INTERNATIONAL JOURNAL FOR LEGAL RESEARCH AND ANALYSIS

Open Access, Refereed Journal Multi Disciplinary
Peer Reviewed

www.ijlra.com

ISSN: 2582-6433

DISCLAIMER

The views, opinions, interpretations, and conclusions expressed in the articles published in this journal are solely those of the respective authors. They do not necessarily reflect the views of the Editorial Board, Editors, Reviewers, Advisors, or the Publisher of IJLRA.

Although every reasonable effort has been made to ensure the accuracy, authenticity, and proper citation of the content published in this journal, neither the Editorial Board nor IJLRA shall be held liable or responsible, in any manner whatsoever, for any loss, damage, or consequence arising from the use, reliance upon, or interpretation of the information contained in this publication.

The content published herein is intended solely for academic and informational purposes and shall not be construed as legal advice or professional opinion.

ISSN: 2582-6433

ABOUT US

The *International Journal for Legal Research & Analysis (IJLRA)* (ISSN: 2582-6433) is a peer-reviewed, academic, online journal published on a monthly basis. The journal aims to provide a comprehensive and interactive platform for the publication of original and high-quality legal research.

IJLRA publishes Short Articles, Long Articles, Research Papers, Case Comments, Book Reviews, Essays, and interdisciplinary studies in the field of law and allied disciplines. The journal seeks to promote critical analysis and informed discourse on contemporary legal, social, and policy issues.

The primary objective of IJLRA is to enhance academic engagement and scholarly dialogue among law students, researchers, academicians, legal professionals, and members of the Bar and Bench. The journal endeavours to establish itself as a credible and widely cited academic publication through the publication of original, well-researched, and analytically sound contributions.

IJLRA welcomes submissions from all branches of law, provided the work is original, unpublished, and submitted in accordance with the prescribed submission guidelines. All manuscripts are subject to a rigorous peer-review process to ensure academic quality, originality, and relevance.

Through its publications, the *International Journal for Legal Research & Analysis* aspires to contribute meaningfully to legal scholarship and the development of law as an instrument of justice and social progress.

ISSN: 2582-6433

PUBLICATION ETHICS, COPYRIGHT & AUTHOR RESPONSIBILITY STATEMENT

The *International Journal for Legal Research and Analysis (IJLRA)* is committed to upholding the highest standards of publication ethics and academic integrity. All manuscripts submitted to the journal must be original, unpublished, and free from plagiarism, data fabrication, falsification, or any form of unethical research or publication practice. Authors are solely responsible for the accuracy, originality, legality, and ethical compliance of their work and must ensure that all sources are properly cited and that necessary permissions for any third-party copyrighted material have been duly obtained prior to submission. Copyright in all published articles vests with IJLRA, unless otherwise expressly stated, and authors grant the journal the irrevocable right to publish, reproduce, distribute, and archive their work in print and electronic formats. The views and opinions expressed in the articles are those of the authors alone and do not reflect the views of the Editors, Editorial Board, Reviewers, or Publisher. IJLRA shall not be liable for any loss, damage, claim, or legal consequence arising from the use, reliance upon, or interpretation of the content published. By submitting a manuscript, the author(s) agree to fully indemnify and hold harmless the journal, its Editor-in-Chief, Editors, Editorial Board, Reviewers, Advisors, Publisher, and Management against any claims, liabilities, or legal proceedings arising out of plagiarism, copyright infringement, defamation, breach of confidentiality, or violation of third-party rights. The journal reserves the absolute right to reject, withdraw, retract, or remove any manuscript or published article in case of ethical or legal violations, without incurring any liability.

ISSN: 2582-6433

AI MODERATION OF SOCIAL MEDIA POSTS AND FREEDOM OF SPEECH AND EXPRESSION UNDER ARTICLE 19(1)(A)

AUTHORED BY - ADITRIE BASU, NAMRATA GOMES,
ANUPAM SAHU & ARGHYASMIT DUTTA
2nd year, BBA LLB(H),
Sister Nivedita University, Kolkata, West Bengal.

Abstract

The rapid proliferation of user-generated content on social media has compelled platforms to rely heavily on Artificial Intelligence (AI) for automated content moderation. While AI algorithms offer scalability in curbing hate speech, misinformation, and harmful content, their inherent inability to comprehend context, nuance, sarcasm, and cultural subtleties poses a severe threat to legitimate public discourse. The statement of the problem lies in this systemic over-censorship and algorithmic bias, which increasingly infringes upon the fundamental right to freedom of speech and expression guaranteed under Article 19(1)(a) of the Constitution of India. The primary objective of this study is to examine the constitutional validity of AI-driven moderation and evaluate whether delegating speech regulation to private algorithmic systems violates democratic free speech paradigms. To achieve this, the paper addresses three critical research questions:

I. How do algorithmic errors and biases in AI moderation disproportionately suppress lawful expression under Article 19(1)(a)?
II. To what extent do the current intermediary guidelines fail to protect users against arbitrary private censorship?
III. How can a balanced legal framework be structured to harmonize technological gatekeeping with constitutional safeguards?

The methodology employed is strictly doctrinal, relying on a qualitative analysis of primary legal sources, including constitutional provisions, IT regulations, and landmark judicial precedents. Secondary sources like legal commentaries, academic journals, and global tech policy reports are also analyzed. Ultimately, the study argues for greater algorithmic

ISSN: 2582-6433

transparency, robust appeal mechanisms, and co-regulatory frameworks to ensure that AI serves as a shield for public safety rather than a sword against constitutional liberties.

Keywords- Artificial Intelligence, Automated Content Moderation, Algorithmic Bias, Freedom of Speech and Expression.

Chapter 1: Introduction

1.1 Philosophical Foundations

The idea of free speech is not merely a contemporary legal construct; it is rooted in a rich philosophical tradition. From the dialogues of Socrates in ancient Athens, advocating for the right to question, to John Stuart Mill's influential text "On Liberty," the rationale for free expression has been established on several foundational principles. Mill emphasized that open discourse is crucial for uncovering truth, as even dissenting or erroneous views prompt society to reassess its beliefs. He also highlighted the importance of individual autonomy, asserting that the ability to form and express personal convictions is essential for a fulfilling life. Moreover, democratic theory, spanning from Rousseau to modern thinkers, asserts that for self-governance to be valid, citizens must have unrestricted access to information and the freedom to engage in public debate.

