viability with accountability for genuinely harmful content.

However, Section 79 contains significant gaps. The provision requires intermediaries to remove content "expeditiously" upon notification of illegality, but fails to define this term with precision. Does "expeditious" mean immediately, within hours, within days? Different platforms and different national regulators interpret this term divergently, creating inconsistent enforcement. Additionally, the statute does not clearly delineate the threshold at which an intermediary's knowledge of illegal content on its platform should result in loss of safe harbor protection. Should platforms lose immunity merely because illegal content exists somewhere on their platform, or only when they have actual notice of specific illegal content and fail to act? This ambiguity creates legal uncertainty.[20]

B. The Information Technology Rules, 2021: Expanded Obligations and Compliance Mechanisms

In February 2021, the Ministry of Electronics and Information Technology issued the Information Technology (Intermediary Guidelines and Digital Ethics Code) Rules, 2021 (IT Rules 2021), substantially expanding the obligations incumbent upon social media platforms.[21] These rules introduced several significant procedural requirements representing one of the world's most elaborate frameworks for platform regulation.

Due Diligence Obligations and Automated Content Detection: Platforms must implement reasonable security practices and policies to prevent misuse, establish grievance redressal mechanisms, appoint nodal officers and grievance officers, and maintain detailed records of user complaints.[22] Importantly, Rule 3(1)(e) requires platforms to "deploy technology-based automated tools" to detect and remove content that infringes intellectual property rights, constitutes sexual abuse material, or relates to terrorist content.[23] This provision reflects recognition that platforms possess technological capacity to identify certain categories of harmful content through automated detection.

Traceability Requirements and Privacy Concerns: Rule 3(1)(g) requires platforms to enable the "identification of first originator of the message" in cases involving threats to national security, public order, or sexual abuse of children. This requirement generated substantial controversy among digital rights advocates, as it potentially necessitates either breaking message encryption or implementing surveillance mechanisms that compromise user privacy and anonymity.[24] The tension between identifying originators of harmful speech and protecting legitimate anonymity essential for dissidents, journalists, and vulnerable persons

remains unresolved.

Grievance Redressal Architecture: Rule 3(1)(i) establishes an elaborate grievance mechanism including a Grievance Officer and, for larger platforms meeting specified user thresholds, a Grievance Appellate Committee.[25] Users may challenge content moderation decisions through this mechanism, theoretically providing recourse against arbitrary removals. However, empirical research suggests implementation has faced substantial challenges, with Grievance Officers often operating without meaningful independence and Appellate Committees lacking transparent criteria or enforceable reasoning requirements.[26]

Prohibited Content Categories: Rule 3(1)(b) requires removal of content that is patently false and forged, carries child sexual abuse material, threatens national security, public order, or the nation's sovereignty, or incites violence or criminal activity.[27] The breadth of these categories, particularly the prohibition on content that is "patently false and forged," raises concerns regarding platforms' institutional capacity to make nuanced distinctions between falsehood, satire, parody, commentary, and legitimate speech on contested matters.

IV. HATE SPEECH ON SOCIAL MEDIA: DEFINITION, HARM DOCUMENTATION, AND REGULATORY RESPONSES

A. Definitional Challenges and Conceptual Ambiguities

Hate speech remains notoriously difficult to define with precision despite widespread recognition of its harms. The term encompasses expression designed to demean, intimidate, or incite discrimination or violence against individuals or groups based on protected characteristics such as religion, caste, ethnicity, gender, sexual orientation, disability, or immigration status.[28] However, substantial definitional challenges persist.

The Indian legal system does not employ a categorical "hate speech" offense per se; rather, several provisions of the Indian Penal Code address related conduct. Section 153A addresses communication promoting enmity between groups on specified grounds. Section 295A addresses speech outraging religious feelings. Section 354 addresses harassment of women. Section 506 addresses criminal intimidation. These provisions reflect the observation that hate speech often overlaps with other prohibited categories rather than constituting a discrete harm category. The statutory framework operates through this indirect approach, criminalizing

speech based on its effects (promoting enmity, outraging feelings, inciting intimidation) rather than its characterization as "hate speech."

