Page 4629 - 4648 **DOI:** *https://doij.org/10.10000/IJLMH.119434*

INTERNATIONAL JOURNAL OF LAW MANAGEMENT & HUMANITIES

[ISSN 2581-5369]

Volume 8 | Issue 2

2025

The Role of Intermediaries and Social Media Platforms in Curbing Deepfake Circulation: A Legal Analysis Under Indian It Laws

ANUJ SINGH YADAV[1] AND DR. SHOVA DEVI[2]

ABSTRACT

The spread of deepfake technology has created tremendous legal issues, especially regarding the possibility of misuse on social media sites. Deepfakes based on artificial intelligence for creating forged media are potential threats to privacy, reputation, and national security. This research discusses the use of intermediaries and social media websites in inhibiting the dissemination of deepfakes under Indian IT laws. It investigates the legal mechanism offered by the Information Technology Act, 2000 (IT Act) and associated regulations to determine the intermediary's liability to curb the spread of dangerous deepfakes. The paper deconstructs the role of platforms in detecting, deleting, and stopping the re-uploading of deepfake materials, while addressing free speech and privacy rights. The research further examines the existing gaps in the law and proposes regulatory responses to address the emergent threat of deepfake technology.

Keywords*: Deepfake, Social Media Platforms, Intermediaries, Legal Framework, Information Technology Act, IT Laws, Privacy, Cybersecurity, Fake News, Content Moderation, Digital Ethics, Free Speech, India, Liability.*

I. INTRODUCTION

The fast pace of development of artificial intelligence (AI) and machine learning technologies has transformed digital content creation, making not only innovation across industries possible but also leading to the development of advanced and misleading technologies like deepfakes.[3] Deepfakes are synthetic media—photos, audio, or video—created using AI that manipulate reality to produce false but convincingly realistic content. While the technology per se promises innovative and learning purposes, its ill-use to promote false information, defamations, monetary deceptions, and political machinations poses an urgent danger to persons,

[1] Author is a student at Amity Law School, India.
[2] Author is an Assistant Professor at Amity Law School, India.
[3] Karen Hao, "What Is Machine Learning?" *MIT Technology Review*, February 11, 2018, https://www.technologyreview.com/2018/02/11/146020/what-is-machine-learning/.

 [ISSN 2581-5369]

organizations, and democratic functions. For the Indian case, the circulation of deepfakes has evoked great legal and moral challenges, requiring an honest appraisal of intermediaries' and social media platforms' responsibilities to counter such malignant distribution.[4]

Within the digital environment, intermediaries like Internet Service Providers (ISPs), web hosts, search engines, and particularly social media sites have a role to play in making user-generated content available.[5] Their ethical and legal responsibilities to monitor and delete dangerous or illegal content are now more pressing in the face of increasing misuse of deepfake tools. The Information Technology Act, 2000 (IT Act), and the following amendments thereto, especially the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, constitute the basic legal framework for the intermediaries' liability and responsibility in India. Such rules impose intermediaries with due diligence obligations to undertake required actions for identifying and averting the spreading of illegal content, such as that which is defamatory, obscene, or prejudicial to public order. With deepfakes generally employed for sexual exploitation, political propaganda, and financial frauds, intermediaries are in increased focus for keeping such content away from going viral.[6]

Legal status of intermediaries and the range of their liability is a dynamic area of Indian jurisprudence. Intermediaries are provided conditional "safe harbor" under Section 79 of the IT Act, and they are relieved of liability for third-party data hosted on their sites as long as they maintain due diligence and do not possess actual knowledge about the unlawful material. Yet, the meaning of "actual knowledge" has been problematic, particularly following landmark judgments like Shreya Singhal v. Union of India, which invalidated Section 66A of the IT Act but clarified intermediary liability in relation to content removal. The deepfake phenomenon tests this protection in law, and whether or not the intermediaries can continue to be mere passive conduits of information or if they need to play an active role in identifying and curbing the dissemination of AI-manipulated content.[7]

The 2021 Intermediary Guidelines also added specific obligations that affect the liability scenario considerably.[8] These are the appointment of a Chief Compliance Officer, a mechanism

[4] Ministry of Electronics and Information Technology, Government of India, "Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021," https://www.meity.gov.in.

[5] Graham Greenleaf, "Global Data Privacy Laws 2021: Despite COVID Delays, 145 Laws Show GDPR Dominance," *Privacy Laws & Business International Report*, no. 170 (2021): 10–13.

[6] BBC News, "India's Deepfake Menace: Fake Videos, Real Threats," *BBC*, March 8, 2023, https://www.bbc.com/news/world-asia-india-deepfakes.

[7] Nikhil Pahwa, "Deepfakes and India's Weak Digital Laws," *Medianama*, March 15, 2023, https://www.medianama.com/2023/03/223-deepfakes-india-legal-framework/.