1.2 Indian Perspective

In India, the fight for free speech was deeply intertwined with the quest for independence. Under British colonial rule, various laws were enacted to stifle dissent, including the Indian Press Act of 1910 and the Sedition Act of 1870, which led to the suppression of newspapers and the imprisonment of prominent figures like Bal Gangadhar Tilak and Mahatma Gandhi. Gandhi articulated the essence of this struggle by asserting that true liberty of speech must endure even when it is uncomfortable. This experience of oppression galvanized the Constituent Assembly to enshrine free speech in the Constitution. However, the trauma of Partition and the complexities of uniting a diverse nation prompted caution, resulting in specific provisions for limitations. The Assembly debates reflect a clear intention: to establish a right that is strong enough to support democratic health while being mindful of the potential for chaos or violence that absolute freedom could invite[1].

[1]Freedom of Speech & Expression (Article 19(1)(a), Law Crub, https://www.lawcurb.in/post/freedom-of-speech-expression-article-19-1-a.

ISSN: 2582-6433

1.3 Contemporary Framework

In the contemporary digital era, online platforms such as X (formerly Twitter), Facebook, and YouTube have become primary spaces for public discourse, political debate, and information sharing. To manage the enormous volume of user-generated content, these platforms increasingly rely on Artificial Intelligence (AI) based moderation systems that automatically detect and remove content considered harmful, such as hate speech, misinformation, or incitement to violence. While these technologies aim to maintain safer online environments, they also raise serious constitutional concerns regarding over-censorship, algorithmic bias, lack of transparency, and potential suppression of legitimate speech.

The tension between technological governance and constitutional rights has become particularly significant in India, where digital regulation operates within the framework of the Information Technology Act, 2000 and the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021. The judiciary has also played a crucial role in shaping this balance, most notably in the landmark judgment of **Shreya Singhal v. Union of India**[2], where the Supreme Court (consisting of a two-judge bench of Justice Jasti Chelameswar and Justice Rohinton Fali Nariman) struck down Section 66A of the IT Act for violating free speech protections.

Against this backdrop, the increasing reliance on AI-driven content moderation raises critical questions about the relationship between intermediary liability, algorithmic governance, and constitutional guarantees of free expression. This study seeks to examine how AI moderation mechanisms interact with the protections under Article 19(1)(a), the legal obligations imposed on intermediaries, and the broader challenge of balancing freedom of expression with public order and societal interests in the digital age. It further explores judicial perspectives, comparative global approaches, and potential regulatory reforms to ensure that technological regulation does not undermine democratic freedoms.

Chapter 2: Constitutional Basis

"If freedom of speech is taken away, then dumb and silent we may be led, like sheep to the slaughter." - George Washington

The freedom of speech and expression is one of the most fundamental rights guaranteed under the Article 19(1)(a) of the Constitution of India. It ensures that every citizen has the liberty to express opinions, ideas, and information without undue interference from the State. However,

[2] 2015 5 S.C.C 1 (India).

ISSN: 2582-6433

this right is not absolute and is subject to reasonable restrictions under Article 19(2) of the Constitution of India in the interests of sovereignty and integrity of India, security of the State, public order, decency, morality, and other specified grounds. With the rapid growth of digital communication and social media platforms, the scope and application of this constitutional freedom have entered a new and complex phase.

2.1 Scope of Article 19 (1)(a)

Article 19(1)(a) includes the freedom to:

1. Express opinions through words (spoken or written);
2. Communicate ideas through print, electronic media, or digital platforms;
3. Express views through art, gestures, or symbolic expression;
4. Receive and disseminate information;
5. Freedom of the press (recognized by judicial interpretation).

2.2 Restrictions on Article 19(1)(a)

However, the right is not absolute. Under Article 19(2) of the Constitution of India, the State may impose reasonable restrictions by law on specific grounds like Sovereignty and integrity of the country, security of the state, friendly relations with foreign states, public order, decency or morality, contempt of court, malafide defamation, incitement to an offence[3].

These restrictions must to satisfy two conditions:

a) They must be imposed by law, and
b) They must be reasonable, meaning they should not be arbitrary or excessive.

Chapter 3: AI Moderation Mechanisms on Social Media Platforms

3.1 Mechanics of AI Content Moderation

AI engines employed in content moderation use a blend of advanced techniques, each playing a distinct role in identifying harmful or inappropriate content.

i) Machine Learning: Machine learning models are trained on massive datasets of text, images, and videos. These models learn patterns that help classify whether content is safe or problematic. As more data is processed, the models continuously improve, leading to higher accuracy and less reliance on manual review[4].

[3] India Const. art. 19(2).
[4]S.P. Pattyam, AI-Driven Data Science for Environmental Monitoring: Techniques for Data Collection, Analysis, and Predictive Modeling, 1(1) Australian Journal of Machine Learning Research & Applications, 132-169 (2021).

ISSN: 2582-6433

ii) Natural Language Processing (NLP): NLP enables AI to understand the nuances of human language. It goes beyond keyword detection by interpreting grammar, tone, slang, and even intentional misspellings that users may use to evade detection. By analysing vast amounts of text at lightning speed, NLP makes it possible to moderate real-time conversations, comments, and posts efficiently[5].

iii) Large Language Models (LLMs): LLMs extend NLP's capabilities by offering deeper contextual understanding. Instead of only spotting individual words, they can analyse the meaning of entire sentences, conversations, or threads. This allows moderation systems to detect subtle cases of harassment, misinformation, or hate speech that keyword filters might miss. Their ability to process content quickly and in context makes them especially valuable for scaling moderation on fast-moving platforms[6].

iv) Image and Video Recognition: Beyond text, AI also moderates visual content. Image and video recognition technologies can identify explicit imagery, violent material, or even subtle visual cues that violate community guidelines. Combined with contextual understanding, these systems enable platforms to address inappropriate media at scale, complementing text-based moderation for a more comprehensive approach[7].

