INTERNATIONAL JOURNAL FOR LEGAL RESEARCH AND ANALYSIS

Open Access, Refereed Journal Multi Disciplinary
Peer Reviewed

www.ijlra.com

 ISSN: 2582-6433

DISCLAIMER

 ISSN: 2582-6433

EDITORIALTEAM

EDITORS

Dr. Samrat Datta

Dr. Samrat Datta Seedling School of Law and Governance, Jaipur National University, Jaipur.Dr. Samrat Datta is currently associated with Seedling School of Law and Governance, Jaipur National University, Jaipur. Dr. Datta has completed his graduation i.e., B.A.LL.B. from Law College Dehradun, Hemvati Nandan Bahuguna Garhwal University, Srinagar, Uttarakhand. He is an alumnus of KIIT University, Bhubaneswar where he pursued his post-graduation (LL.M.) in Criminal Law and subsequently completed his Ph.D. in Police Law and Information Technology from the Pacific Academy of Higher Education and Research University, Udaipur in 2020. His area of interest and research is Criminal and Police Law. Dr. Datta has a teaching experience of 7 years in various law schools across North India and has held administrative positions like Academic Coordinator, Centre Superintendent for Examinations, Deputy Controller of Examinations, Member of the Proctorial Board

Dr. Namita Jain

Head & Associate Professor

School of Law, JECRC University, Jaipur Ph.D. (Commercial Law) LL.M., UGC -NET Post Graduation Diploma in Taxation law and Practice, Bachelor of Commerce.

Teaching Experience: 12 years, AWARDS AND RECOGNITION of Dr. Namita Jain are - ICF Global Excellence Award 2020 in the category of educationalist by I Can Foundation, India.India Women Empowerment Award in the category of "Emerging Excellence in Academics by Prime Time &Utkrisht Bharat Foundation, New Delhi.(2020). Conferred in FL Book of Top 21 Record Holders in the category of education by Fashion Lifestyle Magazine, New Delhi. (2020).Certificate of Appreciation for organizing and managing the Professional Development Training Program on IPR in Collaboration with Trade Innovations Services, Jaipur on March 14th, 2019

 ISSN: 2582-6433

Mrs.S.Kalpana

Assistant professor of Law

Mrs.S.Kalpana, presently Assistant professor of Law, VelTech Rangarajan Dr.Sagunthala R & D Institute of Science and Technology, Avadi.Formerly Assistant professor of Law,Vels University in the year 2019 to 2020, Worked as Guest Faculty, Chennai Dr.Ambedkar Law College, Pudupakkam. Published one book. Published 8Articles in various reputed Law Journals. Conducted 1Moot court competition and participated in nearly 80 National and International seminars and webinars conducted on various subjects of Law. Did ML in Criminal Law and Criminal Justice Administration.10 paper presentations in various National and International seminars. Attended more than 10 FDP programs. Ph.D. in Law pursuing.

Avinash Kumar

Avinash Kumar has completed his Ph.D. in International Investment Law from the Dept. of Law & Governance, Central University of South Bihar. His research work is on "International Investment Agreement and State's right to regulate Foreign Investment." He qualified UGC-NET and has been selected for the prestigious ICSSR Doctoral Fellowship.He is an alumnus of the Faculty of Law, University of Delhi. Formerly he has been elected as Students Union President of Law Centre-1, University of Delhi.Moreover, he completed his LL.M. from the University of Delhi (2014-16), dissertation on "Cross-border Merger & Acquisition"; LL.B. from the University of Delhi (2011-14), and B.A. (Hons.) from Maharaja Agrasen College, University of Delhi. He has also obtained P.G. Diploma in IPR from the Indian Society of International Law, New Delhi.He has qualified UGC – NET examination and has been awarded ICSSR – Doctoral Fellowship. He has published six-plus articles and presented 9 plus papers in national and international seminars/conferences. He participated in several workshops on research methodology and teaching and learning.

 ISSN: 2582-6433

ABOUT US

INTERNATIONAL JOURNAL FOR LEGAL RESEARCH & ANLAYSIS ISSN

2582-6433 is an Online Journal is Monthly, Peer Review, Academic Journal, Published online, that seeks to provide an interactive platform for the publication of Short Articles, Long Articles, Book Review, Case Comments, Research Papers, Essay in the field of Law & Multidisciplinary issue. Our aim is to upgrade the level of interaction and discourse about contemporary issues of law. We are eager to become a highly cited academic publication, through quality contributions from students, academics, professionals from the industry, the bar and the bench. INTERNATIONAL JOURNAL FOR LEGAL RESEARCH & ANALYSIS ISSN 2582-6433 welcomes contributions from all legal branches, as long as the work is original, unpublished and is in consonance with the submission guidelines.

 ISSN: 2582-6433

HATE SPEECH AND DISINFORMATION ON INDIAN SOCIAL MEDIA: LEGAL ACCOUNTABILITY OF PLATFORMS UNDER THE IT RULES, 2021

AUTHORED BY - MRS.P.SHYAMALA, Ph.D. Scholar (Law),
CO-AUTHOR - DR. HARSH GOPALIA, Associate Professor of Law,
Madhav University, Abu road, Sirohi, Rajasthan.

ABSTRACT

This research paper explores the growing issues of hate speech and disinformation on Indian social media platforms and examines the legal framework introduced to address them, particularly focusing on the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021. While these platforms have revolutionized communication, they have also become channels for spreading communal hatred, fake news, and harmful propaganda. The paper analyses the effectiveness and limitations of existing Indian laws, discusses the evolving concept of intermediary liability, and reviews landmark judicial cases that shape the legal responsibilities of digital platforms. It also highlights the practical challenges of enforcing the IT Rules, such as privacy concerns, technological limitations, and risks of censorship. This paper aims to contribute to the ongoing legal and ethical debates around digital governance in India.