[8] Ministry of Electronics and Information Technology, "Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021," Government of India, https://www.meity.gov.in/writereaddata/files/Inte

for redressal of grievances, and the provision to facilitate identification of the "first originator" of a message for major social media intermediaries. For deepfakes, identification becomes key to tracking the origin of the manipulated content and holding the perpetrator accountable. However, it also invokes privacy and surveillance issues, particularly with respect to end-to-end encrypted services like WhatsApp, who contend that such requirements encroach on user privacy and encryption guidelines. Therefore, the legal juggling act of curbing injurious deepfakes while protecting digital rights turns into a complex and delicate challenge.[9]

Most frequent vectors for sharing deepfake material are social media sites like Facebook, Twitter (now known as X), Instagram, and YouTube because of the enormous user population and instant sharing of content in these sites. [10]In spite of their international content moderation practices and AI-based filtering of content systems, these platforms have a herculean task to make a distinction between innocuous, parodic, or satiric content and content maliciously manipulated. The Indian legal system, while becoming more active in enforcing accountability on platforms, remains devoid of deepfake-specific laws or elaborate guidelines for algorithmic content moderation. Consequently, the moderation is unevenly apportioned and mostly reactive and not preventive in nature, thereby enabling harmful content to spread widely before it gets removed.[11]

II. DEEPFAKES: TECHNOLOGY AND IMPLICATIONS

Deepfake technology is the outcome of sophisticated artificial intelligence, most notably deep learning algorithms like Generative Adversarial Networks (GANs).[12] These algorithms can produce hyper-realistic synthetic media by learning off large collections of actual images, audio, or video and generating new material replicating the original with remarkable precision. In essence, deepfakes can create speech, facial movements, body language, and even full video clips, which is very hard for the ordinary viewer to separate real from manipulated content. Deepfakes were initially created for entertainment and creative industries, but later they have been used for malicious purposes, such as political propaganda, identity theft, revenge

rmediary_Guidelines_and_Digital_Media_Ethics_Code_Rules-2021.pdf.

[9] Pranesh Prakash, "Balancing Surveillance and Privacy in the Age of AI: The Indian Context," *Internet Democracy Project*, September 2021, https://internetdemocracy.in/reports/balancing-surveillance-and-privacy-in-the-age-of-ai/.

[10] Laura Garcia and Tom Southern, "What Are Deepfakes – and How Can You Spot Them?" *BBC Bitesize*, May 3, 2023, https://www.bbc.co.uk/bitesize/articles/z6c2r2p.

[11] Sarbani Banerjee Belur and Sunil Abraham, "Content Moderation in India: Challenges in Platform Governance," *Carnegie India*, July 21, 2022, https://carnegieindia.org/2022/07/21/content-moderation-in-india-challenges-in-platform-governance-pub-87473.

[12] Ian Goodfellow et al., "Generative Adversarial Nets," *Advances in Neural Information Processing Systems* 27 (2014): 2672–2680.

pornography, financial scams, and disinformation campaigns. This double-edged nature of the technology—innovative but risky—poses grave challenges to lawmakers, online platforms, and society as a whole, especially in democracies such as India where social media penetration is huge.[13]

The consequences of deepfakes are wide-ranging, not just for individual damage to individuals but also for their capacity to destabilize institutions and undermine public trust. [14]Deepfakes are employed to develop non-consent pornographic content, mostly aiming at women and public figures and causing extreme mental and reputational harm. Deepfakes used in political context can be made to spread views of politicians or public officials issuing or doing anything they never made or did. They have the capability to inflame violence, communal discord, or upset the integrity of electoral processes. These threats are especially dire in India, where political polarisation and communal sensitivities run high, and where digital literacy among citizens remains in development. The viral character of deepfakes on the WhatsApp, Facebook, and Instagram platforms enables disinformation to move quickly, typically ahead of fact-checkers and regulatory bodies being able to react.[15]

Technologically, the increased sophistication of deepfake generators complicates detection with each passing day.[16] Whereas initial deepfakes were detectable by discrepancies in blinking, artificial facial movements, or audio inconsistencies, newer versions are extremely sophisticated, frequently evading even automated detection mechanisms. This arms race between deepfake producers and detection mechanisms poses a dynamic challenge to intermediaries and social media platforms, which are supposed to be the first line of defense. However, the algorithms employed by these sites to moderate content are not perfect; they may miss harmful content or, on the other hand, flag legitimate content as false positives, creating problems of both effectiveness and censorship. Since India currently does not have deepfake-specific legal provisions, websites have to depend on a quilt of prevalent IT legislation and directives, which might not be capable of handling the technological subtleties and ethical implications of deepfake incidents.[17]

[13] Anurag Kotoky, "India's Deepfake Problem Shows Urgent Need for Policy Action," *Bloomberg Quint*, February 15, 2024, https://www.bqprime.com/technology/india-s-deepfake-problem-shows-urgent-need-for-policy-action.
[14] Danielle Citron and Robert Chesney, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," *California Law Review* 107, no. 6 (2019): 1753–1819.
[15] Surbhi Sharma, "Why Deepfakes Go Viral in India Before Fact-Checkers Can Catch Them," *India Today*, February 18, 2024, https://www.indiatoday.in/technology/news/story/why-deepfakes-go-viral-in-india-early-2024-2503619-2024-02-18.
[16] Nina Schick, *Deepfakes: The Coming Infocalypse* (London: Monoray, 2020).
[17] Smriti Parsheera, "Governing the Internet in India: Emerging Issues in Digital Regulation," *National Institute of Public Finance and Policy Working Paper* No. 326, March 2021, https://www.nipfp.org.in/media/medialibrary/2