3.2 Technical Challenges in AI Content Moderation

i) Contextual Understanding: AI struggles to grasp context, often missing nuances like sarcasm or cultural references. These complexities can lead to misclassification, where benign content gets flagged, or harmful content goes unnoticed. Algorithms lack the human-like ability to discern intent and subtleties. As human speech is not objective and the process of content moderation is inherently subjective, these tools are limited in that they are unable to comprehend the nuances and contextual variations present in human speech[8]. These tools are limited in their ability to understand variances in language and behavior that may result from different demographic and regional factors. For example, excessively liking someone's pictures or using certain slang words may be construed as harassment on one platform or in one region of the world. However, these behaviors and speech may take on an entirely different meaning on another

[5]M.H. Huang & R.T. Rust, A strategic framework for artificial intelligence in marketing, 49 Journal of the Academy of Marketing Science, 30-50 (2021).
[6]*Supra Note 4.*
[7] K.K. Ng, et.al., A Systematic Literature Review on Intelligent Automation: Aligning Concepts from Theory, Practice and Future Perspectives, 47 Advanced Engineering Informatics, 101246 (2021).
[8]J. Grimmelmann, The Virtues of Moderation, 17 Yale Journal of Law & Technology, 42 (2015).

ISSN: 2582-6433

platform or in another community[9]. In addition, automated tools are also limited in their ability to derive contextual insights from content. For example, an image recognition tool could identify an instance of nudity, such as a breast, in a piece of content. However, it is unlikely to be able to determine whether the post depicts pornography or perhaps breastfeeding, which is permitted on many platforms[10]. In addition, automated content moderation tools can become outdated rapidly.

ii) Language Nuances: Languages are rich with idioms and slang that evolve quickly. AI must adapt to these changes to ensure accurate moderation. However, models sometimes fail with multilingual content, missing harmful phrases or words due to limited training in less common dialects. These automated tools also need to be updated as language and meaning evolves. For example, in an attempt to avoid moderation, some hateful groups have adopted new methods of slang and representations for indicating hate. One example of this is white supremacists using the names of companies, such as “Google” and “Yahoo” to replace ethnic slurs. In order to keep up, automated tools would have to adapt quickly and be trained across a wide range of domains. However, users could continue developing new forms of speech in response, thus limiting the ability of these tools to act with significant speed and scale[11]. On some platforms when human moderators engage in content moderation, they are able to combat the rapidly changing nature of speech by viewing additional information on the case, such as information on the user who is accused of violating the platform’s rules. However, incorporating such assumptions and processes into an automated tool runs the risk of enhancing biases around particular groups of individuals and could result in skewed or even discriminatory enforcement of content policies[12].

Content that's acceptable in one culture might be offensive in another. AI requires diverse training data to recognize these differences. Yet, creating a universally applicable model is challenging, risking over-censorship in culturally sensitive areas or under-censorship elsewhere.

Research indicates that AI systems are 30% less accurate in moderating content in languages other than English. This statistic is crucial for global platforms and creators who operate in multilingual environments, emphasizing the need for improved

[9] R. Caplan, Content or Context Moderation: Artisanal, Community-Reliant, and Industrial Approaches (2018).
[10] J. Vincent, AI Won't Relieve the Misery of Facebook's Human Moderators, The Verge (Feb. 27, 2019).
[11] N. Duarte, et.al, Mixed Messages? The Limits of Automated Social Media Content Analysis, Conference on Fairness, Accountability, and Transparency (2018).
[12] *Ibid.*

ISSN: 2582-6433

language processing capabilities in AI systems.

As of now, AI researchers have been unable to construct comprehensive enough datasets that can account for the vast fluidity and variances in human language and expression. As a result, these automated tools cannot be reliably deployed across different cultures and contexts, as they are unable to effectively account for the various political, cultural, economic, social, and power dynamics that shape how individuals express themselves and engage with one another.

iii) Accuracy and Error Rates: In the case of content such as extremist content and hate speech, there are a range of nuanced variations in speech related to different groups and regions, and the context of this content can be critical in understanding whether or not it should be removed. As a result, developing comprehensive datasets for these categories of content is challenging, and developing and operationalizing a tool that can be reliably applied across different groups, regions, and sub-types of speech is also extremely difficult. In addition, the definition of what types of speech fall under these categories is much less clear[13]. Although smaller platforms may rely on off-the-shelf automated tools, the reliability of these tools to identify content across a range of platforms is limited. In comparison, proprietary tools developed by larger platforms are often comparatively more accurate, as they are trained on datasets reflective of the types of content and speech they are meant to evaluate.

A recent study by the Center for Democracy & Technology reports that AI content moderation systems can have an error rate of between 5% to 10% in identifying harmful content. This statistic underscores the potential for significant false positives (flagging non-harmful content) and false negatives (missing harmful content) which can lead to either wrongful content removal or harmful content slipping through, affecting user trust and platform integrity. This imbalance can disrupt user experience and leave platforms vulnerable to harmful content, stressing the need for continuous model refinement and human oversight.

iv) Bias and Fairness Issues

One of the key concerns around algorithmic decision-making across a range of industries is the presence of bias in automated tools. Decisions based on automated tools, including in the content moderation space, run the risk of further marginalizing and censoring groups that already face disproportionate prejudice and discrimination

[13] P. Parycek, Artificial Intelligence (AI) and automation in administrative procedures: Potentials, limitations, and framework conditions, 15(2) Journal of the Knowledge Economy, 8390-8415 (2024).

ISSN: 2582-6433

online and offline[14]. As outlined in a report by the Center for Democracy & Technology, there are many types of biases that can be amplified through the use of these tools. NLP tools, for example, are typically used to parse text in English. Tools that have a lower accuracy when parsing non-English text can therefore result in harmful outcomes for non-English speakers, especially when applied to languages that are not very prominent on the internet, as this reduces the comprehensiveness of any corpora that models are trained on. Given that a large number of the users of major internet platforms reside outside English-speaking countries, this is highly concerning. The use of such automated tools in decision-making should therefore be limited when making globally relevant content moderation decisions[15]. These tools are also unable to effectively process differences in dialect and language use that may result from demographic differences[16]. In addition, the personal and cultural biases of researchers are likely to find their way into training datasets. For example, when a corpus is being created, the personal judgments of the individuals annotating each document can impact what is constituted as hate speech, as well as what specific types of speech, demographic groups, and so on are prioritized in the training data. This bias can be mitigated to some extent by testing for intercoder reliability, but it is unlikely to combat the majority view on what falls into a particular category[17].