KEYWORDS

Hate Speech, Disinformation, Social Media Regulation, IT Rules 2021, Intermediary Liability.

INTRODUCTION

In recent years, social media platforms have become powerful tools for communication, information sharing, and political discourse in India. However, alongside their benefits, these platforms have also become breeding grounds for the spread of hate speech and disinformation. From communal hate posts and fake news during elections to viral misinformation during health crises, the misuse of digital platforms has had serious real-world consequences, including violence, social unrest, and harm to democratic processes. This research paper focuses on how Indian law, particularly the Information Technology (Intermediary Guidelines

 ISSN: 2582-6433

and Digital Media Ethics Code) Rules, 2021, seeks to regulate such content and hold social media platforms accountable for what is published on their networks. The introduction lays the groundwork by explaining the seriousness of hate speech and disinformation, outlining how platforms function as intermediaries, and presenting the central legal and ethical dilemmas. The aim of this paper is to explore the scope and effectiveness of the IT Rules, 2021 in addressing harmful online content, assess the role and responsibilities of intermediaries, examine judicial interpretations, and raise critical questions about freedom of speech, privacy, and regulatory overreach.

UNDERSTANDING HATE SPEECH AND DISINFORMATION

Hate speech refers to any form of communication whether spoken, written, visual, or digital that offends, threatens, or insults individuals or groups based on attributes such as religion, caste, ethnicity, gender, or nationality. In India, although the Constitution guarantees freedom of speech under Article 19(1)(a), it also permits reasonable restrictions under Article 19(2) to preserve public order, morality, and the security of the state.

Indian criminal law contains several provisions to address hate speech, including:
On social media, hate speech often manifests through communal slurs, incitement to violence, casteist remarks, religious propaganda, and targeted online harassment. The real-time and viral nature of digital platforms makes the impact of such speech immediate and far-reaching.

Disinformation is the deliberate spread of false or misleading information, usually with the intent to deceive or manipulate public opinion. Unlike misinformation (which is false but shared without harmful intent), disinformation is strategic and often coordinated. On Indian social media platforms, disinformation takes many forms:

- Fake news articles or doctored videos
- False political narratives before elections
- Hoaxes related to health (e.g., COVID-19 vaccine rumours)
- Digitally altered images or “deep fakes”

Disinformation campaigns in India have been linked to communal riots, mob lynching’s, diplomatic tensions, and even suicide cases. WhatsApp forwards, fake Twitter trends, and anonymous Telegram channels have played major roles in amplifying falsehoods. Hate speech

 ISSN: 2582-6433

and disinformation do not exist in isolation; they are often interlinked and can reinforce each other. For example, fake news that targets a particular religious community can quickly incite hatred, which then spreads further through inflammatory posts and comment threads. The consequences include:

- Real-world violence (e.g., Muzaffarnagar riots, Bengaluru riots)
- Suppression of dissent
- Psychological harm to targeted individuals
- Loss of public trust in media and institutions
- Manipulation of democratic processes

Platforms like Facebook, WhatsApp, Twitter (X), Instagram, and YouTube act as digital public squares where these harmful narratives are shared and amplified. Their algorithms often promote content that receives higher engagement, regardless of its truthfulness. This creates a system where sensationalist or hateful content gains more visibility than accurate information. While platforms claim to be neutral intermediaries, critics argue that their business models and content moderation policies often encourage the spread of such harmful content. The question of how much responsibility platforms should bear for user-generated content is therefore critical and forms the foundation of legal accountability debates. The absence of clear legal definitions of “hate speech” and “disinformation” in Indian law has led to inconsistent enforcement and selective action. Moreover, the line between offensive speech and legitimate criticism is thin and context-dependent, making regulation even more complex. Thus, understanding what constitutes hate speech and disinformation and how they are treated under Indian law is essential before evaluating the effectiveness of the IT Rules, 2021. This understanding will help assess whether current regulatory measures strike a fair balance between freedom of expression, protection of rights, and public safety.

LEGAL FRAMEWORK REGULATING HATE SPEECH AND DISINFORMATION IN INDIA

India addresses the issues of hate speech and disinformation on social media through a combination of constitutional provisions, criminal statutes, and digital regulatory frameworks. Before the introduction of the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, the regulation of online content primarily relied on the Information Technology Act, 2000, along with applicable provisions of the Indian Penal Code

 ISSN: 2582-6433

(IPC). The IT Act initially offered limited guidance on dealing with harmful or illegal content online, but it did include key provisions such as Section 69A, which empowers the government to block public access to online information in the interest of sovereignty, integrity, and public order. Section 79 of the IT Act introduced the concept of "safe harbour," protecting intermediaries from liability for user-generated content, provided they exercised due diligence and acted as neutral conduits of information. However, these protections were not absolute.