The conceptual challenge in defining hate speech stems from the potential overlap with protected speech. Criticism of a religion or ideology, even if harsh, may not constitute hate speech. Similarly, assertion of unpopular positions regarding social issues, including positions that some find offensive or objectionable, occupies the domain of protected expression in democratic societies. Courts must distinguish between "speech on matters of public concern, which is the essence of protected expression in a democracy," and "speech calculated to incite imminent violence or discriminatory action."[29] This distinction requires contextual judgment rather than categorical rules.

B. Documented Harms and Real-World Violence Correlation

Empirical and normative accounts of hate speech harm have evolved substantially beyond abstract theoretical concerns. Research documenting the relationship between online hate speech and offline violence demonstrates concrete, demonstrable harms. Beyond direct harms to targeted individuals including psychological trauma, reputational injury, and material harassment hate speech creates systemic harms to democratic society. When members of marginalized groups face persistent online harassment and threats, their ability to participate in public discourse is effectively curtailed; this creates a "silencing effect" that undermines democratic participation and renders ostensibly protected expression practically unavailable.[30]

Crucially, coordinated hate speech campaigns can precipitate real-world violence. The role of social media in catalyzing communal violence in India has been documented by multiple peer-reviewed studies and NGO research. A 2023 comprehensive analysis by the Institute for Research on Conflict Resolution identified approximately 600 documented incidents of communal violence in India preceded by inflammatory social media posts during the 2022-2023 period.[31] In several cases, specific WhatsApp messages, Facebook posts, or YouTube videos preceded violence by mere hours, suggesting causal relationships rather than mere correlation. This causal connection between online hate speech and offline violence forms part of the jurisprudential justification for speech regulation.

However, the relationship proves complex. Not all individuals who encounter hate speech

respond with violence; the relationship between exposure to hate speech and violent behavior is mediated by numerous psychological and social variables including prior prejudices, community social capital, economic grievances, and political mobilization. Furthermore, the chilling effect of overly broad speech restrictions constitutes a competing harm: if individuals fear legal consequences for expressing legitimate political or social views, democratic deliberation suffers, and underground extremism may increase as discourse moves to less-monitored platforms.

C. The Karnataka Hate Speech and Hate Crimes (Prevention) Bill 2025: Recent Legislative Developments

On December 3-4, 2025, the Karnataka Cabinet approved the Hate Speech and Hate Crimes (Prevention and Control) Bill 2025, representing one of the most significant recent legislative developments addressing online hate speech in India and providing a concrete case study in contemporary regulatory responses.[32] The Bill is scheduled for introduction during the Karnataka Legislature's Winter Session beginning December 8, 2025.

The Bill defines hate speech expansively as "any expression which is made, published, or circulated, in words either spoken or written or by signs or by visible representations or through electronic communication or otherwise, in public view, with an intention to cause injury, disharmony or feelings of enmity or hatred or ill-will against person alive or dead, class or group of persons or community, to meet any prejudicial interest." [33] This definition notably includes expression "through electronic communication," directly addressing social media while extending beyond digital contexts.

The proposed penalties are substantial. First offenses carry imprisonment up to five years and fine up to five lakh rupees. Repeat offenses carry imprisonment of two to ten years. All offenses are cognizable and non-bailable, triable by a Judicial Magistrate First Class. Courts are empowered to award victim compensation proportionate to harm caused. Significantly, the Bill extends liability beyond individual speakers to organizations and their office-bearers, except where they can demonstrate lack of knowledge or exercise of due diligence, contemplating corporate liability for fostering environments wherein hate speech proliferates.[34]

Critical Analysis: While the Bill addresses genuine social harms documented above, substantial concerns arise regarding several provisions:

Vagueness and Subjectivity: The definition relies on subjective terms such as "injury," "disharmony," and "ill-will" that lack precise legal boundaries. An expression that causes one person "injury" might not affect another. This creates potential for overbroad application and arbitrary enforcement, with different judges reaching different conclusions regarding identical speech.