A 2023 survey found that 60% of AI moderation tools exhibited some form of bias, particularly against marginalized communities. Understanding bias in AI moderation systems is essential for developers aiming to create fair and equitable technologies, and for agencies committed to ethical content practices.

3.3.A) Role of Human Oversight in addressing Over-Moderation Issues

Human oversight helps counteract over-moderation by adding contextual judgment that algorithms often lack. It allows reviewers to reinstate flagged content that was erroneously removed, restoring balance and user trust.

i) Contextual Review

Humans excel at discerning intent, nuance, and cultural context in edge cases—like sarcasm, reclaimed slurs, or benign discussions—that AI misflags as violations. Moderators review AI-flagged items, overturning over-moderation decisions to prevent

[14] *Supra Note 11.*
[15] *Ibid.*
[16] *Supra Note 10.*
[17] *Ibid.*

ISSN: 2582-6433

stifling legitimate speech, as seen in gaming communities where hybrid systems cut harassment without excessive censorship[18].

ii) Appeal and Escalation Processes

Oversight includes structured appeals where users challenge removals, with human teams providing final rulings to correct false positives and ensure fairness. This reduces self-censorship by signalling accountability, especially during crises when AI spikes can overwhelm without human triage[19].

iii) Training and Iteration

Human feedback loops refine AI models over time, minimizing future over-moderation by incorporating real-world insights into training data. Diverse moderator teams and clear guidelines further curb biases, fostering consistent enforcement without blanket suppression[20].

3.3.B) Examples of Platforms successfully reducing Over-Moderation with Human Oversight

Human oversight significantly reduces over-moderation by providing nuanced review of AI-flagged content, preventing erroneous removals of legitimate speech. Platforms like Meta (Facebook) and Uber exemplify this hybrid success through structured human intervention.

i) Meta/Facebook Hybrid Model

Meta employs tens of thousands of human moderators alongside AI to handle ambiguous cases, such as sarcasm or cultural references, ensuring harmless posts aren't deleted while catching nuanced violations. This approach has minimized wrongful takedowns in high-volume user-generated content, balancing scale with accuracy[21].

Meta's Oversight Board advocates auditing AI systems with human input, as in cases of over-enforcement on health education or anti-trans speech, leading to refined policies that cut false positives through ongoing human refinement[22].

ii) Uber's Trust & Safety Teams

Uber integrates AI for initial flagging of unsafe messages or behaviours, with 24/7

[18] M. F. Peñalver, Keeping AI in Check: The Critical Role of Human Agency and Oversight, (Feb. 26, 2024).
[19] EU Artificial Intelligence Act, 2024, art. 14.
[20] *Ibid.*
[21] A.Mohamed, et.al., The State of the Art and Taxonomy of Big Data Analytics: View from New Big Data Framework, 53 Artificial Intelligence Review, 989-1037 (2020).
[22] Y. Yang, et.al., Multiple Knowledge Representation for Big Data Artificial Intelligence: Framework, Applications, and Case Studies, 22(12) Frontiers of Information Technology & Electronic Engineering, 1551-1558 (2021).

ISSN: 2582-6433

human teams reviewing reports and patterns to enforce guidelines without overreach. Post-safety incidents, this reduced excessive censorship in rider-driver communications while prioritizing real threats like abuse or fraud[23].

3.4 Use of AI for Moderating Large-Scale Content by Different Social Media Platforms

i) Social Media

Social media platforms like Facebook, Instagram, and Twitter face the daunting task of moderating billions of posts daily. Their AI systems are highly advanced, designed to analyze text, images, and videos in real time.

- Cyberbullying prevention

  By analysing language patterns, sentiment, and user interactions, AI can detect instances of harassment or bullying. Platforms like Instagram use AI to automatically flag harmful comments or messages, offering real-time protection to users[24].

- Hate speech and toxic content

  AI algorithms can identify offensive language, discriminatory remarks and threats, remove such content or flag it for human review. For example, Facebook's AI systems are trained to recognize and mitigate the spread of hate speech in multiple languages, across diverse cultural contexts[25].

- Graphic and violent content

  By analysing images and videos, AI can automatically take down content that depicts violence, gore, or other harmful visuals, protecting users from exposure to disturbing material[26].

ii) Video Sharing Platforms

Platforms like YouTube and TikTok are centred around video content, requiring AI models that specialize in visual and audio analysis. These platforms face unique challenges, like: -

- Violent content and deep fakes

[23] *Ibid.*
[24] S.Tatineni & V.R. Boppana, AI-Powered DevOps and MLOps Frameworks: Enhancing Collaboration, Automation, and Scalability in Machine Learning Pipelines, 1(2) Journal of Artificial Intelligence Research and Applications, 58-88 (2021).
[25] W. Liang, et al., Advances, Challenges and Opportunities in Creating Data for Trustworthy AI, 4(8) Nature Machine Intelligence, 669-677 (2022).
[26] Y. J. Alawneh, et.al., A Detailed Study Analysis of Artificial Intelligence Implementation in Social Media Applications, 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), IEEE pp. 1191-1194 (2023).

ISSN: 2582-6433

AI is also used to detect violent acts or manipulated videos (such as deepfakes) that spread misinformation. The ability to analyse both visual and audio cues in videos allows AI to flag suspicious content quickly.