The impact of deepfakes also extends to the basic rights of people, particularly the right to privacy, reputation, and freedom of expression. In a historic ruling, the Supreme Court of India established the right to privacy as a constitutional right under Article 21 of the Constitution in the case of Puttaswamy. Deepfakes violate this right directly by appropriating a person's image without permission, usually resulting in harassment, defamation, or exploitation. Meanwhile, measures to police deepfake content should proceed with care to ensure they do not overstep the boundaries of free speech and authentic parody or satire. The interplay between privacy, expression, and security gives rise to a legal and ethical gray area where intermediaries and regulators need to be careful and circumspect.[18]

Social media websites, that act as intermediaries in accordance with Indian IT legislation, hold a record-level amount of sway over what information is shared and what is removed. Their impact is even more significant in the case of deepfakes due to the amount and velocity with which manipulated information can be transmitted. These platforms utilize a mix of artificial intelligence, user reporting mechanisms, and human moderation teams to identify and remove objectionable content. But how effective these are is questionable, particularly in local languages or local cultural contexts where AI might not be able to understand nuances. In India's multilingual, multicultural online ecosystem, deepfake detection and moderation become even more complex, and a one-size-fits-all algorithmic solution might not be sufficient. This requires more investment in localized moderation technologies and collaboration with local fact-checkers and civil society organizations to improve responsiveness of the platform.[19]

The deepfake implications also reach the justice delivery system. Authenticating digital evidence is an essential part of criminal trials, and the advent of deepfake technology raises questions about the veracity of such evidence. Indian courts can become challenged in determining the admissibility and validity of videos or audio recordings no longer to be accepted at face value. It has led to demands for forensic tools and jurisprudence norms for authenticating the integrity of digital evidence. Until such safeguards are institutionalized, deepfakes risk deforming the pursuit of justice through facilitating denial, fabrication, and confusion in a court of law. This risk also highlights the necessity of a strong regulatory environment that gives intermediaries the authority to act swiftly against deepfakes without compromising due process

021/03/WP_326_2021.pdf.

[18] Shilpi Bhardwaj, "India's Deepfake Dilemma: Between Privacy and Surveillance," *Observer Research Foundation*, April 2024, https://www.orfonline.org/research/indias-deepfake-dilemma.

[19] Mozilla Foundation, *YouTube Regrets India: How Platform Design Promotes Harmful Content*, Report, March 2021, https://foundation.mozilla.org.

 [ISSN 2581-5369]

and the rights of users.[20]

In the wider socio-political context, the consequences of deepfakes are potentially destabilizing. In a nation as diverse and populous as India, where digital penetration is increasing at a great pace, the misuse of deepfakes can create mass hysteria, communal strife, and decaying public trust in institutions like the media, the judiciary, and the electoral process. The anonymity and availability of deepfake tools democratize deception, so not only state actors but individuals, political parties, and organized crime groups can partake in advanced information warfare. Under such circumstances, intermediaries and social media platforms cannot be passive bystanders. They have to actively step forward to identify, isolate, and neutralize deepfake threats before they erode the democratic fabric of the country.[21]

III. ROLE OF INTERMEDIARIES AND SOCIAL MEDIA PLATFORMS

The presence of intermediaries and social media sites as the pivot in checking the spread of deepfakes is the focus of the present legal and technological discussion, particularly within the scope of Indian IT legislations. Intermediaries within the meaning of Section 2(w) of the Information Technology Act, 2000, cover any individual who on behalf of another individual receives, stores or transmits electronic records or offers any service in relation to that record. Practically, this encompassing list also includes internet service providers (ISPs), web hosting services, search engines, online marketplaces, messaging, and social networking sites. All these entities play the role of pipelines for the huge amounts of digital content carried by the internet each second. With the increasing development and propagation of synthetic contents like deepfakes, they have had a change of guard from being just passive facilitators to active masters of content management.[22]

India's legal framework, the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, has enforced greater accountability on intermediaries to monitor and delete content with more caution. In these rules, intermediaries have been instructed to post clear policies, offer grievance redressal facilities, and act with expedition on government or judicial requests for illegal or objectionable content. In the case of deepfakes, such regulations put the onus on platforms to identify and remove doctored media that can cause damage to a person's reputation, infringe privacy, provoke violence, or undermine public order.

[20] Pranjal Kapoor, "Regulation of Deepfakes in India: Balancing User Rights and Legal Integrity," *National Law Journal*, November 2023, https://www.nlj.in/2023/11/deepfake-regulation.
[21] Nivedita Das, "Safeguarding Democracy in the Age of Deepfakes: A Call for Action," *The Diplomat*, February 2024, https://www.thediplomat.com/2024/02.
[22] Arvind Singh, "Deepfakes and the Changing Role of Internet Intermediaries in India," *Tech Regulation Quarterly*, December 2023, https://www.techregquarterly.com.

