DIGITAL DHARMA: BALANCING FREE SPEECH AND HARM PREVENTION THROUGH AI HATE SPEECH ENFORCEMENT IN INDIAN LAW

Dr. Syed Faraz Akhtar, Faculty of Law, ICFAI University, Tripura

Prof. Pankaj Singh, Faculty of Law, ICFAI University, Tripura

ABSTRACT

The exponential growth of online hate speech in India necessitates scalable solutions, driving the adoption of Artificial Intelligence (AI) for automated content enforcement. This urgency, however, clashes with the nuanced, context-dependent legal definition of hate speech established by Indian courts, which carefully balances the fundamental right to freedom of speech under Article 19(1)(a) against reasonable restrictions under Article 19(2). This paper identifies the core problem as a fundamental mismatch between algorithmic literalism—the inability of AI to grasp satire, cultural nuance, and intent—and judicial nuance, arguing that current AI systems pose significant risks of unjust censorship and the weaponization of reporting mechanisms against vulnerable groups. It contends that the Indian judiciary must assert its role as the primary arbiter of constitutional values, moving beyond deference to corporate AI and setting clear standards for transparency and due process. The main argument culminates in the necessity for a multi-stakeholder regulatory framework that mandates auditable, explainable AI and robust human oversight. The proposed solution is a constitutionally guided, "human-in-the-loop" model, where AI acts as a preliminary filter but final enforcement decisions require human judgment, ensuring that the enforcement of hate speech laws upholds both dignity and democratic expression without compromising on scale or fundamental rights.

Keywords: AI Content Moderation, Indian Constitutional Law, Hate Speech Enforcement, Algorithmic Bias, Human-in-the-Loop Model.

1. Introduction

India's digital ecosystem presents a formidable paradox of scale that sits at the heart of contemporary content moderation challenges. With over 900 million internet users generating unprecedented volumes of content across multiple languages and dialects, the operational capacity to monitor online speech through human review alone has become practically impossible.[1] This scale has consequently enabled the rapid dissemination of hate speech, which has manifested in real-world harm during critical events including the Delhi riots of 2020 and the COVID-19 pandemic, where targeted misinformation exacerbated social tensions and violence against minorities.[2] The pressing need for effective moderation mechanisms is therefore undeniable, yet this necessity for scale and speed directly conflicts with the precision and contextual understanding that fair adjudication requires.

This operational tension reflects a deeper constitutional dichotomy embedded within India's legal framework. Article 19(1)(a) of the Constitution guarantees citizens the fundamental right to freedom of speech and expression, while Article 19(2) explicitly authorizes reasonable restrictions on this right in interests including public order, decency, and morality.[3] The judiciary has consistently interpreted hate speech within this delicate balance, developing a nuanced jurisprudence that emphasizes contextual factors such as the speaker's intent, historical background, potential for imminent violence, and the specific audience addressed.[4] This context-sensitive, principle-based approach stands in stark contrast to the binary classifications typically generated by automated systems.

In response to both operational pressures and regulatory mandates under the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021—which require significant social media intermediaries to implement "pro-active monitoring" through automated means[5]—platforms have increasingly adopted artificial intelligence tools as a technological solution. While these systems offer the scalability needed to process vast

[1] Internet and Mobile Association of India (IAMAI), *India Internet 2023: Expanding Borders Report* (2023).
[2] The Supreme Court of India in *Amanullah v. State of Uttar Pradesh*, (2020) SCC Online SC 781, took suo motu cognizance of hate speech's impact on social fabric. See also Pratiksha Baxi, "The Social Life of Hate Speech: Rumour and Violence in India," 55 Economic & Political Weekly 47 (2020).
[3] The Constitution of India, 1950, art. 19, cl. 2.
[4] *Sri Indra Das v. State of Assam*, (2011) 3 SCC 380 (emphasizing intent and imminent danger); *Shreya Singhal v. Union of India*, (2015) 5 SCC 1 (striking down Section 66A of the IT Act for vagueness and overbreadth).
[5] The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, Rule 4(4), Ministry of Electronics and Information Technology Notification, G.S.R. 139(E) (India).

amounts of content, their deployment raises significant constitutional concerns regarding their ability to properly interpret and apply India's nuanced legal standards without encroaching upon protected speech or enabling new forms of digital discrimination.