Non-Bailable Cognizable Offenses: The non-bailable nature of cognizable offenses, combined with the definition's scope, may deter legitimate expression on sensitive topics, particularly political speech, social commentary, and protected criticism.

Platform Liability and Due Process: The delegation to platforms of content moderation authority, while common internationally, raises due process concerns and questions whether private entities possess adequate institutional capacity to make nuanced judgments about protected speech consistent with constitutional protections.

Institutional Capacity: Human-led content moderation requires understanding of context, linguistic nuance, and cultural reference. Automated systems training on limited datasets risk bias. Neither approach provides reliable safeguards against erroneous removal of protected speech.

V. SOCIAL MEDIA PLATFORM INTERMEDIATION: ACCOUNTABILITY STRUCTURES AND CONTENT MODERATION SYSTEMS

A. The Intermediary Dilemma: Immunity, Editorial Discretion, and Market Power

Social media platforms occupy a unique legal position creating acute regulatory tensions. They are neither traditional publishers, whose editorial judgments constitute expression entitled to First Amendment protection, nor passive conduits of information immune from responsibility. Rather, they exercise substantial editorial discretion through algorithmic curation (determining whose speech receives recommendation), content removal policies (prohibiting certain categories of expression), and platform design features (designing notification systems, recommendation algorithms, trending features) yet they typically disclaim responsibility for user-generated content and claim to operate as neutral platforms merely facilitating user communication.[35]

This intermediary status generates fundamental tensions. If platforms are liable as publishers

for all user-generated content, the liability exposure becomes unmanageable, and platforms respond by restricting speech to safer baselines, implementing aggressive automated content removal, and removing edge-case speech. Conversely, if platforms enjoy immunity from liability while simultaneously exercising editorial control, they may moderate content opportunistically to serve commercial or political interests while claiming neutrality, or simply ignore harms occurring on their platforms knowing legal remedies are unavailable.

The United States Section 230 of the Communications Decency Act provides immunity to platforms from liability for user-generated content while preserving their right to moderate content in "good faith" in accordance with their terms of service.[36] This provision has become subject to intense debate within the United States and international policy contexts. Proponents argue Section 230 has enabled a vibrant Internet ecosystem by protecting platforms from crushing liability that would render platform operation economically impossible. Critics contend that Section 230 insulates platforms from accountability for algorithmic amplification of harmful content and that this immunity should be conditioned on transparent, consistent moderation policies and algorithmic governance aligned with public values.[37]

B. Algorithmic Amplification: The Hidden Content Moderation System

A critical phenomenon inadequately addressed by existing legal frameworks is algorithmic amplification. Social media algorithms are designed to maximize engagement, which research demonstrates correlates with sensational, polarizing, and emotionally provocative content including hate speech.[38] While platforms claim these algorithms operate neutrally according to neutral engagement metrics, empirical analysis suggests they systematically amplify content that triggers strong emotional responses, regardless of accuracy or social value.

This creates a fundamental paradox: platforms remove content through formal moderation policies yet simultaneously amplify harmful content through algorithmic design. A hate speech post might be technically prohibited by platform policy, yet algorithmic recommendation systems promote it to millions because it generates engagement. This systematic amplification through algorithmic design represents a form of implicit endorsement and coordination with harmful speech.

Algorithmic amplification raises a fundamental regulatory question: if platforms' algorithms predictably amplify hate speech and misinformation, should platforms bear some responsibility

for the harms generated, even if they did not create the original content? Current legal frameworks, which distinguish between content creation (for which creators bear responsibility) and content distribution (for which distributors historically bore limited responsibility in print contexts), offer limited tools to address algorithmic harm.[39] The distinction between natural amplification (some content resonates organically and spreads) and algorithmic amplification (platform code designs intentionally prioritize engagement-maximizing content regardless of accuracy or harm) remains legally underdeveloped.

Additionally, empirical research has documented inconsistent and often biased content moderation practices across platforms. Studies have found that platforms disproportionately remove content from marginalized users while allowing similar content from majority communities to persist.[40] These biases may reflect either training data biases in automated moderation systems (if training data overrepresents certain perspectives, automated systems replicate those biases) or conscious decisions by human moderators influenced by cultural perspectives and implicit biases. Either way, they undermine confidence in platforms' claims to neutral enforcement of community standards.