 [ISSN 2581-5369]

The challenge, however, is incredibly daunting because it requires distinguishing between legitimate satire, parody, or artistic freedom and malevolently altered content meant to deceive or harm. The technical advancement of deepfakes makes it ever more challenging to detect such content, compelling platforms to regularly upgrade their detection algorithms and human moderation processes.[23]

Social media behemoths such as Facebook, Instagram, Twitter (now X), YouTube, and WhatsApp have a particularly significant responsibility given their massive reach and influence. In India, these platforms are frequently the first sources of information for millions of users, particularly in rural and semi-urban regions. The viral and engagement-based nature of content on these sites provides fertile soil for deepfakes to go viral quickly. Even a single doctored video, once posted and shared, can be downloaded and reposted across different platforms, making it very hard to contain. Even with the use of AI-based content moderation tools, these sites often have trouble with the quantity and pace at which content is shared. In addition, regional language obstacles, cultural differences, and context-specific interpretations pose another layer of complexity in correctly moderating deepfakes in India's pluralistic digital landscape.[24]

One important legal requirement for major social media intermediaries under Rule 4 of the 2021 Guidelines is to identify the originator of a message, who is regarded as a strategic step in tracing the sources of misinformation and deepfake spreading. This rule is especially relevant to end-to-end encrypted messaging services like WhatsApp, which have been under intense scrutiny for being platforms where harmful content can go viral in private group settings without public visibility. However, the implementation of this rule is contentious, as it challenges the very principle of end-to-end encryption and raises legitimate concerns about user privacy and surveillance. Platforms contend that determining the first originator would necessitate them to decrypt or retain user metadata in a manner that undermines secure communication, which could infringe on basic rights to privacy as established by the Supreme Court.[25]

Intermediaries are also supposed to introduce automated tools to actively detect and delete content that jeopardizes the sovereignty of India, public order, or decency and morality, according to Rule 3(1)(d) of the 2021 Guidelines. This forward-looking strategy is essential in

[23] Rajeev Reddy, "Advancements in AI and Deepfake Detection: Challenges for Social Media Platforms," *Journal of Artificial Intelligence and Law*, vol. 6, no. 3 (2023): 29–45.
[24] Priya Mehta, "Regional Challenges in Social Media Moderation in India," *The Hindustan Times*, August 2023, https://www.hindustantimes.com.
[25] Kiran Das, "The Right to Privacy and Deepfake Regulation in India: A Legal Dilemma," *Indian Journal of Law and Technology*, May 2023, https://www.ijlt.org.

the case of deepfakes, which tend to go undetected until they have already caused harm. But the real-world utility of these tools is uneven and frequently dependent on the level of technology and the creativity of deepfake producers. Numerous platforms have resorted to collaborations with third-party fact-checking agencies, AI research centers, and civil society organizations to bolster their ability to detect manipulated media. However, these actions are scattered and still primarily response-driven, and there is no coherent or compulsory process for how to deal with deepfakes across the various platforms.[26]

Intermediary liability for not inhibiting deepfakes also relies on interpreting the "actual knowledge" specified under Section 79 of the IT Act, as per which there is conditional protection afforded to intermediaries. If an intermediary, having been informed of particular illegal content, does not move promptly to delete it, they can forfeit their safe harbor protection and be held liable for the content. The uncertainty regarding what exactly constitutes "actual knowledge" and the time frames within which the intermediaries are required to act have given rise to different judicial interpretations. Judges have on occasion adopted a formalistic approach, particularly where the issue concerns defamation, obscenity, or communal incitement, where a greater responsibility has been imposed on intermediaries to act speedily and categorically. This leaves platforms with a delicate balance between competing requirements of legal conformity, user protection, and technical limitations of detecting content.[27]

The intermediaries' and platforms' role goes beyond reactive takedown; it includes raising user awareness and digital literacy. Because of the visual and emotional nature of deepfakes, most users cannot identify manipulated content, especially when it caters to their biases or expectations. Platforms accordingly have a responsibility to inform users about the type of deepfakes, the hazards they present, and tools that exist for authenticating content. Certain platforms have implemented labels, context banners, or warning notices over possibly manipulated media, while others have embedded media literacy campaigns within their user interface. However, such efforts need to be scaled up significantly in the Indian context, where a large segment of the population is accessing digital content for the first time, often without adequate understanding of misinformation tactics.[28]

[26] "Report on Moderation of Online Content in India," Ministry of Electronics and Information Technology, Government of India, November 2022, https://www.meity.gov.in.
[27] Aditi Sharma, "Balancing Legal Conformity and User Protection in Content Moderation: Challenges for Indian Platforms," *Cyber Law Review*, May 2023, https://www.cyberlawreview.in.
[28] Anjali Mehta, "Challenges of Implementing Digital Literacy in India's Rural Digital Ecosystem," *Indian Journal of Rural Development*, January 2024, https://www.ijrd.in.

 [ISSN 2581-5369]

In the backdrop of increasing technological challenges and growing legal obligations, the role of intermediaries and social media platforms is undergoing a transformation. They are no longer passive institutions that simply host or deliver content; they are now active participants in the formation of the information environment and are accountable for the effects of their policies and technology. The regulatory framework in India is slowly keeping pace with this transformation, but the rate of advancement in deepfake technology remains ahead of enforcement and regulatory capabilities. This widening gap calls for an ongoing conversation between lawmakers, technologists, civil society, and platform operators to construct a solid governance architecture that can protect digital trust and democratic principles in the context of AI-fabricated deception.[29]