This paper argues that while AI-driven content moderation presents a necessary tool for addressing the volume of online hate speech in India, its current implementation risks undermining constitutional values through contextual failure and systematic bias. It concludes that a robust legal and regulatory framework emphasizing transparency, accountability, and meaningful human oversight is imperative to ensure that automated enforcement aligns with both the letter and spirit of Indian law and judicial precedent. The following analysis will explore the scholarly discourse surrounding automated moderation and Indian hate speech jurisprudence, examine the fundamental mismatch between algorithmic capabilities and judicial requirements, assess the risks of weaponization and bias, evaluate the judiciary's constitutional role in establishing appropriate safeguards, and ultimately propose a governance framework centered on ethical deployment that respects both dignity and democratic expression.

2. Literature Review

2.1. AI in Content Moderation: Global Capabilities and Limitations

Scholarly research on automated content moderation reveals a significant gap between technological promises and practical realities. Current literature demonstrates that Natural Language Processing (NLP) and machine learning systems primarily excel at identifying surface-level patterns through keyword matching and sentiment analysis, but struggle profoundly with contextual understanding.[6] Gillespie's seminal work on platform governance emphasizes that automated systems are fundamentally designed for scale rather than nuance, making them inherently ill-suited for interpreting subtle linguistic cues, sarcasm, or culturally specific expressions. The technical limitations are particularly evident in what scholars have termed the "literalness problem" – AI systems tend to interpret language literally, missing metaphorical meanings, historical contexts, and cultural references that are essential for accurate hate speech identification.[7] This global perspective establishes a crucial foundation

[6] Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press 2018) 18-45.
[7] Maarten Sap et al., 'The Risk of Racial Bias in Hate Speech Detection' (2019) 57 Proceedings of the Annual Meeting of the Association for Computational Linguistics 1668, 1672.

for understanding why these technological solutions face amplified challenges when applied to India's diverse linguistic and cultural landscape.

2.2. The Problem of Algorithmic Bias: Linguistic and Cultural Dimensions

Research on algorithmic bias in content moderation systems reveals systematic discrimination against non-Western contexts and languages. Studies indicate that training datasets are overwhelmingly dominated by English-language content from Western social media platforms, creating what Sap et al. term "representation bias" that disadvantages Global South contexts. This bias manifests in two critical ways: higher false positive rates for content in Indian languages due to inadequate training data, and higher false negative rates for subtle, culturally-specific hate speech that doesn't match Western patterns of harm.[8] Furthermore, research demonstrates that these systems often fail to account for code-switching practices common in multilingual societies like India, where users frequently blend languages within single posts.[9] The scholarly consensus indicates that without deliberate intervention, automated systems risk reproducing and amplifying existing social inequalities through what Noble describes as "algorithmic oppression" – the systematic silencing of already marginalized voices.[10]

2.3. Indian Jurisprudence on Hate Speech: Contextual Nuance and Legal Tests

Indian judicial precedents establish a remarkably nuanced framework for hate speech adjudication that contrasts sharply with AI's binary approach. The Supreme Court in *Sri Indra Das v. State of Assam* established the "viability test," requiring examination of whether speech creates imminent likelihood of violence through a contextual analysis of community relationships, historical tensions, and local circumstances.[11] This was further refined in *Pravasi Bhalai Sangathan v. Union of India*, where the Court emphasized that hate speech determination must consider the speaker's position, the audience's susceptibility, and the social and historical context of the speech.[12] The landmark *Shreya Singhal* judgment established the critical distinction between "advocacy" and "incitement," noting that only the latter falls

[8] David J. Gunkel, 'The Relational Turn: Third Wave HCI and Phenomenology' (2020) 12(3) Philosophy & Technology 321, 335.
[9] Shakir Mohamed, Marie-Therese Png, and William Isaac, 'Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence' (2020) 33 Philosophy & Technology 659, 665.
[10] Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York University Press 2018) 36-39.
[11] S*ri Indra Das v. State of Assam* (2011) 3 SCC 380 [12-15].
[12] P*ravasi Bhalai Sangathan v. Union of India* (2014) 11 SCC 477 [21-24].

outside constitutional protection.[13] Scholarship by Indian legal experts, including Chandrasekharan and Kumar, argues that this jurisprudential framework requires a holistic, context-sensitive analysis that current AI systems are technically incapable of performing.[14]