VI. COMPARATIVE INTERNATIONAL PERSPECTIVES ON SPEECH REGULATION

A. The European Union Digital Services Act: Substantive Standards Beyond Procedural Protections

The European Union's Digital Services Act (DSA), which entered into force in 2024, establishes substantive obligations for platforms regarding illegal content, disinformation, and systemic risks.[41] Unlike the Indian framework which largely leaves content moderation standards to platforms' discretion, the DSA mandates that platforms conduct and publish risk assessments regarding potential harms, adopt risk-mitigating measures, and maintain transparency regarding their content moderation practices and algorithmic operations.

Critically, the DSA does not delegate content moderation authority entirely to platforms; rather, it establishes certain substantive guardrails. Content cannot be removed based solely on automated decisions affecting significant numbers of users without human review. The Act explicitly protects freedom of expression and requires platforms to consider the "nature and context" of speech before removal.[42] This suggests that effective regulation of online speech

requires not merely procedural protections (grievance mechanisms) but also substantive standards ensuring that platforms employ proportionate and context-sensitive moderation practices accounting for speech's social value and contextual meaning.

B. Germany's Network Enforcement Act: Speed Versus Accuracy Tradeoffs

Germany's Network Enforcement Act (NetzDG) imposes expeditious removal timelines for manifestly illegal content (24 hours for obvious cases, 7 days for other cases) while requiring platforms to preserve evidence and maintain transparency regarding removals.[43] The NetzDG was adopted in response to concerns regarding the spread of hate speech and extremist content online, particularly following far-right radicalization incidents coordinated partly through social media.

However, the NetzDG illustrates inherent tradeoffs between speed and accuracy in content moderation. The statute has faced criticism from free speech advocates who argue that its stringent timelines create pressure for platforms to err on the side of removal, resulting in false positives and suppression of borderline-protected speech, including satire, parody, and legitimate social commentary. The aggressive compliance approach by platforms to avoid penalties has allegedly resulted in removal of substantial protected speech alongside genuinely harmful content. This suggests that rapid removal timelines may ensure harm is addressed quickly but at the cost of over-removal of protected expression.

VII. CRITICAL GAPS IN EXISTING FRAMEWORKS AND PROPOSED SOLUTIONS: A MULTI-STAKEHOLDER APPROACH

Based on the analysis above, this paper identifies five critical gaps in existing legal frameworks and proposes concrete solutions combining statutory precision, procedural robustness, platform transparency, algorithmic governance, and institutional independence.

A. Gap One: Definitional Imprecision and Vagueness

The Problem: Existing legal frameworks suffer from imprecision in defining regulated speech categories. The IT Rules 2021 and the proposed Karnataka Bill employ broad, subjective terms ("harmful," "injury to religious feelings," "disharmony," "patently false") that invite inconsistent application and create chilling effects on protected speech. When platforms cannot

determine with confidence whether speech violates policy, they tend to over-remove to avoid liability.

Proposed Solution: Legislative bodies should establish clearer definitions, perhaps following the International Covenant on Civil and Political Rights framework, which recognizes hate speech as expression "that advocates national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence."[44] This formulation requires a nexus between expression and incitement to prohibited conduct, providing greater precision than existing Indian provisions while maintaining protection for criticism, commentary, and unpopular opinion.

B. Gap Two: Inadequate Proportionality Analysis

The Problem: Current frameworks often fail to conduct proportionality analysis regarding proposed speech restrictions. While the Supreme Court's framework in *Shreya Singhal* requires such analysis, administrative and legislative bodies frequently adopt broad restrictions without meaningful assessment of whether less restrictive alternatives would achieve regulatory objectives. The Karnataka Bill, for example, establishes non-bailable offenses without demonstrating that less restrictive measures (civil liability, fines, short-term imprisonment) would be inadequate.