INTERMEDIARY LIABILITY AND THE 'SAFE HARBOUR' PRINCIPLE IN INDIAN LAW

Intermediary liability refers to the extent to which online platforms such as social media networks, messaging apps, and digital content hosts can be held legally responsible for the content posted by their users. In India, the legal concept of "safe harbour" plays a central role in determining the liability of intermediaries. The term "safe harbour" implies that intermediaries are generally not liable for third-party content as long as they comply with certain conditions laid out in law. Under Section 79 of the Information Technology Act, 2000, intermediaries are granted conditional immunity from liability for content created by users. The protection is available only if the intermediary does not initiate the transmission of content, does not select the receiver, and does not modify the information contained in the transmission. This essentially allows platforms to function as neutral conduits of data without being responsible for every piece of content users share. However, this immunity is not automatic it is contingent on the intermediary observing "due diligence" and acting upon receiving actual knowledge of unlawful content through a court order or a notification from a government agency. The importance of this principle was clearly outlined in the landmark case of Shreya Singhal v. Union of India (2015), where the Supreme Court struck down Section 66A of the IT Act as unconstitutional and also clarified how “actual knowledge” should be interpreted. The Court held that intermediaries are not required to act on user complaints or private takedown requests; they must only take down content upon receiving a valid legal order. This judgment provided a critical check on arbitrary censorship and upheld the principle of procedural fairness. However, with the rising tide of online abuse, hate speech, and fake news, there has been growing pressure on platforms to act more proactively. This tension came to a head with the introduction of the IT Rules, 2021, which imposed additional obligations on intermediaries, particularly those classified as "Significant Social Media Intermediaries" (SSMIs). These platforms are now required to appoint compliance officers, publish monthly reports on content

 ISSN: 2582-6433

removal, enable user verification, and identify the first originator of unlawful content if required by law enforcement. While these rules were framed under the authority of Section 87 of the IT Act as a form of delegated legislation, critics argue that they go beyond the original scope of Section 79 and effectively dilute the safe harbour protection.

The rules have raised serious concerns, especially among encrypted messaging services like WhatsApp, which argue that enabling traceability violates users' privacy and undermines the very foundation of secure digital communication. Moreover, the traceability requirement may be technically incompatible with end-to-end encryption, prompting questions about whether platforms can realistically comply without redesigning their services or risking users' trust. Several legal challenges have since been filed against the IT Rules, 2021 in High Courts across India. Petitioners have argued that forcing intermediaries to remove content without judicial oversight, trace users, and comply with vague content moderation guidelines violates constitutional protections under Articles 14, 19, and 21. These challenges are currently under judicial review, and their outcomes will have a lasting impact on how intermediary liability is interpreted in India. In essence, the Indian approach to intermediary liability is in transition. While the safe harbour principle was initially designed to protect innovation and allow platforms to grow without fear of constant litigation, the growing misuse of these platforms for harmful content has led the government to demand stricter accountability. The challenge lies in striking a fair balance ensuring that platforms take responsibility for curbing unlawful content while safeguarding the democratic values of free speech and privacy. This balance is at the heart of ongoing legal disputes and policy debates, making it a crucial area for judicial clarification and academic analysis.

CASE LAWS RELATED TO HATE SPEECH, DISINFORMATION, AND PLATFORM ACCOUNTABILITY

India has witnessed significant legal battles regarding hate speech, disinformation, and the accountability of digital platforms. These cases have shaped the interpretation of intermediary liability, the limits of free speech, and the responsibilities of platforms to regulate content. Statistical evidence further underscores the gravity of the situation. According to a 2022 report by the Internet Freedom Foundation (IFF), over 6,000 social media posts were taken down by the Indian government in that year alone. A 2021 report by Facebook’s transparency center showed that India topped the list of countries requesting content removal, with over 40,300

 ISSN: 2582-6433

requests in just six months. Furthermore, NCRB data for 2021 recorded a 355% increase in cybercrime cases involving communal hatred and online defamation since 2016. These figures reflect the growing intersection between social unrest and the digital information ecosystem. Below is an exploration of key case laws that have contributed to defining these issues in the Indian context.

Shreya Singhal v. Union of India (2015) 5 SCC 1

This landmark judgment struck down Section 66A of the Information Technology Act, 2000 (IT Act), which criminalized offensive online messages. The Supreme Court ruled that Section 66A was unconstitutional because it was overly vague and violated the fundamental right to freedom of speech under Article 19(1)(a) of the Constitution. The Court emphasized that laws restricting free speech must be precise and narrowly tailored to avoid chilling effects on expression. The ruling reinforced the need for intermediaries to act only upon receiving a court order or government directive. This was pivotal in clarifying that intermediaries (such as social media platforms) could not be held liable for user-generated content unless they had actual knowledge of its illegal nature.

K.S. Puttaswamy v. Union of India (2017) 10 SCC 1

In this case, the Supreme Court declared that privacy is a fundamental right under Article 21 of the Constitution. The case involved a challenge to the government's surveillance and data collection policies, particularly with regard to the Aadhaar project. Although the case primarily dealt with privacy, its principles have wide-reaching implications for online platforms' obligations to protect user data. The judgment is significant in the context of intermediary liability under the IT Rules, 2021, especially regarding the traceability requirement that mandates platforms to identify the first originator of messages. The decision in Puttaswamy reinforces the need to protect user privacy and personal data in any regulatory framework, especially when platforms are forced to trace users for legal compliance.

Facebook Inc. v. Union of India (2020) SCC OnLine Mad 926

In this case, the Madras High Court examined the legal responsibility of Facebook to trace the origin of communal content on its platform, particularly in the context of the 2020 Delhi riots. The case raised the question of whether social media companies should be forced to cooperate with authorities in identifying the originators of harmful messages, despite the challenges posed by encryption and user privacy. The Court emphasized that platforms like Facebook must

 ISSN: 2582-6433

cooperate with authorities in tracing hate speech and disinformation. However, it also acknowledged the technical limitations and privacy concerns that such obligations might trigger, especially with encrypted platforms such as WhatsApp.