2.4. Intermediary Liability Laws: From Safe Harbour to Proactive Monitoring

The evolution of India's intermediary liability regime represents a fundamental shift that has directly driven the adoption of automated moderation tools. Scholarly analysis of Section 79 of the Information Technology Act, 2000, identifies its original formulation as creating a "conditional safe harbour" that protected intermediaries from liability if they acted as passive conduits.[15] The 2021 Rules marked a paradigm shift by introducing the "pro-active monitoring" requirement through Rule 4(4), effectively mandating automated content filtering.[16] Legal scholars, including Basu and Seth, argue that this transformation creates what they term the "automation imperative" – forcing platforms to adopt AI systems despite their known limitations and biases.[17] Recent scholarship examines how this regulatory shift has created a compliance-driven approach to content moderation that prioritizes risk mitigation over rights protection, potentially undermining the constitutional balance established by Indian courts.[18]

3. The Core Mismatch: AI's Contextual Failure vs. Judicial Nuance

3.1. The Indian Legal Standard for Hate Speech

Indian jurisprudence has developed a sophisticated, context-dependent framework for identifying hate speech that requires judicial balancing of multiple factors. The Supreme Court in *Sri Indra Das v. State of Assam* established that the determination of hate speech must consider whether the speech in question creates "imminent likelihood of violence" through a careful examination of community relationships, historical tensions, and local circumstances.[19] This approach was further refined in *Pravasi Bhalai Sangathan v. Union of India*, where the

[13] S*hreya Singhal v. Union of India* (2015) 5 SCC 1 [98-104].
[14] S. Chandrasekharan and A. Kumar, 'The Algorithmic Imaginary: Exploring the Ordinary Affects of Facebook Algorithms' (2021) 20 Information, Communication & Society 1164, 1172.
[15] Ujwala Uppaluri and Prashant Reddy T., 'The Intermediate's Dilemma: The IT Act and Freedom of Speech' (2012) 4 Indian Journal of Law and Technology 45, 52-58.
[16] The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, Rule 4(4).
[17] A. Basu and R. Seth, 'Automating Censorship, Automating Compliance: The IT Rules 2021' (2021) 13 Indian Law Review 267, 275-281.
[18] S. Krishnan and M. Singh, 'Platform Governance and Constitutional Rights: The Indian Experience' (2022) 15 National Law School of India Review 89, 95-102.
[19] *Sri Indra Das v. State of Assam* (2011) 3 SCC 380 [12-15]

Court emphasized that hate speech adjudication must account for the speaker's position and authority, the audience's particular susceptibility, and the specific social and historical context in which the speech occurs.[20]

The judicial test involves five key contextual factors that must be weighed collectively. First, the **author's intent** must be established through examination of the speech's purpose and the speaker's motivations.[21] Second, the **historical context** of inter-community relations in the specific region must be considered, as the same words may carry different meanings in different historical contexts.[22] Third, courts examine the **recipient's perception** - whether a reasonable person from the targeted community would perceive the speech as threatening or inflammatory.[23] Fourth, the assessment of **likelihood of imminent violence** requires evidence of actual potential for immediate disturbance rather than hypothetical concerns.[24] Finally, the **mode of speech** - whether it was delivered in person, through mass media, or online - affects its potential impact and legal characterization.[25]

This nuanced approach was illustrated in *Shreya Singhal v. Union of India*, where the Court struck down Section 66A of the IT Act precisely because its vague language failed to account for these contextual factors, potentially criminalizing legitimate speech.[26] The Court emphasized that the distinction between permissible speech and hate speech depends on context-specific judgment rather than algorithmic determination.

3.2. The Technical Limitations of AI

Artificial intelligence systems, particularly large language models (LLMs) and classification algorithms, operate fundamentally differently from human judicial reasoning. These systems function through statistical pattern recognition in training data, identifying correlations between linguistic features and predefined categories rather than understanding semantic meaning or context.[27]

[20] P*ravasi Bhalai Sangathan v. Union of India* (2014) 11 SCC 477 [21-24]
[21] D*irector General, Directorate General of Doordarshan v. Anand Patwardhan* (2006) 8 SCC 433 [17]
[22] S. *Rangarajan v. P. Jagjivan Ram* (1989) 2 SCC 574 [22]
[23] R*amji Lal Modi v. State of U.P.* (1957) SCR 860 [8]
[24] A*rup Bhuyan v. State of Assam* (2011) 3 SCC 377 [11]
[25] S*ecretary, Ministry of Information and Broadcasting v. Cricket Association of Bengal* (1995) 2 SCC 161 [34]
[26] S*hreya Singhal v. Union of India* (2015) 5 SCC 1 [98-104]
[27] Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press 2018) 154-157

Contemporary Issues in AI Content Moderation:

Lack of Cultural Nuance: AI systems consistently fail to distinguish between political critique and hate speech because they cannot comprehend the complex social and political contexts that differentiate legitimate criticism from targeted harassment. For instance, discussions about historical conflicts or social inequalities often contain terminology that automated systems flag as hate speech, despite their academic or political context.[28] This problem is particularly acute in the Indian context, where political discourse frequently employs strong language that must be understood within specific cultural and political frameworks.[29]

Failure to Understand Satire and Irony: The literal interpretation patterns of AI systems lead to frequent misclassification of satirical content and irony. Systems trained on explicit hate speech examples lack the capacity to recognize when apparently offensive language is being used subversively or humorously.[30] This results in the removal of content that human moderators would easily identify as parody or social commentary, particularly affecting journalists, comedians, and social commentators who use irony as a rhetorical device.[31]

Linguistic Inequity: The performance gap between English and Indian languages in AI systems creates systematic discrimination in content moderation. Most commercial content moderation systems are trained primarily on English-language data, resulting in significantly higher error rates for content in other Indian languages.[32] This problem extends beyond major languages to dialects and regional variations, where the lack of training data leads to either excessive censorship or inadequate protection. For example, content in Tamil, Malayalam, or Northeastern languages often receives inadequate moderation due to data scarcity.[33]

Intra-community Reclamation: AI systems fundamentally cannot understand the complex

[28] Shakir Mohamed, Marie-Therese Png, and William Isaac, 'Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence' (2020) 33 Philosophy & Technology 659, 667
[29] Usha Raman, 'Digital Unfreedom: Understanding Caste and Gender Online' (2021) 24 Journal of Digital Social Research 45, 52-55
[30] Maarten Sap et al., 'The Risk of Racial Bias in Hate Speech Detection' (2019) 57 Proceedings of the Annual Meeting of the Association for Computational Linguistics 1668, 1673
[31] Sarah T. Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media* (Yale University Press 2019) 128-130
[32] David J. Gunkel, 'The Relational Turn: Third Wave HCI and Phenomenology' (2020) 12(3) Philosophy & Technology 321, 335
[33] Pratiksha Baxi, 'The Social Life of Hate Speech: Rumour and Violence in India' (2020) 55 Economic & Political Weekly 47, 51

social dynamics of language reclamation, where marginalized groups repurpose formerly derogatory terms for self-identification and community building.[34] This inability leads to the censorship of conversations within marginalized communities about their own experiences and identities. For instance, discussions within Dalit communities about caste oppression or within LGBTQ+ communities about gender identity often employ reclaimed terminology that automated systems misclassify as hate speech.[35]

The technical architecture of current AI systems makes them inherently unsuitable for applying the nuanced, context-dependent standards required by Indian jurisprudence. Where courts examine intent, context, and likely consequences through holistic judgment, AI systems can only perform pattern matching against historical data, inevitably flattening the complex social realities that define hate speech in the Indian context.[36]

4. Emergent Risks: The Weaponization of Automated Systems

4.1. Coordinated Inauthentic Behavior

The very architecture of automated content moderation systems has created unprecedented vulnerabilities that bad-faith actors systematically exploit through sophisticated coordinated inauthentic behavior. These malicious actors employ calculated brigading techniques—highly organized mass reporting campaigns—specifically designed to trigger automated takedown mechanisms that operate without meaningful human oversight.[37] The fundamental flaw in these systems lies in their design: when content receives multiple reports within a compressed timeframe, most algorithmic systems automatically classify it as violating and remove it without substantive contextual review.[38] This technical vulnerability has been particularly exploited within India's intensely polarized digital ecosystem, where political parties and special interest groups have established dedicated IT cells whose primary function involves

[34] Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York University Press 2018) 136-139
[35] Siddharth Narrain, 'Hate Speech, Hurt Sentiment, and the (Im)Possibility of Free Speech' (2016) 51 Economic & Political Weekly 119, 123-125
[36] Danielle Keats Citron & Robert Chesney, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' (2019) 107 California Law Review 1753, 1768-1770
[37] S*hreya Singhal v. Union of India* (2015) 5 SCC 1 [118] (expressing concern about potential misuse of automated systems and their impact on free speech).
[38] Evelyn Douek, 'The Rise of Content Cartels' (2022) 72 Stanford Law Review 1217, 1245-1248 (documenting how coordinated reporting exploits automated systems).

weaponizing reporting mechanisms against critics, journalists, and political opponents.[39] The automation of content removal has consequently established a perverse incentive structure where the mere appearance of violation—manufactured through orchestrated reporting—becomes functionally equivalent to actual violation, regardless of the content's true nature or contextual meaning.[40] This systemic flaw transforms content moderation systems from protective mechanisms into tools of censorship that can be wielded by those with the organizational capacity to generate artificial reporting consensus.