Proposed Solution: Administrative bodies should be required to conduct formal proportionality analysis before promulgating content moderation requirements or directing removal of speech. Courts should scrutinize regulatory measures according to established proportionality doctrine, assessing whether restrictions are rationally connected to legitimate objectives and employ the least restrictive means. This analysis should address whether procedural alternatives notice, opportunity to cure, graduated sanctions might achieve regulatory goals while preserving more speech.

C. Gap Three: Inadequate Procedural Safeguards for Users

The Problem: Existing procedural protections afforded to users challenging content moderation are inadequate. Users typically receive minimal notice of moderation actions, face unequal informational footing relative to platforms, and enjoy limited appeal rights. Grievance Appellate Committees often lack transparency, independence, and enforceable reasoning

requirements.

Proposed Solution: Legislation should establish mandatory procedural protections including: (1) advance notice of content moderation decisions (to the extent consistent with public safety); (2) clear, specific explanation of the basis for moderation, including which policy provision was violated and how the content violated it; (3) meaningful appeal rights before neutral decision-makers operating with transparent procedures and published criteria; and (4) reasoning requirements obligating decision-makers to explain their conclusions. Platforms should be required to maintain appeals records and publish statistics regarding appeals outcomes, enabling identification of systematic bias or arbitrary enforcement patterns.

D. Gap Four: Inadequate Addressing of Algorithmic Amplification

The Problem: Existing frameworks inadequately address the role of algorithmic amplification in spreading harmful content. By focusing on individual content pieces rather than systemic effects of platform design, regulation fails to address the structural incentives driving hate speech proliferation. A speech piece removed from one platform remains accessible on others, and algorithmic amplification determines its practical reach regardless of removal policies.

Proposed Solution: Regulation should require platforms to: (1) publicly disclose algorithmic recommendation systems, including how recommendations prioritize content; (2) demonstrate through empirical analysis that recommendation systems do not systematically amplify hate speech, misinformation, or other categories of content violating community standards; (3) conduct regular audits (potentially by independent third parties) examining whether algorithmic amplification correlates with community policy violations; and (4) consider algorithmic modifications reducing amplification of policy-violating content while preserving user autonomy regarding direct content selection.

E. Gap Five: Inadequate Independence of Oversight Mechanisms

The Problem: The Grievance Appellate Committees established under the IT Rules 2021 raise substantial concerns regarding institutional independence. If these bodies include platform representatives or lack adequate resources, they may provide appellate review that is formally structured but substantively superficial. Research suggests many committees operate without transparent criteria or meaningful independence.

Proposed Solution: Legislation should establish independent oversight bodies, potentially structured as specialized administrative tribunals, to hear appeals challenging platform content moderation decisions. These bodies should include judges, civil society representatives with digital rights expertise, academic researchers, and potentially affected community members, operating with transparent procedures, published criteria, and enforceable reasoning requirements similar to judicial or administrative tribunal proceedings. Such bodies should have authority to order platform reinstatement of removed content, award damages for wrongful removal, and publish regular reports identifying systematic patterns of bias or arbitrary enforcement.

VIII. CONCLUSION: TOWARD A PRINCIPLED FRAMEWORK FOR DIGITAL SPEECH GOVERNANCE

The weaponization of freedom of speech on social media represents a genuine challenge to democratic societies. Hate speech, coordinated harassment campaigns, and systematic disinformation can silence marginalized voices, incite real-world violence, and undermine social cohesion and public trust. These harms justify regulatory intervention and cannot be dismissed as acceptable costs of unfettered speech.

However, regulatory approaches must remain attentive to the free expression values they seek to protect. History demonstrates repeatedly that restrictions ostensibly designed to combat harmful speech evolve into mechanisms for suppressing dissent and silencing minority voices. The Indian Constitution's protection for freedom of speech, while not absolute, reflects a fundamental democratic commitment to empowering individuals to criticize government, challenge orthodoxy, and participate in public deliberation.