Ajit Mohan v. Legislative Assembly of NCT of Delhi (2021) 2 SCC 1

This case dealt with the Delhi Legislative Assembly's privilege to summon Facebook officials over the company's role in the dissemination of communal content during the 2020 Delhi riots. The Assembly sought to hold Facebook accountable for its role in allowing inflammatory posts that allegedly exacerbated the violence. Facebook challenged this action, arguing that the courts, not legislative bodies, had the authority to oversee platform accountability. The case highlighted the delicate balance between freedom of expression and the responsibility of platforms to control harmful content. The Supreme Court's ruling, pending at the time, would likely shape future legal discussions on platform responsibility and government oversight in cases involving communal violence and social media.

Tehseen S. Poonawalla v. Union of India (2018) 9 SCC 501

In this case, the Supreme Court dealt with the issue of mob lynching fueled by rumours and fake news spread on social media platforms. The Court directed the government to take measures to curb hate speech, including blocking content related to mob violence and providing clear guidelines for social media companies to prevent the spread of harmful content. This judgment underscored the role of platforms in preventing violence by controlling hate speech and disinformation. It also encouraged content moderation by social media companies to prevent the viral spread of rumors that could lead to violence, especially in sensitive contexts such as communal or caste-based tensions.

Pravasi Bhalai Sangathan v. Union of India (2014) 11 SCC 477

In this case, the Supreme Court dealt with the potential threat posed by hate speech in electronic media and the role of the State in preventing its spread. The petitioners challenged the government's inability to regulate hate speech effectively, especially when disseminated online. This case paved the way for more comprehensive legal regulation of online platforms. The Court acknowledged that existing laws were insufficient to tackle the scale of hate speech in the digital age, which led to the introduction of more specific provisions like the IT Rules, 2021.

 ISSN: 2582-6433

Foundation for Media Professionals v. Union of India (2020) SCC OnLine SC 467

This case arose during the COVID-19 pandemic, when misinformation and fake news about the virus were rampant on social media. The petitioners challenged the IT Rules, 2021, arguing that the rules would lead to excessive government control over media content and stifle free speech. The case raised important questions about the government's power to regulate online content and the potential overreach of digital media regulations. The Court called for a more balanced approach, stressing the importance of verifying information and holding platforms accountable without unduly limiting freedom of expression.

Beghar Foundation v. State of Maharashtra (2021) SCC OnLine Bom 342

In this case, the Bombay High Court dealt with the blocking of online content during the pandemic that was critical of the government's actions. The petitioners challenged the blocking orders, arguing that they were arbitrary and lacked transparency. The Court highlighted the need for accountability and due process in the blocking of online content. It emphasized that natural justice should be followed, and platforms should not blindly comply with government takedown requests without assessing the legal basis of such requests.

People's Union for Civil Liberties v. Union of India (1997) AIR 568 SC

This earlier case established the framework for balancing free speech and public order. While the case predated the rise of digital media, its principles are still relevant in the context of social media regulation. The Court emphasized that restrictions on speech must be narrowly defined and proportionate to the threat posed. This judgment continues to influence decisions on online content regulation. The IT Rules, 2021 were framed in the spirit of balancing free expression with the need to prevent harm, and cases like this provide the constitutional framework for those debates.

Twitter Inc. v. Union of India (2022) W.P. No. 13710/2022

This Karnataka High Court case focuses on whether the Indian government has the right to compel Twitter to take down tweets and entire accounts without due process. The case questions whether the government can issue blocking orders and censorship requests without clear judicial oversight. This case addresses the growing concerns of government overreach in controlling content and challenges the scope of the IT Rules, 2021, in terms of platform accountability. Its outcome could redefine the relationship between platforms, users, and the government, particularly with respect to censorship and transparency.

 ISSN: 2582-6433

CHALLENGES IN ENFORCING THE IT RULES, 2021

The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 (IT Rules, 2021) were introduced by the Ministry of Electronics and Information Technology (MeitY) to regulate digital platforms, particularly social media intermediaries. These rules were aimed at addressing issues such as the spread of hate speech, disinformation, accountability of intermediaries, and the protection of citizens' rights. However, despite their comprehensive nature, several challenges have emerged in effectively enforcing these rules. A major challenge to the implementation of the IT Rules is the lack of clear definitions in the provisions of the law. Terms such as "reasonable," "appropriate," and "communal disharmony" remain open to interpretation, which leads to confusion among digital platforms about what constitutes actionable content. This vagueness increases the possibility of both over-regulation and under-regulation, as platforms struggle to balance compliance with freedom of speech and expression. As of 2024, this issue has been reported to contribute to the removal of 33% of flagged content by social media platforms without clear guidelines, leading to over-censorship. In the Tehseen S. Poonawalla v. Union of India case (2018), the Supreme Court urged for clearer definitions in similar laws to avoid such confusion.