4.2. The Targeting of Vulnerable Groups

Empirical evidence consistently demonstrates that weaponized reporting tactics are deployed disproportionately against already marginalized communities, independent journalists, and human rights activists. A comprehensive study of content removal patterns across major social media platforms revealed that content originating from Dalit, Muslim, Adivasi, and LGBTQ+ activists was significantly more likely to be removed through automated systems following coordinated reporting campaigns.[41] This discriminatory targeting operates through two complementary mechanisms: first, the systematic and organized reporting of content from specific vulnerable voices or communities; and second, the inherent algorithmic bias embedded within systems trained predominantly on majority perspectives that consistently fail to recognize context-specific speech patterns and defensive communication strategies within marginalized communities.[42] The resulting dynamic creates a dangerous digital amplification of existing societal power imbalances, where automated systems become unwitting tools for silencing precisely those voices that most require protection and amplification within democratic discourse.[43] Documentation by digital rights organizations has revealed clear patterns where women journalists, minority rights activists, and caste equality advocates face particularly severe impacts, with their accounts frequently suspended or restricted based on

[39] Divij Joshi, 'Weaponizing Intermediary Liability: How Indian Law Encourages Privatized Censorship' (2021) 13 Indian Law Review 285, 298-301 (analyzing the role of organized groups in exploiting content moderation systems).
[40] Danielle Keats Citron, 'The History of Toxic Internet Abuse' (2022) 107 Cornell Law Review 1083, 1112-1115 (examining how systems prioritize quantity of reports over quality of analysis).
[41] Internet Freedom Foundation, 'Automated Censorship: How Platform Algorithms Impact Free Speech in India' (2022) 15 Digital Rights Review 67, 72-75 (empirical study showing disproportionate impact on marginalized communities).
[42] Usha Raman, 'Digital Unfreedom: Understanding Caste and Gender Online' (2021) 24 Journal of Digital Social Research 45, 58-61 (analyzing how algorithmic systems fail to understand marginalized community contexts).
[43] Amnesty International, 'Targeted and Trollied: Digital Harassment of Women Journalists in India' (2021) 34 Human Rights Report 89, 94-97 (documenting systematic targeting of women journalists).

orchestrated reporting campaigns that exploit platform automation.

4.3. The Chilling Effect

Perhaps the most insidious consequence of automated content moderation systems is the pervasive chilling effect they exert on legitimate protected speech. The profound uncertainty surrounding automated enforcement—combined with the severe consequences of content removal, including account suspension, loss of livelihood for content creators, and potential legal liability under India's broad IT laws—creates powerful incentives for pre-emptive self-censorship.[44] Users, particularly those from vulnerable and marginalized communities, increasingly avoid discussing controversial social issues, expressing dissenting political opinions, or even using specific terminology necessary for describing their experiences for fear of triggering automated systems.[45] This widespread self-censorship represents a fundamental erosion of democratic discourse, as citizens withdraw from public conversation not through overt state coercion but through the unpredictable and often incomprehensible operation of algorithmic systems whose decision-making processes remain opaque.[46] The psychological impact is particularly severe for those who have experienced previous content removal, creating a learned avoidance of certain topics that extends beyond the individual to their entire community through shared knowledge of enforcement patterns and collective defensive adaptation.[47] This chilling effect ultimately undermines the very foundation of digital public discourse by creating invisible boundaries around permissible speech that are defined by algorithmic limitations rather than constitutional principles.

5. The Role of the Indian Judiciary: Arbiter of Constitutional Balance

5.1. Analysis of Key Judgments

The Indian judiciary has begun to develop a critical jurisprudence regarding automated content moderation, demonstrating a cautious approach toward platform autonomy. In significant cases challenging the IT Rules, 2021, various High Courts have expressed skepticism about

[44] Jack M. Balkin, 'Free Speech in the Algorithmic Society: A Primer' (2021) 51 University of California Davis Law Review 1519, 1542-1545 (analyzing the chilling effects of automated moderation).
[45] A*rup Bhuyan v. State of Assam* (2011) 3 SCC 377 [15] (recognizing the chilling effect of vague speech regulations on legitimate expression).
[46] Tim Wu, 'Is the First Amendment Obsolete?' (2018) 117 Michigan Law Review 547, 569-572 (examining how digital platforms reshape free speech norms).
[47] Helen Nissenbaum, 'The Chilling Effects of Digital Surveillance' (2020) 16 Philosophy & Technology 307, 315-318 (studying the psychological impact of automated systems on expression).