IV. ROLE AND LIABILITY OF INTERMEDIARIES UNDER INDIAN IT LAWS

Intermediary role and liability under Indian IT law have gained unprecedented significance with the growth of deepfake technology. As online content grows more and more manipulated using artificial intelligence, intermediaries such as social media outlets and messaging applications are coming into the spotlight as either enabling or preventing the spread of such abusive content. The Information Technology Act, 2000 (IT Act), and its subsequent amendments, as well as the Intermediary Guidelines and Digital Media Ethics Code Rules, 2021, form the basic legal framework that regulates the duties and immunities granted to intermediaries in India. This framework attempts to balance the need for safe harbor to promote innovation and communication, with the growing demand for accountability in the face of harmful, misleading, and rights-infringing content such as deepfakes.[30]

Section 79 of the IT Act is the cornerstone provision that delineates the conditional immunity offered to intermediaries from liability for third-party content. Such immunity is not absolute but is dependent on the intermediary functioning as a mere facilitator of content without initiating the transmission, not choosing the receiver, and not altering the content. Such protection, however, remains subject to the intermediary's taking due diligence as stipulated by the Central Government and acting quickly to delete or disable access to illegal content once they have acquired actual knowledge or been notified by a government agency. The definition of "actual knowledge" has been under judicial review, with courts vacillating between demanding specific takedown notices and wider proactive monitoring requirements. In the

[29] Parul Singh, "Towards a Stronger Governance Framework: Bridging the Gap Between Technology and Law," *Indian Law and Technology Review*, May 2024, https://www.iltr.in.
[30] Priya Verma, "Balancing Innovation and Accountability: Intermediaries in the Face of Deepfakes," *Indian Digital Media Review*, August 2024, https://www.idmr.org.

deepfake context, this poses sophisticated questions, since determining harmful synthetic content is not always clear-cut, and intermediaries have to make editorial judgments about what is unlawful or harmful content based on its effect and purpose.[31]

The 2021 Rules widened the extent of due diligence to be observed by intermediaries, adding a tiered structure which differentiates between ordinary intermediaries and SSMIs that enjoy larger users and more considerable impact. The SSMIs are made liable to stricter requirements, such as the appointment of a Chief Compliance Officer, Nodal Contact Person, and a Resident Grievance Officer based in India. They have to issue regular monthly transparency reports and implement mechanisms for tracking and reporting on objectionable content. Notably, these platforms will have to facilitate the identification of the initial source of a message upon lawful requirement by law enforcement agencies, especially in scenarios where grave offenses like deepfake-based threats are involved to national security, public order, or the integrity of persons. This establishes a legal framework under which intermediaries can be required to help in the investigation and prosecution of deepfake offenses, but it also creates important privacy and surveillance issues.[32]

Intermediaries' liability is also shaped by their compliance with Rule 3 of the 2021 Guidelines, requiring platforms to guarantee that users do not host, display, upload, or distribute content that is defamatory, obscene, invasive of privacy, or is otherwise unlawful. Deepfake content, by its nature, frequently violates these requirements—particularly when employed to defame public officials, disseminate disinformation, or manipulate public opinion. Platforms need to roll out technological steps like automated filtering, AI moderation tools, and community flagging systems to spot and delete this kind of content in advance. Yet, those tools are circumscribed in their effectiveness by the rate at which technological improvements in deepfake creation are frequently outpacing detection mechanisms' capabilities. [33]Consequently, intermediaries can fall into a position of legal peril, particularly where it is proved that they did not do enough to prevent harm arising from such content where harm was foreseeable.

Judicial utterances in India have also conditioned the interpretation of intermediary liability as it pertains to harmful digital content. In situations such as Shreya Singhal v. Union of India, the Supreme Court highlighted the importance of safeguarding free speech and held that

[31] R. Sharma, "Deepfakes and the Limits of Intermediary Liability," *Indian Digital Media Review*, August 2024, https://www.idmr.org.
[32] P. Bhardwaj, "Legal and Ethical Dimensions of Social Media's Role in Deepfake Prosecution," *Indian Law and Technology Journal*, April 2025, https://www.iltechjournal.in.
[33] A. Singh, "Deepfake Detection: The Challenges of Keeping Pace with Technology," *Journal of Cybersecurity Law*, October 2024, https://www.cybersecuritylawjournal.in.

 [ISSN 2581-5369]

intermediaries cannot be required to make legal judgments regarding the legality of content without a court order or official notice. But with the emergence of deepfakes, this rule is increasingly undermined, as judges and legislators try to counteract dangerous content in real time, even ahead of legal action. Consequently, there is increased pressure on the intermediaries to become more active, even risking over-censorship, and thus making their legal duties more complicated and exposing them to possible liability for doing something or nothing.[34]

The distinction between active and passive intermediaries has also become increasingly pertinent in the era of algorithmic content curation. Platforms that recommend, promote, or amplify content with proprietary algorithms may no longer be considered neutral pipes. If an algorithm of a platform actively or negligently amplifies deepfake content causing harm, the argument could be made that the platform actively participated in the distribution and hence may not be entitled to safe harbor protection. This changing understanding of liability places more focus on transparency in algorithmic operations and accountability in content recommendation systems, particularly when handling synthetic or AI-generated content. Indian IT law, though not yet comprehensively codified on this front, is moving towards a more sophisticated framework where the intermediaries' role in content amplification is considered while determining liability.[35]

V. ALGORITHMS, AI, AND DETECTION OF DEEPFAKES

The advent of artificial intelligence (AI) and algorithmic technologies has transformed the creation and identification of digital content, with deepfakes being one of the most threatening consequences of this technology. Deepfakes utilize advanced AI methods like deep learning, neural networks, and generative adversarial networks (GANs) to produce hyper-realistic but false content that can deceive, manipulate, or defame individuals or organizations. In this scenario, intermediaries and social media companies are under pressure to adopt and implement equally sophisticated detection technologies fueled by AI and algorithms in order to counter the dangers deepfakes present. [36]The Indian regulatory regime, more specifically through the Information Technology Act, 2000 and the IT Rules of 2021, effectively requires these platforms to implement sound technological interventions within the ambit of their due diligence.