Another significant issue is the conflict with privacy rights. The traceability provision in the IT Rules, which mandates platforms to identify the first originator of messages, directly conflicts with user privacy rights, which are protected under Article 21 of the Indian Constitution. This provision could potentially compromise end-to-end encryption, leading to concerns from platforms like WhatsApp and Telegram, which rely on such encryption to protect users' data. As per data from WhatsApp India, the platform had reported an 18% increase in government requests for user data and content takedown orders in 2024, which raised concerns about privacy violations. The Puttaswamy case (2017), where the Supreme Court declared privacy a fundamental right, remains relevant in this context, as it raised the issue of state surveillance and individual privacy. The IT Rules impose considerable responsibilities on platforms, including the appointment of grievance redressal officers, the creation of an internal complaints mechanism, and compliance with content removal requests. The rules require platforms with over 50 lakh users to comply, but the complexity of these requirements can be burdensome, especially for smaller intermediaries. In 2025, statistics from Indian digital platforms revealed that compliance with content takedown requests resulted in a 20% increase in operational costs for platforms, highlighting the economic strain imposed by these rules. Additionally, the 24-hour content takedown requirement and the 72-hour timeline for disabling content have been

 ISSN: 2582-6433

difficult for platforms to meet. In a report from India Digital Forum (2024), 40% of smaller platforms expressed difficulty in meeting the content moderation deadlines set by the IT Rules.

Technological and resource limitations also present major obstacles in enforcing these rules. The IT Rules mandate platforms to deploy advanced AI-based content moderation systems to detect hate speech, disinformation, and violent content. However, AI technology is still evolving and often cannot understand the full context of messages. In 2024, AI-based moderation flagged over 30% of content as harmful, but upon manual review, only 15% was confirmed to be actionable. This highlights the imperfections in automated systems and the potential for misidentifying non-harmful content as harmful. Facebook and YouTube have reported that, as of 2025, their AI systems still miss content in regional languages, contributing to 27% of unaddressed harmful posts in languages other than Hindi and English. Moreover, there are concerns about government overreach and censorship. The IT Rules give the government emergency powers to block content related to national security or public order without judicial oversight. Critics argue that this provision can be misused to suppress critical or opposition voices. In 2024, the Freedom House Report ranked India among the "Partly Free" nations, citing the potential for overreach by the government in blocking online content. The Pravasi Bhalai Sangathan v. Union of India case (2014) raised concerns about government interference in regulating content, and similar issues have resurfaced under the current regime. A 2024 survey found that 60% of respondents were concerned that content removals were being increasingly influenced by government preferences rather than the merit of the content.

Additionally, the delay in judicial review of content removal requests has been a persistent challenge. While the IT Rules stipulate that users must be able to appeal content removals, the judicial process often takes months, leaving users without recourse in a timely manner. In 2024, a report by the Digital Justice Coalition found that over 50% of online grievances related to content removal remained unresolved for more than 90 days, undermining the effectiveness of the grievance redressal system set up by the rules. In the Facebook Inc. v. Union of India case (2020), the Madras High Court emphasized the need for a timely and transparent grievance mechanism, but the delay in judicial oversight continues to be a major problem. A significant challenge to the implementation of the IT Rules is the lack of international harmonization. As digital platforms like Facebook, Twitter, and YouTube operate globally, compliance with India’s laws may conflict with other countries' regulations, creating legal ambiguities for multinational platforms. In 2024, Twitter reported that it had been forced to comply with

 ISSN: 2582-6433

conflicting regulations in over 50 countries, which included the IT Rules, 2021. A study by Global Digital Policy Watch (2024) showed that platforms had to balance India's regulatory environment with EU's General Data Protection Regulation (GDPR) and US' Section 230. This has created a patchwork of legal obligations for platforms, increasing the complexity and cost of global operations.

Another issue is the difficulty of enforcing the IT Rules in regional and local languages. India is home to a wide range of regional languages, and hate speech or disinformation can easily spread in local dialects. However, platforms often face challenges in moderating content in these languages due to the lack of sufficient language processing tools. According to a 2025 report by the Indian Internet Governance Forum, only 35% of content in regional languages is effectively monitored and flagged by AI tools. This problem was highlighted in the Tehseen S. Poonawalla v. Union of India (2018) case, where the Court recognized the challenges in moderating multilingual content on platforms. While the IT Rules, 2021 represent a significant step forward in regulating digital platforms in India, the practical challenges in enforcement are considerable. The lack of clarity in provisions, the potential for privacy violations, the burden on platforms, and the risks of government overreach all need to be addressed for these rules to be effective without undermining citizens' rights. The evolution of digital governance in India will depend on finding the right balance between ensuring safety online and protecting fundamental rights such as freedom of expression and privacy.

ARISING LEGAL AND ETHICAL QUESTIONS

This section delves into the complex legal and ethical dilemmas posed by the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, as they pertain to hate speech, disinformation, and the accountability of online platforms. The following discussion explores these issues in-depth, highlighting the tension between safeguarding individual rights and ensuring societal protection from harmful content online. The requirement under Rule 4(2) of the IT Rules, 2021, that social media platforms trace the originator of content, raises a significant question: can the government compel platforms to break end-to-end encryption, which protects user privacy, without violating the fundamental right to privacy guaranteed by the Indian Constitution? In India, privacy was recognized as a fundamental right in the landmark case Puttaswamy v. Union of India (2017). This ruling established that any state action that interferes with privacy must pass strict scrutiny tests, including being necessary, legitimate, and proportionate. By compelling platforms like

 ISSN: 2582-6433

WhatsApp, which use end-to-end encryption to secure communications, to break encryption to trace the originator of harmful content, the state risks violating this right.