Chapter 4: Legal Framework for Intermediaries

4.1 Impact of Recent Statutory Regulations on AI Deployment for Content Removal

The February 2026 amendment to the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 represents India's most assertive regulatory intervention into AI-generated content to date. By compressing takedown timelines, mandating technical traceability, and redefining intermediary obligations, the government has shifted from reactive moderation toward proactive algorithmic governance. These revisions mark a significant shift - aimed squarely at controlling how AI-made material spreads online. Instead of broad oversight, the focus zeroes in on manipulated videos and fake audio clips flooding social networks. The amendment introduces the following key structural changes:

i) Statutory definition of “deepfake”: The guidelines define "deepfake” as content generated using algorithmic or computational techniques - to produce sound, visuals, or both, so convincingly that they could pass as authentic representations of real individuals[27]. Such material includes pictures made by artificial intelligence, replicated voices, or video clips crafted to mislead an audience.

For the first time in India, the law formally recognizes **Synthetically Generated Information (SGI)[28]** – content such as videos, images, or audio that is created or altered using artificial intelligence but appears real. To prevent misuse of such technology, the amendment introduces:

- **Mandatory labelling:** AI-generated content must clearly display visible labels so users know that the content is artificial.
- **Audio disclosures:** AI-generated audio must include clear voice disclosures so listeners understand it is synthetic.
- **Traceability measures:** Platforms must embed permanent digital identifiers (metadata or fingerprints) into AI-generated files to help trace their origin if misuse occurs.

[27]Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026, Rule 2(1)(d), G.S.R. 120(E), Ministry of Electronics and Information Technology (India).
[28]IT Rules Amendment, 2026 – provisions relating to Synthetically Generated Information (SGI).

ISSN: 2582-6433

- **Tampering prohibited:** Removing or disabling these identifiers is strictly prohibited.
- **Reasonable exemptions:** Normal editing practices like colour correction, noise reduction, accessibility tools, or academic training material that do not create false impressions are excluded from SGI regulation.

ii) Drastic reduction in the takedown timeline: Perhaps the most striking shift comes with the takedown timeline. Earlier versions had required intermediaries to remove unlawful content within 36 hours of receiving notice. The new amendment slashes that to merely 3 hours—a 92% reduction in response time, reflecting government concern about how quickly misinformation spreads on social media[29]. This shortened period applies upon receiving either of two types of notice: knowledge through a court order or a written communication issued by an authorized officer not below the rank of Joint Secretary, specifying legal grounds and identifying the exact web address of the material in question. Notably, all such notices must undergo a three-level review before being acted upon. If especially delicate material appears - like non-consensual private photos, altered pictures depicting terrorist actions, or anything likely to provoke harm - platforms must report it straightaway and remove it within two hours. When children are involved, as defined under the Bharatiya Nyaya Sanhita 2023 (formerly governed by the Protection of Children from Sexual Offences Act, 2012)[30] information has to reach the National Commission for Protection of Child Rights within one day. Crucially, these reports should keep evidence intact, avoiding intrusive markings such as visible stamps. User complaints must now be acknowledged and resolved within seven days instead of fifteen. Complaints related to identity theft or serious harm must be resolved within 36 hours**.**

iii) Mandatory technical disclosure and traceability measures for AI-generated content: To enable users to recognize and assess AI-generated content, platforms must now deploy technical disclosure measures[31]. This marks India's first regulatory step beyond basic content moderation. Specifically, intermediaries must embed metadata or use other unique technical markers that allow tracing AI-generated content back to its original source. Unlike basic watermarks that users can easily see, this approach cannot be

[29] Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026, Rule 3(1)(b)(ii), G.S.R. 120(E), Ministry of Electronics and Information Technology (India).
[30] Bharatiya Nyaya Sanhita, 2023, § 69-72 (India); Protection of Children from Sexual Offences Act, 2012,(India).
[31] Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026, Rule 4(4)(a), G.S.R. 120(E), Ministry of Electronics and Information Technology (India).

ISSN: 2582-6433

bypassed with simple editing tools. It helps verify authenticity while combating coordinated disinformation campaigns[32]. Significantly, if usage exceeds internally declared thresholds, automated systems must then check content for at least 50% accuracy. Any content that fails this threshold must be flagged. The amendment tightens language from the previous draft text of 'endeavour to use' to simply 'must'—no longer allowing intermediaries to claim they demonstrated good faith effort[33]. Instead, functional deployment must be operational at least every three months, with simple audits available upon request. However, the amendment does not clearly define how "accuracy" is to be measured — whether through machine confidence scoring, human verification, or third-party auditing — leaving operational ambiguity that may complicate enforcement. The applicability of these rules extends across certain categories of intermediaries: significant social media intermediaries as defined under Rule 3(1)(w), and community-facing online gaming intermediaries rather than passive infrastructure hosts.

iv) Expanded compliance and dispute resolution obligations for significant intermediaries: Dispute Resolution must now be resolved within 7 days, down from the previous 15-day window[34]. Instead of relying on voluntary self-regulatory relief mechanisms, this codifies a default dispute protocol. However, the safe harbour provisions referring to due diligence remain intact, though notably fail to specify what happens if procedures go awry and lead to errors—potentially exposing intermediaries to liability and provoking a chilling effect. This approach represents a hybrid regulatory model. The rules aim to strike a balance between government oversight and platform-level pre-publication structural controls, but the architecture itself pushes toward algorithmic content moderation.

v) Stronger Responsibilities for Large Platforms (SSMIs)

Platforms with more than five million users, classified as Significant Social Media Intermediaries (SSMIs**)**, now face stricter duties even before content becomes public. New obligations include:

- **User disclosure:** Users uploading content must declare whether it is AI-

[32] Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026, Rule 4(4)(b), G.S.R. 120(E), Ministry of Electronics and Information Technology (India).
[33] Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026, Rule 4(4)(c), G.S.R. 120(E), Ministry of Electronics and Information Technology (India).
[34] Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026, Rule 4(2)(d), G.S.R. 120(E), Ministry of Electronics and Information Technology (India).

ISSN: 2582-6433

generated.

- **Technical verification:** Platforms must use automated tools to verify such declarations before allowing publication.
- **Risk of liability:** If platforms fail to label AI-generated content or miss the strict takedown timelines, they lose **safe harbour protection** under Section 79 of the IT Act[35]. This means they can become legally responsible for harmful user content.