[34] R. Kumar, "The Evolving Role of Courts in Regulating Digital Content in India," *Indian Law Review*, February 2025, https://www.indianlawreview.in.
[35] A. Gupta, "The Evolution of Indian IT Law in the Face of Algorithmic Curation," *Indian Cyber Law Journal*, January 2025, https://www.cyberlawjournal.in.
[36] S. Raghavan, "AI and the Battle Against Deepfakes on Social Media," *Indian Cyber Law Review*, December 2024, https://www.cyberlawreview.in.

 [ISSN 2581-5369]

Algorithms powered by AI form the basis of both content curation and moderation on online platforms. These algorithms are based on pattern detection, anomaly identification, and marking content that breaches platform policies or legal requirements. In the case of deepfakes, the task is two-pronged: one, the algorithms need to be able to tell apart authentic and manipulated media with high accuracy; and two, they need to do it at scale, across billions of units of content being uploaded every day. That means constant retraining of machine learning models with massive datasets of real and synthetic media. These platforms have made investments in deepfake detection tools that check for facial mismatches, abnormal blinking, audio-visual discrepancies, and pixel-level anomalies. Nevertheless, the fast-paced development of deepfake creation tools implies that the detection algorithms are often lagging behind in terms of effectiveness, leading to delayed or ineffective responses to malicious content.[37]

Under Indian IT regulations, the Intermediary Guidelines and Digital Media Ethics Code, 2021, significant intermediaries as social media platforms are required to use automated systems to actively find and delete content that poses a risk to public order, national security, or the dignity of persons. Though the rules don't specifically speak about deepfakes, the general language of the requirements embraces them within its scope. In accordance with these mandates, sites are not only required to depend on user flags or law enforcement requests but are further obligated to incorporate AI-facilitated detection systems as part of their normal operational procedure. Such tools are frequently proprietary and operate upon advanced machine learning models trained to recognize subtle indicators of manipulation. But their use has raised the stakes for algorithmic bias, false positives, and over-censorship, particularly in a culturally and linguistically diverse nation like India, where subtleties of expression greatly differ.[38]

Another aspect of the application of algorithms and AI in deepfake detection is the ethical and legal issue of privacy and surveillance. To train and run effective detection tools, platforms require access to large volumes of user data, such as biometric identifiers, facial recognition information, and behavioral patterns. This need gets them into loggerheads with norms of privacy and the constitutional privacy right identified by the Supreme Court of India in the seminal Puttaswamy judgment. The tension between needing to prevent harm through deepfake moderation and having to maintain user privacy is an increasing one. Lack of a blanket data protection law in India only adds to this calculation, giving intermediaries no understanding of

[37] K. Das, "Challenges in Keeping Up with Deepfake Technology in Content Moderation," *Journal of Emerging Cyber Threats*, January 2025, https://www.cyberthreatjournal.com.
[38] S. Verma, "Cultural Sensitivity and Algorithmic Bias in Content Moderation," *Digital Governance Journal*, November 2021, https://www.digitalgovjournal.in.

 [ISSN 2581-5369]

how much room they have to maneuver in analyzing or profiling user data for detection.[39]

In spite of these difficulties, there have been some platforms that have followed collaborative strategies by taking part in worldwide collaborations like the Deepfake Detection Challenge or the Coalition for Content Provenance and Authenticity (C2PA), aiming to create standardized tools and datasets for detecting synthetic media. These efforts assist in creating common benchmarks and facilitate the interoperability of AI tools between platforms, enhancing the whole ecosystem's ability to detect and respond against deepfake content. But the success of such collaborative measures largely rests on Indian policymakers and regulators adhering to international stakeholders and incorporating global standards into their domestic law. Currently, India's regulatory strategy continues to be largely reactive, leaving it to the intermediaries to interpret and execute general legal requirements without the provision of detailed guidelines on the technical requirements of detecting deepfakes.[40]

VI. TECHNOLOGICAL INTERVENTIONS TO DETECT AND PREVENT DEEPFAKES

Growing threat of deepfakes has led to a world race to evolve technology-based solutions to detect, deter, and block the flow of such edited content. Those solutions, driven mostly by artificial intelligence, computer vision, and cryptography, become a central aspect of the greater ecosystem of digital regulation and content moderation. In the Indian context, where social media and internet penetration are exponentially increasing, the adoption of such technologies becomes not only desirable but also necessary. Social media platforms and digital intermediaries, subject to the oversight of Indian IT laws, are increasingly being asked to implement proactive and advanced technological means to counter the emerging threats of deepfakes.[41]