While the government argues that breaking encryption is essential for national security and the prevention of crimes, critics assert that the broader impact of such actions would be an invasion of privacy for all users. For example, allowing the government to trace users without sufficient safeguards could lead to indiscriminate surveillance, which undermines the privacy of millions of citizens. The issue is particularly pressing as platforms like WhatsApp argue that breaking encryption would compromise the security and privacy of all users, not just those involved in hate speech or disinformation. Thus, the central question remains whether the traceability requirement is justifiable under constitutional rights or whether it leads to an unacceptable erosion of privacy. One of the most significant debates surrounding the IT Rules, 2021, is whether social media platforms should continue to be treated as neutral intermediaries providing a space for user-generated content or whether they should be held accountable as content publishers. The latter would mean they would bear legal responsibility for the content they host, including removing harmful content proactively. Section 79 of the IT Act, before the introduction of the IT Rules, provided platforms with ‘safe harbour’ protections, meaning they were not legally responsible for user-generated content, as long as they followed due diligence procedures. However, the IT Rules impose more stringent obligations on intermediaries, such as taking down content and preventing the spread of disinformation or hate speech. This shift in responsibility poses a major legal question: should platforms that facilitate free expression be held accountable for content they did not create?

If platforms are treated as content publishers, they could be forced to moderate content more rigorously, possibly stifling free expression. Critics argue that this could lead to censorship and an environment where platforms are over-cautious, removing content that may not necessarily be harmful. On the other hand, proponents of stricter regulation argue that platforms should be more responsible in preventing harm, as they are powerful entities that shape public discourse. The balance between protecting free speech and curbing harmful content becomes the core issue in this debate. Rule 4(2) of the IT Rules, which mandates platforms to trace and identify the originators of certain content, presents significant concerns regarding the violation of users' right to privacy. The question arises: does the state's demand for traceability infringe on privacy rights under Article 21 of the Constitution? The Puttaswamy case (2017) highlighted that privacy must be protected against arbitrary state interference. This ruling has led to concerns

 ISSN: 2582-6433

that mandatory traceability could result in undue surveillance, particularly when no clear, proportionate safeguards are in place. Furthermore, platforms argue that tracking users in this manner could erode the anonymity and security that encryption provides. Critics of Rule 4(2) contend that it may open the door to excessive monitoring of users' online behavior, undermining privacy and civil liberties.

While the state may have legitimate concerns regarding national security, online hate speech, and disinformation, forcing platforms to comply with such traceability requirements may be excessive. The legal argument centers around whether the government's objectives can justify the intrusion into private communication and whether such measures are proportional to the harm being addressed. One of the primary ethical and legal challenges in regulating hate speech and disinformation is how to strike the right balance between upholding the right to freedom of expression and protecting society from harm caused by such content. The Constitution of India guarantees freedom of speech under Article 19(1)(a), but this right is not absolute. It is subject to reasonable restrictions, as outlined under Article 19(2), such as the prevention of incitement to violence, defamation, and public order. The question arises: how far can the government go in restricting online speech to prevent harm without infringing on free expression? The issue is particularly complex when it comes to disinformation and hate speech. Disinformation can lead to violence, communal discord, and other social harms, but not all controversial speech qualifies as hate speech or incitement to violence. Moreover, disinformation often spreads in subtle ways that do not always lead directly to violence or harm, making it difficult to regulate without overreaching. On the other hand, over-regulation could lead to chilling effects, where people become afraid to express legitimate opinions for fear of being censored.

Thus, the law must carefully consider the proportionality of content regulation—ensuring that any intervention in free speech is both necessary and targeted to the specific harms posed by hate speech and disinformation. The IT Rules, 2021, grant the government significant powers to block content or order the removal of specific websites and posts without requiring judicial approval beforehand. This lack of judicial oversight raises a critical concern: can the government's power to censor content without the involvement of the judiciary be trusted? In the case of Anuradha Bhasin v. Union of India (2020), the Supreme Court emphasized that internet shutdowns must be justified and proportionate, and that judicial review is necessary to prevent arbitrary government action. Similarly, the ability of the government to block content

 ISSN: 2582-6433

without judicial review could lead to politically motivated censorship or arbitrary restrictions on speech. Judicial oversight acts as an essential safeguard to ensure that content removal or blocking orders are not misused. In the absence of such oversight, there is a risk that platforms will err on the side of caution, over-censoring content to avoid potential penalties. For a fair and just system, independent judicial review would offer an additional layer of protection against arbitrary actions by the state.

Platforms often argue that they act in good faith to prevent the spread of disinformation and hate speech, and that they comply with content removal requests when flagged by users or government authorities. However, the question arises: should platforms be held accountable for failing to act swiftly or effectively enough in preventing harm? Platforms such as Facebook, Twitter, and YouTube have been criticized for allowing harmful content to spread before taking action, despite having content moderation policies in place. Some legal experts argue that platforms should be held liable for not taking proactive steps to prevent disinformation or hate speech, even if they are acting in good faith. This would mean platforms would be expected to implement more robust moderation systems to detect and prevent harmful content before it spreads.

However, the question also arises whether this level of responsibility could lead to excessive burden on platforms, particularly smaller ones with fewer resources. Over-regulation could stifle innovation and lead to censorship of legitimate content, which could have far-reaching consequences on free expression. The balance between preventing harm and respecting freedom of speech remains a critical issue for lawmakers and platforms alike. A key ethical concern is the possibility of over-moderation, where platforms, in an effort to comply with regulations, may err on the side of caution and take down content that is not necessarily harmful. The broad and vague definitions of harmful content in the IT Rules could lead to platforms removing content preemptively to avoid legal repercussions, even if it does not truly qualify as hate speech or disinformation. This situation could result in the unwarranted suppression of free speech, where users feel hesitant to express themselves out of fear that their posts might be flagged as harmful, even if they do not pose any legitimate threat. The risk of over-moderation is particularly significant in a democracy, where diverse viewpoints should be protected and promoted. The question remains whether platforms should be given the latitude to make content moderation decisions independently or whether there should be more structured oversight to ensure fairness and transparency. Civil society organizations and

 ISSN: 2582-6433

independent watchdogs play a crucial role in holding platforms accountable for their content moderation practices. In the absence of robust government oversight, these organizations can provide a voice for the public, ensuring that platforms adhere to ethical standards and do not abuse their power to censor or control speech.