4.2 Intersection with Article 19(1)(a)

From a legal perspective, these rules engage Article 19(1)(a) of the Indian Constitution, which guarantees freedom of speech and expression[36]. Any restriction on this fundamental right must satisfy strict constitutional tests: Under the Supreme Court's proportionality framework — articulated in **Modern Dental College v. State of Madhya Pradesh**[37] and reaffirmed in **K.S. Puttaswamy v. Union of India**[38] — any restriction must pursue a legitimate aim, adopt rational means, be necessary in the absence of less restrictive alternatives, and include adequate procedural safeguards. Critics argue the three-hour takedown mandate imposes burdens, implementing what amounts to a prior restraint, and avoids meaningful procedural safeguards—all of which could be construed as censorship and could create a chilling effect on free expression[39]. Satirical content, for instance, may become more difficult to distinguish from genuine deepfakes, raising concerns if enforcement lacks nuance and subjects material to pre-emptive removal. If enforced strictly as written, vagueness in constitutional jurisprudence about what must be narrowly tailored and adequately filtered will emerge. Meeting filing demands requires significant effort, while seeking redress unfolds through layered prior analysis - proportionality assessments come first, followed by four-step connections, arguments based on essential need, alongside options for milder, step-by-step penalties. After-the-event scrutiny offers minimal safeguarding, shifting discussions about personal privacy or societal damage into broad regulatory promises without clear boundaries.

[35]Information Technology Act, 2000, § 79 (India).
[36]India Const. art. 19(1)(a).
[37] (2016) 7 SCC 353 (India).
[38] (2017) 10 SCC 1 (India).
[39] Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026, Rule 3(1)(b)(ii), G.S.R. 120(E), Ministry of Electronics and Information Technology (India).

ISSN: 2582-6433

Chapter 5: Conflicts, Violation and Reasonable Restrictions

Article 19(1) (a) of the Indian Constitution guarantees every citizen "the right to freedom of speech and expression"[40]. This right is not absolute; Article 19(2) permits the State to impose *"reasonable restrictions"* on speech on specified grounds (sovereignty, security of the State, friendly relations, public order, decency, contempt of court, defamation, and incitement to offence)[41]. In a landmark judgment,[42] the Supreme Court reaffirmed that any law curtailing speech must be "approximately related" to one of these eight grounds and clearly defined (otherwise it is unconstitutionally vague or overbroad). The Court struck down Section 66A of the IT Act on exactly these grounds – it had no clear nexus to the Article 19(2) exceptions and chilled a "very large amount of protected and innocent speech"[43]. In yet another landmark judgment[44], a five-judge bench likewise held that free speech "could not be restricted for any reason other than those under Article 19(2)"[45]. Thus, Indian law requires any restriction (including content moderation) to be narrowly tailored to a legitimate aim and not infringe the broad core of speech.

5.1 Content-Takedown Laws

Social media platforms in India are regulated as "intermediaries" under the IT Act, 2000. Section 79 of the IT Act grants intermediaries **safe-harbor immunity**: they "shall not be liable" for third-party content as long as they act as neutral hosts and follow "due diligence" norms[46]. However, immunity is lost if an intermediary fails to remove unlawful content after receiving a court order or government notification[47]. For example, Section 69A empowers the government to block online content on narrow grounds (sovereignty, security, public order, etc. – **grounds mirrored from Article 19(2)**[48]. The Supreme Court upheld that power in *Shreya Singhal*[49], but struck down overly broad prior restraints. In December 2021 the Government also notified new Intermediary Guidelines (the IT Rules 2021) requiring platforms to establish grievance officers, acknowledge takedown orders within 36 hours, and even trace the "first

[40] *Supra Note 36.*
[41] *Supra Note 3.*
[42] *Supra Note 2.*
[43] *Ibid.*
[44] Kaushal Kishore v. State of Uttar Pradesh, (2023) 4 S.C.C. 1 (India).
[45] Arpan Banerjee, Supreme Court Review Digital Platform Regulation and Freedom of Expression in India, 15 *J. Creative Commc'n* 215, 220–22 (2023).
[46] *Supra Note 35.*
[47] *Ibid.*
[48] *Supra Note 3.*
[49] *Supra Note 2.*

ISSN: 2582-6433

originator" of a message (with significant privacy implications)[50].Critics warn these rules may violate free speech and privacy by imposing retroactive traceability and incentivizing blanket filtering[51]. In sum, Indian statutes compel intermediaries to remove certain content on demand, balancing platform immunity against state-mandated censorship.

5.2 Key Indian Case Laws on Online Speech

The Indian judiciary has generated several landmark decisions on online expression:

- **Shreya Singhal v. Union of India (2015)[52]** – Struck down Section 66A of the IT Act as vague and overbroad. The Court held any restriction must fit Article 19(2) and provide clear standards. Section 66A's criminalization of "grossly offensive" online speech had no proximate link to security, public order, or the other specified grounds, and its undefined terms curtailed "a very large amount of protected and innocent speech".
- **Anuradha Bhasin v. Union of India (2020)[53]** – Held that indefinite internet shutdowns violate Article 19(1) (a) unless strictly necessary and proportionate. The Court affirmed that digital communication is a principal medium of free expression, and any suspension must satisfy necessity and proportionality tests. It ruled that while the government could order a complete shutdown, such orders must be public, time-bound and subject to judicial review.
- **Wikimedia Foundation Inc. v. ANI Media (2025)[54]** – The Supreme Court quashed a Delhi High Court order forcing Wikipedia to remove an article about a media lawsuit. The Court found the takedown request "disproportionate" and likely to have a "chilling effect on free speech"[55]. It held that lower courts may not order deletion of online content about sub judice matters unless it scandalizes the court or prejudices proceedings. Emphasizing the public's "right to know", the Court reaffirmed Article 19(1) (a)'s reach even in digital forums.
- **Kunal Kamra v. Union of India (Bombay HC, 2024)[56]**– A split Bombay High Court struck down a 2023 rule empowering a government-appointed "fact-check unit" to flag

[50]Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, Rule 7, G.S.R. 139(E), Ministry of Electronics and Information Technology (India).
[51]*Ibid.*
[52]*Supra Note 2.*
[53](2020) 3 S.C.C. 637 (India).
[54]2025 INSC 656 (India).
[55]*Ibid.*
[56]2024: BHC-OS: 1575 -DB (India).