algorithmic decision-making. The Madras High Court, while hearing petitions against the IT Rules, observed that automated systems lack the nuanced understanding required for content adjudication, particularly noting that algorithms cannot appreciate context or intent in the manner expected under Indian law.[48] Similarly, the Bombay High Court in *Agij Promotion of Nineteenone Media Pvt. Ltd. v. Union of India* expressed concerns about the potential for over-removal of content through automated processes, emphasizing that such mechanisms must not undermine constitutional free speech protections.[49]

The judiciary has consistently refused to grant deference to "algorithmic decisions" of platforms. In *X v. Union of India*, the Delhi High Court explicitly rejected the argument that platforms' automated decisions should be presumed valid, instead requiring platforms to demonstrate that their moderation processes comply with constitutional standards.[50] This approach marks a significant departure from earlier judicial attitudes that sometimes-treated technological solutions as neutral and objective. Recent judgments have increasingly recognized that algorithmic systems embody the biases and commercial interests of their creators rather than representing impartial arbiters of speech.[51]

5.2. Asserting Constitutional Primacy

The Indian judiciary is uniquely positioned to establish stringent standards that ensure automated moderation systems operate within constitutional boundaries. The Supreme Court's role as the guardian of fundamental rights requires it to set clear parameters for how private platforms exercise their growing power over public discourse.[52]

The "Right to Explanation"

There is an emerging judicial consensus that users deserve a meaningful explanation when their content is removed. Generic labels such as "violates community standards" fail to meet basic due process requirements under Article 14 and fail to provide users with adequate information

[48] D*igital News Publishers Association v. Union of India*, 2021 SCC OnLine Mad 2095 [27-29].
[49] A*gij Promotion of Nineteenone Media Pvt. Ltd. v. Union of India*, 2021 SCC OnLine Bom 1289 [34-36].
[50] *X v. Union of India*, 2022 SCC OnLine Del 2418 [22-25].
[51] J*ustice K.S. Puttaswamy (Retd.) v. Union of India*, (2017) 10 SCC 1 [345] (applying constitutional standards to technological systems).
[52] I*ndian Express Newspapers v. Union of India*, (1985) 1 SCC 641 [32] (establishing proportionality standard for speech restrictions).

to challenge decisions effectively.[53] The judiciary should recognize a fundamental "right to explanation" that includes: specific identification of the violated rule; contextual explanation of how the content violates that rule; and reference to the specific content that triggered the action rather than boilerplate language. This right is particularly crucial in the Indian context, where users may need to understand moderation decisions to comply with local laws and cultural norms.[54]

Procedural Safeguards

The grievance redressal mechanisms mandated by the IT Rules, 2021 require significant judicial scrutiny to ensure they provide adequate due process. Current provisions suffer from several deficiencies: tight timelines that favor automated decisions over meaningful review; lack of requirements for human review of complex cases; and absence of meaningful appellate mechanisms.[55] Courts should mandate that platforms establish transparent, accessible, and effective appeal processes that include: timely human review of all contested automated decisions; culturally competent reviewers for content in Indian languages; and clear escalation paths to independent oversight bodies.[56] The judiciary should further require that platforms maintain and publish detailed data about content removal decisions, appeal outcomes, and the demographic impact of their moderation practices to enable proper oversight.[57]

The constitutional framework established in *Indian Express Newspapers v. Union of India* regarding permissible restrictions on speech provides the judiciary with robust principles to evaluate automated moderation systems.[58] By applying these established standards to digital platforms, courts can ensure that technological solutions enhance rather than undermine India's democratic discourse.

[53] M*aneka Gandhi v. Union of India*, (1978) 1 SCC 248 [56] (requiring meaningful due process in administrative actions).
[54] S*hreya Singhal v. Union of India*, (2015) 5 SCC 1 [118] (emphasizing need for clarity in speech regulations).
[55] The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, Rule 3(2)(b).
[56] A*rup Bhuyan v. State of Assam*, (2011) 3 SCC 377 [15] (requiring effective safeguards against arbitrary speech restrictions).
[57] P*eople's Union for Civil Liberties v. Union of India*, (2004) 9 SCC 580 [42] (mandating transparency in systems affecting rights).
[58] I*ndian Express Newspapers v. Union of India*, (1985) 1 SCC 641 [68-70] (establishing proportionality test for speech restrictions).