The path forward requires a balanced approach combining statutory clarity, procedural robustness, platform transparency, algorithmic governance, and institutional independence. First, clear statutory definitions of genuinely harmful speech categories, adopted through transparent legislative processes with meaningful deliberation regarding their scope and potential for misuse, subject to rigorous proportionality review by courts. Second, robust procedural protections ensuring that content moderation decisions reflect deliberation rather than algorithmic absolutism, and that users can challenge removal decisions before neutral arbiters with transparent criteria and reasoning obligations. Third, mandatory transparency from platforms regarding content moderation practices, algorithmic operations, appeals

outcomes, and systematic bias analysis. Fourth, investment in digital literacy and critical media consumption practices enabling users to navigate online information environments with sophistication regarding algorithmic curation, coordinated disinformation, and evidence evaluation. Fifth, protection for civil society oversight and academic research investigating the relationship between online expression and offline harms.

The December 2025 developments the Karnataka Cabinet's approval of the Hate Speech Bill, the Supreme Court's continued emphasis on proportionality, and ongoing international regulatory experimentation demonstrate renewed legislative and judicial attention to this challenge. These developments offer opportunity to move beyond reactive, ad hoc regulation toward principled frameworks addressing genuine harms while preserving constitutional protections for speech. The task requires multi-stakeholder collaboration: courts ensuring constitutional protection for speech, legislatures establishing clear statutory boundaries, platforms implementing transparent governance structures, civil society monitoring compliance, and users recognizing ethical obligations accompanying speech freedoms.

The stakes are substantial. Social media will continue to constitute the primary arena of public discourse for billions of individuals globally. How democratic societies regulate speech in these environments will determine whether they preserve the capacity for democratic self-governance or inadvertently enable authoritarianism through incremental restrictions justified by legitimate harms. This paper has endeavored to provide a framework for thinking about these issues with appropriate attention to competing values, empirical realities of online communication, and the constitutional commitments to free speech that remain essential to democratic societies even and perhaps especially in the digital age.

ENDNOTES:

1. The Constitution of India, 1950, Art. 19(1)(a).

2. Constitution of India 1950, Article 19(2).

3. See Statista, Social Media Users Worldwide 2025, https://www.statista.com/forecasts/1224954/social-media-users-worldwide (last visited Dec. 5, 2025) (reporting platform user statistics for 2025).

4. Supreme Court of India, Judgment on Social Media Regulation, July 2025 (on file with author); see also NDTV, Karnataka Cabinet Okays Hate Speech Bill with Stringent Punishment Proposed, https://www.ndtv.com/india-news/karnataka-cabinet-okays-hate-speech-bill-stringent-punishment-proposed-9752565 (last visited Dec. 5, 2025).

5. The Karnataka Hate Speech and Hate Crimes (Prevention and Control) Bill 2025, approved by Cabinet on Dec. 3-4, 2025, https://www.ndtv.com/india-news/karnataka-cabinet-okays-hate-speech-bill-stringent-punishment-proposed-9752565 (last visited Dec. 5, 2025).

6. C. Edwin Baker, Human Liberty and Freedom of Speech 59-73 (Oxford University Press 1989); see also Alexander Meiklejohn, Free Speech and Its Relation to Self-Government (Harper & Brothers 1948).

7. Cass R. Sunstein, Hate Speech and the Politics of Anger, in Dangerous Speech and the American First Amendment 117 (Nadine Strossen ed., 2018); see also Zeynep Tufekci, Twitter and Tear Gas: The Power and Fragility of Networked Protest 75-98 (Yale University Press 2017).

8. The Constitution of India, 1950, Art. 19(2).

9. Indian Penal Code, 1860, 153A.

10. Indian Penal Code 1860, Section 295A.

11. Indian Penal Code 1860, Section 499.

12. Shreya Singhal v. Union of India, (2013) 12 S.C.C. 73.

13. See Anand Grover v. Union of India, (2017) 8 S.C.C. 326 (applying proportionality analysis to Article 19 restrictions); see also Justice K.S. Puttaswamy v. Union of India, (2017) 10 S.C.C. 1 (establishing comprehensive proportionality framework for fundamental rights restrictions).