ISSN: 2582-6433

government-related social media posts as "fake, false or misleading." The majority (2:1) found the rule prima facie unconstitutional. Justice Patel held it placed an undue burden on platforms (shifting liability from user to intermediary), incentivizing pre-emptive self-censorship and stifling the open marketplace of ideas. Justice Gokhale dissented, viewing the rule as a proportionate anti-misinformation measure consistent with Article 19(2) and providing adequate safeguards. Ultimately, in a tie-breaker, Justice Chandurkar joined the majority: he agreed the rule violated Articles 19(1) (a) and 14 by being vague, one-sided, and chilling speech. (The rule is currently stayed pending appeal).

- **Other Notable Rulings:** In **Kaushal Kishore v. State of Uttar Pradesh (2023)**[57] the Court made clear that freedom of speech "could not be restricted" except on Article 19(2) grounds. It reaffirmed that even powerful public speech (by politicians) falls under Article 19(2) constraints only. In contrast, **Media One v. Union of India (2023)**[58] saw the Court quashes an arbitrary broadcast ban, underscoring that vague national-security claims cannot override speech rights absent clear proof. And in **Foundation for Media Professionals v. Union Territory of Jammu and Kashmir**[59], the Supreme Court ordered guidelines for seizure of journalists' devices, noting that digital evidence seizures can chill the press. Together, these decisions insist that any content restriction must satisfy strict necessity, proportionality and procedural fairness tests.

5.3 Judicial Analysis

From a judicial standpoint, the governance of online speech—especially through AI-driven moderation—has evolved around core constitutional principles such as freedom of expression, proportionality, procedural fairness, and accountability. Courts approach AI moderation not merely as a technological issue, but as a constitutional question involving the protection of fundamental rights under Article 19(1)(a)[60].

A foundational element in judicial reasoning is the recognition that online speech enjoys the same constitutional status as offline expression. This means that any restriction on digital content must meet the standards laid down under Article 19(2)[61], including legality, necessity, and proportionality. Courts have been clear that the medium of expression does not dilute the

[57] *Supra Note 44.*
[58] (2023) 7 S.C.C. 433, 450 (India).
[59] (2020) 5 S.C.C 746, 752 (India).
[60] *Supra Note 36.*
[61] *Supra Note 3.*

ISSN: 2582-6433

strength of the right.

AI-driven content moderation in India is governed by **Article 19(1)(a)**[62], subject to reasonable restrictions under Article 19(2). Indian courts require that any restriction—whether imposed by the State or through platform mechanisms—must be **clear, proportionate, and procedurally fair**.

In **Kunal Kamra v. Union of India**[63] (prima facie), the Bombay High Court indicated that **platforms cannot exercise unchecked discretion** in moderating content, reinforcing accountability in digital governance. This is significant for AI moderation, where algorithmic opacity can mask arbitrary decisions.

The Supreme Court in **Shreya Singhal v. Union of India**[64] struck down vague online speech restrictions, holding that **uncertain standards lead to over-censorship and chilling effects**—a core risk in automated moderation systems.

Similarly, **Anuradha Bhasin v. Union of India**[65] affirmed that restrictions on digital expression must satisfy **proportionality**, meaning they must be necessary and the least restrictive means.

In **Wikimedia Foundation v. ANI**[66], the Court warned against excessive takedowns, emphasizing the **public's right to know**.

In contrast, international regulatory approaches—particularly the **European Union's AI Act (2024)**[67]—take a more **structured compliance-based route**, mandating **transparency in algorithmic decision-making, documentation of moderation processes, and user rights to contest automated decisions**. Similarly, global frameworks such as **UNESCO guidelines**[68] emphasize that AI moderation must be **human rights–compliant, non-discriminatory, and accountable**.

- **India relies on constitutional adjudication**, where courts intervene to strike down vague or disproportionate restrictions.
- **International regimes focus on ex-ante regulation**, imposing predefined obligations on platforms to ensure transparency and fairness.

Closely linked to this is the issue of the "chilling effect". AI moderation systems, due to their

[62] *Supra Note 36.*
[63] *Supra Note 56.*
[64] *Supra Note 2.*
[65] *Supra Note 53.*
[66] *Supra Note 54.*
[67] EU Artificial Intelligence Act, 2024.
[68] UNESCO, Recommendation on the Ethics of Artificial Intelligence, U.N. Doc. SHS/BIO/PI/2021/1 (Nov. 23, 2021).

ISSN: 2582-6433

limited ability to interpret context, satire, or political nuance, often lead to over-removal of lawful content. Courts recognize that when users anticipate arbitrary or opaque takedowns, they may refrain from expressing themselves freely. This indirect suppression is treated as a serious constitutional concern because it undermines open discourse and democratic participation.

Another key aspect of judicial analysis is procedural due process. Courts emphasize that any restriction on speech must be accompanied by:

- Notice to the affected user,
- Clear and reasoned justification for the removal, and
- An effective opportunity to challenge or appeal the decision.

The judiciary also pays close attention to the problem of vagueness in legal standards. Laws governing online content often use broad and undefined terms such as "harmful," "offensive," or "misleading." When such vague standards are enforced through AI systems, the result is often overbroad and inconsistent censorship. Courts have repeatedly stressed that clarity and precision in law are essential, especially when enforcement mechanisms operate at scale through automated tools.

Chapter 6: Policy Recommendations

To bridge the gap between automated efficiency and constitutional protections, the following policy framework is proposed. These recommendations aim to align AI moderation practices with the Reasonable Restriction standard under Article 19(2) and the principles of natural justice.

6.1 Mandating Human-in-the-Loop for Sensitive Contexts

AI should never be the final arbiter of speech that is inherently contextual.

- Contextual Shield: Policies should mandate human review for content flagged in categories involving political discourse, satire, news reporting, and social activism.
- Threshold for Automated Removal: Purely automated takedowns should be restricted to objective violations, such as non-consensual intimate imagery, known terrorist signatures such as hashing, or child sexual abuse material.

6.2 Algorithmic Transparency and Explainability

The Black Box nature of AI moderation violates the principle of due process.