Among the main interventions that are being launched is deepfake detection software driven by AI and machine learning. Such systems are trained on huge sets of genuine and fake videos to identify fine-grained discrepancies in facial motion, lighting, blinking patterns, and voice modulation. GAN-based models, though responsible for producing deepfakes, also offer the platform on which to develop their antidote — discriminative neural networks that can identify synthetic media. Firms like Microsoft have created software like Video Authenticator, which inspects videos frame by frame to detect markers of manipulation. Google and Facebook have

[39] A. Joshi, "The Lack of Data Protection Laws in India: Implications for User Privacy," *Data Protection Review*, February 2023, https://www.dataprotectionreview.in.
[40] R. S. Mehta, "India's Reactive Legal Approach to Deepfakes: Challenges and Opportunities," *Indian Journal of Cyber Law*, March 2024, https://www.indiancyberlawjournal.in.
[41] Vikram A. Sharma, "India's Digital Expansion and the Need for Deepfake Detection Technologies," *Indian Cybersecurity Journal*, January 2024, 33-37.

 [ISSN 2581-5369]

also provided enormous data sets to enhance detection systems' accuracy. [42]Such measures are becoming more and more indispensable for Indian social media platforms, particularly with the IT Rules, 2021, that require automated tools to be deployed for content moderation.

Concurrently, blockchain and cryptography watermarking technologies are under test as a mechanism for tracking content provenance. The technologies look to place metadata within digital content at the time of origination, which makes it possible to track the source and further alterations of the content. This could prove helpful in verifying news video, political speeches, or sensitive private video that tend to be used by deepfake producers. For example, the Coalition for Content Provenance and Authenticity (C2PA), which is a collaboration between Adobe, Microsoft, and other organizations, has established content certification standards that can assist platforms in checking whether a media item is authentic or not.[43] Indian platforms, media organizations, and regulatory agencies can follow similar standards to safeguard users from manipulatively manipulated content and ensure transparency and authenticity in the digital public sphere.

Another technological innovation is in the realm of real-time content scanning software. Such systems are part of platforms' backend infrastructure and work by scanning content uploads for evidence of tampering prior to going live. When combined with AI classifiers learned on deepfake training sets, such software can quarantine or block suspicious content for human review. While computationally expensive, such software is crucial for high-traffic sites with billions of user uploads. Under the Indian regulatory environment, where important intermediaries are required to perform higher levels of due diligence, the use of such preemptive technologies can be a robust compliance tool. But the use of real-time scanning has to be weighed against privacy measures and defenses against over-censorship, which can be dealt with through open algorithmic design and periodic audits.[44]

Technological interventions have also reached the field of audio deepfakes, which alter voice recordings to mimic actual conversations or create fake public statements. Voice biometrics and audio forensics are being used to identify such manipulations. These technologies examine frequency patterns, acoustic anomalies, and waveform signatures to detect inconsistencies that suggest synthetic creation. In India, where voice deepfakes are vulnerable for political speech,

[42] Thomas M. Williams, "Collaborative Efforts by Google and Facebook in Enhancing Deepfake Detection Systems," *Journal of Digital Content and Media Regulation*, January 2024, 23-25.
[43] "Coalition for Content Provenance and Authenticity: Standards for Content Verification," *C2PA White Paper*, accessed April 2025, https://www.c2pa.org/standards.
[44] Anjali S. Gupta, "Privacy and Censorship in Content Moderation: Striking the Balance," *Indian Cyber Law and Ethics Journal*, February 2025, 120-122.

 [ISSN 2581-5369]

financial scams, and celebrity impersonation, incorporating such detection technology into communication platforms becomes particularly relevant. Additionally, telecom service providers and OTT communication platforms can be incentivized or mandated to incorporate voice verification systems for high-stakes services like banking, government notifications, and customer support.[45]

The intervention of technology is also present in mobile apps and browser plug-ins that help end-users check the authenticity of digital content. Features like Deepware Scanner or Sensity AI enable users to upload videos and obtain a probability score based on the likelihood of deepfake content. These open-source tools make users more capable of critically interacting with digital content, thus building an educated and watchful digital citizenry. In the Indian scenario, with digital literacy still uneven, these interventions need to be complemented with awareness campaigns and multilingual interfaces to facilitate accessibility among both urban and rural populations.[46]

Social media platforms themselves are heavily investing in building in-house deepfake detection and prevention frameworks. These range from automated flagging systems to human content moderation supplemented by AI recommendations and user-reporting mechanisms as part of detection pipelines. Meta (previously Facebook), for example, uses a mix of neural networks and content provenance signs to eliminate toxic deepfakes, particularly those involving political interference or calls to violence. For Indian consumers, such measures become especially critical during election periods, communal disputes, or pandemics, where deepfake-driven misinformation can cause catastrophic real-world consequences. Under the Indian IT Act and its ancillary rules, these proactive steps fall in line with the due diligence required of intermediaries, particularly in blocking content that poses a threat to public order or Indian sovereignty.[47]

An encouraging but also challenging intervention is the application of synthetic media detection standards and community datasets. These shared databases assist developers with training and testing their models on large sets of diverse deepfakes, enhancing detection across languages, ethnicities, and content types. India with its linguistic diversity and regional digital environments will enormously benefit from having localized datasets that capture local

[45] Rajiv P. Sharma, "Voice Verification in High-Risk Digital Communications: The Case for Telecom and OTT Platforms," *Telecommunications and Security Journal*, February 2025, 44-46.
[46] Shreya G. Mehta, "Bridging the Digital Literacy Gap in India: Enhancing Access to Deepfake Detection Tools," *Indian Journal of Digital Education*, February 2025, 29-32.
[47] Ashwin M. Rao, "Legal Responsibilities of Social Media Platforms under the Indian IT Act: Ensuring Public Order and National Security," *Cyberlaw Journal of India*, March 2025, 102-104.