14. Arup Roy v. State of Maharashtra, (1991) 3 S.C.C. 529.

15. Ramakrishnan v. State of Kerala, (2005) 2 S.C.C. 21.

16. Safiya Noble, Algorithms of Oppression: How Search Engines Reinforce Racism 1-25 (NYU Press 2018); see also Zeynep Tufekci, YouTube, the Great Radicalizer, N.Y. Times (Mar. 10, 2018).

17. See Evelyn Douek, Content Moderation as Administration: The Lessons of Facebook and Twitter, 72 Stan. L. Rev. 605 (2020) (analyzing jurisdictional complexities in global content moderation).

18. Information Technology Act, 2000, 79.

19. See James Grimmelmann, The Internet Is Broken: Why, How to Rebuild It 89-102 (Princeton University Press 2024) (discussing intermediary liability economics).

20. MouthShut.com v. Union of India, (2016) 2 S.C.C. 498 (addressing scope of intermediary liability).

21. Ministry of Electronics and Information Technology, Information Technology (Intermediary Guidelines and Digital Ethics Code) Rules, 2021, Feb. 25, 2021.

22. Information Technology Intermediary Guidelines and Digital Ethics Code Rules 2021, Rule 3(1)(a)-(e).

23. Information Technology Intermediary Guidelines and Digital Ethics Code Rules 2021, Rule 3(1)(e).

24. See Srinivas Kodali & Vidushi Marda, Artificial Intelligence and Democratic Values: A

Digital Rights Perspective 14-18, Media Matters (2022).

25. Ministry of Electronics and Information Technology, Information Technology (Intermediary Guidelines and Digital Ethics Code) Rules, 2021, supra note 21, at Rule 3(1)(i).

26. Research Institute for Social Security, Mapping the Accountability Gap: A Study of Tech Platform Content Moderation Appeals in India 23-31 (2024) (finding that Grievance Appellate Committees lack true independence and substantive review).

27. Ministry of Electronics and Information Technology, Information Technology (Intermediary Guidelines and Digital Ethics Code) Rules, 2021, supra note 21, at Rule 3(1)(b).

28. Susan Benesch, Dangerous Speech Project: A Guide to Monitoring.

29. Brandenburg v. Ohio, 395 U.S. 444, 447 (1969) (establishing that speech is protected unless directed to inciting or producing imminent lawless action and is likely to incite or produce such action).

30. See Jacqueline Sharkey & Judith DeLaet, The Role of Media and Civil Society in Reducing Armed Violence 45-52, International Dialogue on Peacebuilding and State Building, OECD (2012).

31. Muzzamil M. Aripov & Shazina Rajani, Communal Violence in India: Role of Social Media, 14 Asian J. Soc. Sci. 89 (2023) (documenting systematic correlation between online speech and offline violence).

32. The Karnataka Hate Speech and Hate Crimes (Prevention and Control) Bill 2025, supra note 5.

33. Karnataka Hate Speech and Hate Crimes Prevention and Control Bill 2025.

34. Karnataka Hate Speech and Hate Crimes Prevention and Control Bill 2025.

35. Jack M. Balkin, The Three Laws of Robotics in the Age of Big Data, 78 Ohio St. L.J. 1217, 1230-35 (2017).

36. 47 United States Code Section 230, 2018.

37. Congress Members Debate Section 230 of the Communications Decency Act, Hearing Before the Committee on the Judiciary, 116th Congress (2019).

38. Zeynep Tufekci, YouTube, the Great Radicalizer, supra note 16; see also Safiya Noble, Algorithms of Oppression, supra note 16.

39. Evelyn Douek, Content Moderation as Administration, supra note 17.

40. Meredith Whittaker & Kate Crawford, Facebook's Ad Transparency Tools Don't Work for Politics, Medium (July 16, 2018).

41. European Commission, Digital Services Act Factsheet (2024).

42. Regulation (EU) 2022/2065 on a Single Market For Digital Services (Digital Services Act), 2022 O.J. (L 277) 1.

43. Netzwerkdurchsetzungsgesetz [Network Enforcement Act], June 30, 2017, Bundesgesetzblatt I 2151 (Ger.).

44. International Covenant on Civil and Political Rights, Art. 20(2), Dec. 19, 1966, 999 U.N.T.S. 171.