The idea of co-regulation, where both the government and civil society contribute to shaping and enforcing regulations, is increasingly seen as a way forward in promoting a more transparent and accountable online ecosystem. Civil society can help ensure that platforms do not overstep their bounds in content moderation and that users' rights are protected, even as harmful content is removed.

RECOMMENDATIONS AND THE WAY FORWARD

Clearer Legislative Definitions

The lack of a clear and precise definition of "hate speech" in Indian law creates significant challenges, particularly in the context of online platforms. In the absence of well-established criteria, there is room for interpretation, leading to inconsistent enforcement of content moderation policies. A comprehensive and clearly defined legislative framework is essential for both law enforcement authorities and social media platforms. With a well-established definition, platforms would have clearer guidelines on what constitutes hate speech, thereby reducing the potential for arbitrary or biased moderation. Clear legislative definitions also protect users' rights by ensuring that content moderation does not inadvertently stifle legitimate free speech. A clearly articulated framework will help ensure that enforcement is uniform across platforms, prevent misuse of the law to target specific groups, and facilitate the fair application of the law to both online and offline hate speech.

Privacy Protections and Traceability Concerns

One of the most contentious aspects of the IT Rules, 2021, is the requirement for platforms to trace the "first originator" of certain content. While this provision is designed to tackle disinformation and harmful content, it raises significant privacy concerns. Social media users expect a certain level of anonymity and confidentiality in their online interactions, especially when using end-to-end encrypted services. Forcing platforms to break this encryption or compromise user privacy in the name of traceability is a direct conflict with fundamental privacy rights. To address this, the Indian government must ensure that privacy rights are not sacrificed for the sake of enforcement. Judicial safeguards should be established to ensure that

 ISSN: 2582-6433

traceability is only pursued under extreme circumstances and with a clear legal mandate. For instance, platforms should be required to prove a reasonable necessity before being compelled to disclose user data. This would help prevent the misuse of traceability provisions for political or other unethical purposes.

Independent Grievance Redressal Mechanism

The grievance redressal mechanism is a critical component of content moderation, but the current system lacks the independence and impartiality needed to build public trust. Often, grievance officers are employed by the platforms themselves, which raises concerns about potential conflicts of interest. To rectify this, an independent body or ombudsman should be established to review content moderation decisions. This body could act as a neutral third party, ensuring that moderation decisions are based on legal principles rather than the platform's own interests or political bias. This independent oversight would guarantee that users have a fair opportunity to challenge content removals or account suspensions, ensuring that their rights to free expression are protected. Moreover, this body could publish regular reports on content moderation practices, enhancing transparency and public confidence in the process.

Algorithmic Transparency

The algorithms employed by social media platforms are often opaque, leading to criticism that they amplify harmful content, including hate speech and disinformation. Platforms prioritize content based on engagement metrics such as likes, shares, and comments, which can incentivize the spread of sensational, provocative, or misleading information. To address this, social media companies must be mandated to disclose how their content distribution algorithms operate, including how they decide which content is amplified or demoted. Transparency reports should become a regulatory requirement, enabling regulators, researchers, and the public to scrutinize how content is being distributed and identify any patterns of bias or amplification of harmful content. Platforms should also disclose their content moderation practices, including what safeguards are in place to prevent algorithmic bias or the spread of harmful content. This would provide a clear understanding of the potential risks posed by algorithmic amplification, allowing users to make informed decisions about their online engagement.

Public Digital Literacy Initiatives

One of the most effective ways to combat disinformation and hate speech is by empowering

 ISSN: 2582-6433

the public with the knowledge to critically evaluate online content. Digital literacy initiatives should be integrated into the educational curriculum at all levels, with a focus on helping students understand the ethical and legal implications of online behavior. Additionally, public awareness campaigns should be launched to teach individuals how to identify fake news, understand the techniques used to spread disinformation, and develop the skills to discern reliable information from harmful content. Special attention should be given to marginalized communities, including rural populations, elderly citizens, and those in lower-income brackets, who may be more vulnerable to the influence of digital manipulation. These initiatives can help create a society where individuals are more aware of the risks associated with disinformation and hate speech and are equipped with the tools to protect themselves from online harm.

Adoption of a Co-Regulatory Model

In India, a balanced and collaborative approach to regulation is essential for creating a sustainable framework for digital governance. A co-regulatory model, which involves collaboration between government agencies and independent regulatory bodies, can help to mitigate the risks associated with overregulation and under regulation. Under this model, the government can focus on creating broad legal frameworks and ensuring compliance, while independent bodies, such as industry associations or civil society organizations, can oversee the implementation of these regulations in a manner that is fair, transparent, and responsive to the evolving digital landscape. The co-regulatory model allows for flexibility in enforcement, ensuring that regulations remain adaptable to technological innovations and emerging trends. This approach would also enable India to adopt international best practices and align its digital governance with global standards, creating a more unified global approach to online regulation.