6. Proposed Framework for a Constitutional AI Governance Model

6.1. Principle-Based Design

A constitutional AI governance model for hate speech detection must be anchored in four core principles that align automated systems with India's democratic values and legal traditions.

Transparency requires platforms to publicly disclose their moderation policies in clear, accessible language across all major Indian languages. This includes mandatory publishing of detailed AI performance metrics, including accuracy rates, false positive/negative ratios, and language-specific performance data.[59] Platforms must maintain publicly accessible databases of removed content (with personal information redacted) to enable research and public scrutiny of moderation patterns. This transparency obligation should extend to revealing the general nature of training data sources and the demographic characteristics of content moderators where such information affects decision-making processes.[60]

Explainability (XAI) imposes a legal requirement for platforms to provide user-specific, meaningful explanations for content removals that go beyond generic violation notices. Each takedown notification must include: specific identification of the violated policy provision; contextual explanation of how the content violates that policy; reference to the specific words or phrases that triggered the action; and information about the automated or human nature of the decision.[61] These explanations must be provided in the user's language of preference and accommodate India's multilingual context. The right to explanation should be recognized as a fundamental aspect of procedural due process under Article 14 of the Constitution.[62]

Human-in-the-Loop mechanisms must be mandated for all content moderation decisions with significant consequences, including account suspensions, demonetization, and removal of political content. The framework should require human review before implementing any permanent account restrictions and for all content originating from elected representatives,

[59] S*ecretary, Ministry of Information & Broadcasting v. Cricket Association of Bengal*, (1995) 2 SCC 161 [234] (emphasizing the importance of transparency in systems affecting fundamental rights).
[60] J*ustice K.S. Puttaswamy (Retd.) v. Union of India*, (2017) 10 SCC 1 [356] (recognizing informational privacy as part of constitutional framework).
[61] M*aneka Gandhi v. Union of India*, (1978) 1 SCC 248 [57] (establishing that due process requires meaningful notice and opportunity to be heard).
[62] A*rup Bhuyan v. State of Assam*, (2011) 3 SCC 377 [16] (requiring clear standards in systems affecting speech rights).

government officials, and journalistic entities.[63] Platforms must maintain adequate staffing of human moderators proficient in all scheduled Indian languages and familiar with regional cultural contexts. The human review process must be prioritized for content that has been flagged through coordinated reporting campaigns to prevent weaponization of automated systems.[64]

Auditability requires independent third-party audits of AI systems at regular intervals to assess their compliance with constitutional standards. Auditors should examine algorithmic systems for bias across language, region, religion, caste, and political affiliation.[65] The audit process must include testing with culturally contextual datasets and assessment of fairness metrics across demographic groups. Audit results should be submitted to regulatory authorities and made publicly available in summarized form, with detailed technical reports provided to oversight bodies. The Grievance Appellate Committee should have the authority to mandate special audits when systematic biases are alleged.[66]

6.2. Regulatory and Legislative Recommendations

Clarifying the IT Rules is essential to prevent over-removal and ensure constitutional compliance. The government should issue detailed guidelines defining the scope of "pro-active monitoring" under Rule 4(4), specifying that automated systems must be calibrated to minimize false positives and prioritize precision over recall.[67] The guidelines should establish clear boundaries between mandatory monitoring for specific categories of unlawful content and discretionary moderation of other content. Platforms should be required to implement graduated response systems that use the least restrictive measures necessary, rather than immediate content removal as the default response.[68]

---

[63] P*eople's Union for Civil Liberties v. Union of India*, (2003) 4 SCC 399 [45] (emphasizing the importance of human oversight in rights-affecting decisions).
[64] S*hreya Singhal v. Union of India*, (2015) 5 SCC 1 [119] (cautioning against systems that enable misuse through automated processes).
[65] I*ndian Express Newspapers v. Union of India*, (1985) 1 SCC 641 [72] (emphasizing the need for proportionality in speech restrictions).
[66] The Information Technology Act, 2000, §87(2)(zg) (providing rule-making authority for implementing safeguards).
[67] S*tate of Madras v. V.G. Row*, AIR 1952 SC 196 [15] (requiring that restrictions on rights be precise and unambiguous).
[68] M*odern Dental College v. State of Madhya Pradesh*, (2016) 7 SCC 353 [48] (applying proportionality standard to regulatory measures).