- Specific Notice: When content is removed, the intermediary must provide a specific

ISSN: 2582-6433

reason beyond violation of community standards. The user should receive a snippet of the specific policy violated and a brief explanation of the AI's logic.

- Public Transparency Reports: Platforms should be required to disclose the error rates—false positives and false negatives of their AI tools across different Indian languages to identify linguistic biases.

6.3 Periodic Algorithmic Audits

Governmental or independent third-party bodies should conduct Constitutional Stress Tests on moderation algorithms.

- Bias Detection: Audits should specifically look for Shadow Banning or systematic suppression of marginalized groups, ensuring that the training data is inclusive and representative of India's pluralistic society.
- Safe Harbour Linkage: Continued Safe Harbour protection under Section 79 of the IT Act should be contingent upon the platform's successful completion of these annual transparency and bias audits.

6.4 Strengthening Grievance Redressal Mechanisms

The right to be heard is fundamental to Article 19(1)(a)

- Fast-Track Appeals: Automated takedowns must have a one-click appeal process that triggers a priority human review.
- Restoration of Speech: If an appeal is successful, the policy should require the platform to not only restore the content but also "re-boost" it to compensate for the algorithmic suppression during the takedown period.

6.5 Digital Literacy and User Empowerment

- Standardized Terms of Service: Mandate that community guidelines be available in all 22 scheduled languages of India, written in plain legal language rather than complex legalese.
- User Controls: Allow users to opt-in or opt-out of certain sensitivity filters. While illegal content must remain moderated, users should have more agency over what borderline content they see, reducing the platform's role as a moral gatekeeper.

6.6 Harmonizing IT Rules with Judicial Standards

The IT Rules, 2021 and subsequent amendments should be refined to ensure that the proactive

ISSN: 2582-6433

monitoring requirement does not devolve into Prior Restraint.

- Narrowing the Scope: Clearly define proactive monitoring to ensure platforms aren't incentivized to over-censor to avoid legal liability.
- Judicial Oversight: For high-stakes takedowns involving public interest or journalistic content, an interim stay mechanism should be available via the Grievance Appellate Committee or a designated digital ombudsman.

By implementing these policies, the legal ecosystem can ensure that Article 19(1)(a) remains a living reality in the digital age, where technology serves as a facilitator of discourse rather than a silent censor.

Chapter 7: Conclusion

The 2026 digital landscape in India represents a watershed moment in the evolution of civil liberties. We have moved decisively from an era of passive intermediary liability to a Zero-Tolerance regime for unregulated Artificial Intelligence. This shift, caused by the February 2026 Amendments to the IT Rules, has fundamentally altered the evolution of the Indian internet. The transition from a 36-hour takedown window to a rigorous 3-hour mandate for AI-generated misinformation and deepfakes has effectively forced social media intermediaries to abandon their roles as transforming them into proactive, algorithm-driven gatekeepers of the public square.

7.1 The Paradox of Protection

While these stringent measures are undeniably vital for protecting Article 21—The Right to Dignity and Privacy—serving as a necessary bulwark against the weaponization of deepfakes and non-consensual synthetic imagery—they have inadvertently created a High-Pressure Environment for Article 19(1)(a). When the cost of a false negative is the total loss of legal safe harbor and potential criminal liability for platform executives, the rational corporate response is defensive over-censorship.

In this climate, algorithmic moderation ceases to be a neutral tool for safety and becomes a structural threat to free expression. The cooling effect is no longer a theoretical concern, it is a systemic byproduct of a legal framework that prioritizes speed over accuracy. The AI, lacking the human capacity for understanding irony, political satire, or cultural nuance, operates as a blunt instrument where a surgical scalpel is required.

The intersection of AI moderation and the rights of the unborn presents a unique challenge:

ISSN: 2582-6433

how to regulate technology to protect potential life while safeguarding the expressive liberties of the living. The following recommendations advocate for a Constitutional-Tech alignment.

7.2 Summary of the Crisis of Scale vs. Rights

The evolution of the "Digital Public Square" has reached a paradox. On one hand, the velocity and volume of online discourse make manual moderation a physical impossibility, necessitating the use of AI. On the other hand, the deployment of these automated systems has introduced a new form of technological censorship. As discussed throughout this paper, AI's inability to grasp the contextual nuances of human speech—such as satire, political dissent, or regional idioms—often leads to the over-blocking of legitimate expression. This results in a chilling effect where the fear of automated shadow-banning or account suspension discourages citizens from exercising their fundamental rights.

7.3 The Constitutional Imperative

The core of this research underscores that Article 19(1)(a) does not lose its potency simply because the medium of expression has changed from paper to pixels. The Indian judiciary, through landmark rulings like Shreya Singhal v. Union of India, has made it clear that restrictions on speech must be narrow and specific. AI moderation, in its current black box form, often operates on broad, vague parameters that would likely fail the test of reasonableness under Article 19(2). When an algorithm suppresses speech without a human understanding of the underlying intent, it ceases to be a tool for public order and becomes an instrument of arbitrary restraint.

7.4 Moving Toward "Rights-by-Design"

The path forward is not to abandon AI, but to subject it to the Rule of Law. This transition requires a shift from purely technical efficiency to Digital Constitutionalism

Key takeaways for the future include:

- Algorithmic Transparency: Platforms must move beyond proprietary secrecy and allow for independent audits of their moderation logic.
- Human Oversight: The principle of "Human-in-the-loop" must be mandated for sensitive content, ensuring that no political or social discourse is terminated by a machine alone.
- Procedural Due Process: Every AI-led takedown must be accompanied by a clear, non-templated explanation and a robust mechanism for appeal, as envisioned by the IT

ISSN: 2582-6433

Rules, 2021.

7.5 Final Outlook

In conclusion, the soul of democracy lies in the ability to disagree, critique, and express. If we allow automated systems to define the boundaries of our speech based on efficiency rather than equity, we risk hollowed-out democratic participation. AI must be viewed as a supplementary tool, not a sovereign judge. To protect the sanctity of Article 19(1)(a), the legal framework must ensure that while the medium of moderation is automated, the standard of moderation remains profoundly human and strictly constitutional. The goal is an internet that is safe, but not silenced.