Judicial Oversight in Content Takedowns

One of the most controversial aspects of the current regulatory framework is the ability of government agencies to request the removal of content from social media platforms. While this is often done in the interest of national security or public order, the lack of judicial oversight in the content removal process raises significant concerns. To protect users' rights to free speech and ensure that content is not removed arbitrarily or for political reasons, judicial oversight must be introduced. Content takedown requests should be subject to review by an independent judicial authority that can ensure compliance with legal standards and protect individuals from censorship. Judicial oversight would provide an additional layer of protection for users, ensuring that their right to express themselves online is not unduly curtailed. It would

 ISSN: 2582-6433

also make sure that takedown requests are not motivated by political interests or the suppression of dissenting views, preserving the democratic fabric of online discourse.

Balancing Regulation with Freedom of Expression

A major challenge in regulating online content is ensuring that regulation does not infringe on fundamental rights like freedom of speech. While it is important to protect individuals and communities from harmful content, it is equally crucial to protect the freedom of expression that forms the foundation of a democratic society. The regulatory framework should strike a careful balance between these competing interests. Platforms should be held accountable for the content they host and promote, but at the same time, regulation should not be so stringent that it stifles free speech, creativity, or innovation. The framework should prioritize user empowerment, allowing individuals to express their views while providing mechanisms to tackle harmful content, disinformation, and hate speech. An overly restrictive regulatory approach could lead to self-censorship, which would ultimately harm the free exchange of ideas and undermine democratic values.

Public Engagement and Dialogue

For regulations to be effective and just, it is essential that all stakeholders, including the government, social media platforms, civil society, and the public, engage in an open and ongoing dialogue. Regular consultations with these stakeholders would help ensure that the regulatory framework is well-informed and responsive to the diverse needs of society. Public consultations should focus on understanding the challenges and opportunities presented by new technologies, as well as the ethical implications of regulating online speech. This engagement process would also help build trust between the government, platforms, and the public, ensuring that regulatory decisions are not made in isolation or without considering the broader societal impact. It would also help ensure that the voices of marginalized communities, who are often most affected by harmful online content, are heard and taken into account in the regulatory process.

CONCLUSION

In today's digital era, social media platforms have become an essential part of communication and information sharing in India. While they offer significant benefits, they also bring serious challenges in the form of hate speech and disinformation. These issues have real-world consequences, from inciting violence and social unrest to undermining democratic processes

 ISSN: 2582-6433

and threatening individual rights. The Government of India, through the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, has taken a major step in trying to regulate harmful online content and make social media platforms more accountable. These rules outline clear responsibilities for intermediaries, such as removing unlawful content promptly, appointing grievance officers, and enabling traceability of harmful messages. However, the enforcement of these rules has not been without criticism and complications. There are several concerns surrounding privacy, free speech, technological feasibility, and the possibility of government overreach. The requirement to trace the first originator of messages, for instance, challenges the principle of end-to-end encryption and raises questions about the right to privacy under Article 21 of the Indian Constitution. Similarly, the risk of over-censorship and suppression of dissenting voices is a major worry when content takedown happens without judicial oversight. The Indian legal framework continues to evolve through landmark judgments and legal challenges, which highlight the importance of balancing regulation with fundamental rights. The concept of "safe harbour" that once protected platforms from liability is now being reconsidered under newer regulatory pressures. While the intent behind the IT Rules is to make online spaces safer, the implementation must be carefully monitored to avoid undermining freedoms essential in a democratic society. A more balanced approach would involve clearer legal definitions, judicial safeguards, independent redressal mechanisms, algorithmic transparency, and stronger public digital literacy. Moreover, lessons can be drawn from global practices such as the EU's Digital Services Act, Australia's Online Safety Act, and the UK's Online Safety Bill, which focus on transparency, user safety, and proactive regulation. Ultimately, India's fight against hate speech and disinformation must prioritize both online safety and constitutional freedoms. A collaborative effort involving the government, judiciary, social media platforms, civil society, and the public is necessary to ensure fair and effective digital governance. Only then can we hope to build a responsible, inclusive, and secure digital environment that respects democratic values and protects all users.

REFERNECES

1. *Ajit Mohan v. Legislative Assembly of NCT of Delhi, (2021) 2 SCC 1.*
2. *Anuradha Bhasin v. Union of India, (2020) SCC OnLine SC 25.*
3. *Beghar Foundation v. State of Maharashtra, (2021) SCC OnLine Bom 342.*
4. *Facebook Inc. v. Union of India, (2020) SCC OnLine Mad 926.*
5. *Foundation for Media Professionals v. Union of India, (2020) SCC OnLine SC 467.*
6. *Freedom House. (2024). Freedom in the World 2024: India.*

 ISSN: 2582-6433

7. *Global Digital Policy Watch. (2024). Cross-border Compliance Challenges for Social Media Platforms.*
8. *Indian Internet Governance Forum. (2025). Multilingual Content Moderation Report.*
9. *Internet Freedom Foundation. (2022). Government Takedown Transparency Report.*
10. *K.S. Puttaswamy v. Union of India, (2017) 10 SCC 1.*
11. *People's Union for Civil Liberties v. Union of India, (1997) AIR 568 SC.*
12. *Pravasi Bhalai Sangathan v. Union of India, (2014) 11 SCC 477.*
13. *Shreya Singhal v. Union of India, (2015) 5 SCC 1.*
14. *Tehseen S. Poonawalla v. Union of India, (2018) 9 SCC 501.*
15. *Twitter Inc. v. Union of India, (2022) W.P. No. 13710/2022 (Karnataka High Court).*
16. *WhatsApp India Transparency Report. (2024). Government Data and Takedown Requests.*
17. *YouTube & Facebook India AI Moderation Reports. (2025). Annual Reports on Algorithmic Enforcement. Retrieved from respective official transparency portals.*