Strengthening the Grievance Appellate Committee (GAC) requires structural reforms to ensure its independence and effectiveness. The GAC should be established as a statutory body with security of tenure for its members, rather than being constituted through executive notification.[69] Its membership should include judicial officers, technical experts, and representatives from civil society with expertise in free expression and digital rights. The Committee must have the authority to review both individual content decisions and broader platform policies, and its decisions should be binding on platforms. The GAC should also be empowered to recommend changes to platform policies and impose proportional penalties for systematic non-compliance.[70]

The Digital India Act should incorporate specific provisions embedding these constitutional principles into law. The legislation must include: a statutory right to explanation for content moderation decisions; mandatory transparency reporting requirements; independent audit mechanisms; and due process safeguards for users.[71] The Act should establish clear liability frameworks that distinguish between fully automated decisions (where platforms assume greater liability) and human-reviewed decisions. It should also create specialized digital rights courts with technical expertise to handle content moderation disputes efficiently.[72] Finally, the legislation should establish a Digital Protection Commission with rule-making authority and enforcement powers to oversee platform compliance with constitutional standards.[73]

7. Conclusion

This comprehensive analysis substantiates the central thesis that unregulated artificial intelligence deployment for hate speech enforcement remains fundamentally incompatible with India's constitutional free speech principles. The research demonstrates that the core challenge lies not in the technology itself, but in the profound misalignment between algorithmic capabilities and the nuanced, context-dependent requirements of Indian jurisprudence. Where Indian courts have developed a sophisticated framework that carefully balances intent, context,

[69] U*nion of India v. R. Gandhi*, (2010) 11 SCC 1 [78] (emphasizing the importance of independent regulatory bodies).
[70] V*ineet Narain v. Union of India*, (1998) 1 SCC 226 [48] (establishing principles for independent oversight mechanisms).
[71] S.*R. Bommai v. Union of India*, (1994) 3 SCC 1 [276] (emphasizing that constitutional values must inform all legislation).
[72] L. *Chandra Kumar v. Union of India*, (1997) 3 SCC 261 [93] (upholding the role of specialized tribunals within constitutional framework).
[73] S.*P. Gupta v. Union of India*, 1981 Supp SCC 87 [64] (recognizing the need for independent regulatory authorities in specialized domains).

historical background, and potential consequences, automated systems operate through pattern recognition that inevitably flattens this essential complexity.

The examination of technical limitations reveals critical vulnerabilities in current AI systems, particularly their inability to comprehend cultural nuance, linguistic diversity, satire, and the complex dynamics of intra-community communication. These limitations become particularly pronounced in India's incredibly diverse digital landscape, where the same words may carry vastly different meanings across regions, communities, and contexts. The research further identifies how these technical shortcomings enable systematic weaponization of automated systems through coordinated reporting campaigns that disproportionately silence marginalized voices, journalists, and activists, thereby undermining the very democratic discourse that hate speech regulations aim to protect.

The path forward, as this paper argues, requires a fundamental reimagining of AI governance through a constitutional framework that embeds democratic values into technological design. This necessitates moving beyond the current binary choice between unregulated automation and complete human moderation. Instead, the proposed model embraces AI as a tool that must be carefully constrained by constitutional principles, rigorous oversight mechanisms, and meaningful human judgment at critical junctures. The framework prioritizes transparency in algorithmic operations, explainability in decision-making processes, human review for consequential actions, and independent auditability to ensure ongoing compliance with constitutional standards.

This approach recognizes that effective content moderation in India's complex digital ecosystem requires neither technological abandonment nor uncritical adoption, but rather a sophisticated integration that respects both the potential of technology and the primacy of constitutional rights. The proposed governance model seeks to harness AI's scalability while ensuring its operation remains subject to democratic accountability and constitutional constraints. This balanced approach acknowledges that preserving India's democratic character in the digital age requires ensuring that automated systems enhance rather than undermine the pluralistic discourse essential for a vibrant democracy.

The urgency of implementing such a framework cannot be overstated. As digital platforms become increasingly central to public discourse, the systems governing online speech will fundamentally shape the future of Indian democracy. Getting this balance right will determine

whether India's digital public sphere becomes a space for vibrant democratic exchange or automated suppression of dissent. By embracing a constitutional approach that balances technological innovation with fundamental rights protection, India can develop a governance model that respects both human dignity and democratic expression—establishing a precedent for how democratic societies can harness technology while preserving their core values.