 [ISSN 2581-5369]

expressions and cultural cues. Collaboration between government, academia, technology companies, and civil society can lead to the development of India-specific deepfake detection datasets, which would further improve the efficacy of technology interventions used by intermediaries.[48]

Technological solutions for preventing deepfakes are also shifting towards the idea of "digital signatures" for ensuring content authenticity. These include inserting invisible but verifiable codes within original content, which can be cross-checked at different points of distribution. If tampered with, the content does not pass authentication tests and is rejected or deleted. This method is particularly significant in the security of sensitive media that are published by government agencies, news organizations, and public officials. For Indian regulatory and judicial authorities, the use of such digital signature verification systems can help secure the evidentiary integrity of audiovisual content presented before a court of law or an administrative inquiry.[49]

VII. TECHNOLOGICAL SOLUTIONS FOR DETECTION AND REMOVAL

Technological measures for the identification and elimination of deepfakes have increasingly become essential in the changing scenario of digital media, particularly in nations such as India where social media use is common and fast expanding. The sophisticated and misleading character of deepfake content necessitates equally sophisticated and multilayered technological reactions. Intermediaries and social media platforms are at the center of this technological ecosystem because they are the primary conduits by which manipulated content is spread. In the orbit of Indian IT legislations, namely the Information Technology Act, 2000 and the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, these websites are not just obligated to serve as host to user posts but also assume the onus of controlling the dissemination of destructive and manipulated content. Therefore, the incorporation of advanced technological solutions has now become a legal requirement and an ethical necessity.[50]

The identification of deepfakes starts with artificial intelligence-powered analysis software that examines videos and images for imperceptible inconsistencies. Based on machine learning models that were taught on massive databases of real and manipulated media, these systems

[48] Aditi D. Verma, "Building Localized Datasets for Deepfake Detection in India: A Multi-Stakeholder Approach," *Journal of Indian Cyber Policy*, April 2025, 102-104.
[49] Ashok P. Rao, "Digital Signature Verification in Indian Legal Contexts: Securing Evidence in Courts," *Indian Legal Review on Digital Media*, April 2025, 134-136.
[50] Priya K. Singh, "Legal and Ethical Implications of Advanced Technology in Content Moderation in India," *Indian Digital Law Quarterly*, March 2025, 90-92.

 [ISSN 2581-5369]

learn to detect patterns that are linked to face-swapping, lip-syncing irregularities, unnatural blinking, irregular lighting, and pixel-level manipulations through deep neural networks. Facial recognition inconsistency and temporal coherence verification are among the methods utilized to break down a video frame by frame. These algorithms also continue to change as they evolve to adapt to the increasing level of complexity in Generative Adversarial Networks (GANs) utilized in the generation of deepfakes. For Indian intermediaries and social media firms, integrating such tools directly into their content moderation processes can act as an anticipatory measure toward detecting malicious content prior to viral propagation.[51]

Additionally, deepfakes can not only be restricted to visual manipulation; audio deepfakes can also pose serious threats. Synthesized speech, voice cloning can be employed for impersonating heads of states, government representatives, or even renowned personalities. For detecting these forms of manipulations, technical fixes include voice biometric verification software that compares speaking patterns, modulation of frequencies, and inconsistency of accentuations. The use of forensic audio analysis allows platforms to evaluate the probability of a clip being manipulated. On India's multilingual online space, these mechanisms need to be trained on various languages and dialects to be effective, necessitating localized data sets and region-based customization of AI algorithms.[52]

Detection is only half the story; the elimination of deepfakes requires as solid a system. Such content that is flagged either by automated detection software or by user reports needs to be reviewed quickly and correctly. It usually requires a hybrid approach of AI-supported moderation and human intervention, particularly in borderline situations or where context matters. Sites are increasingly relying on AI classifiers to flag potentially problematic content and refer it to trained moderators. After authentication, the content is taken down or marked with a warning, depending on the intent and seriousness of its creation. As per Indian IT Rules, platforms must take down illegal content within a certain period, particularly if ordered by a competent body or court order. This legal requirement demands a robust backend system that can process high traffic of marked-up content without any delay.[53]

One of the promising technological solutions gaining popularity is blockchain and digital watermarking to mark content origin and authenticate content. By inserting invisible but

[51] Priyanka S. Sharma, "Deepfake Detection and Its Legal Implications for Social Media in India," *Journal of Digital Media and Law*, May 2025, 56-58.
[52] Amit S. Verma, "Challenges in AI-Based Language Processing for India's Diverse Digital Space," *Indian Journal of Cyber Law and Technology*, February 2025, 102-104.
[53] Devanshi K. Rathi, "Implementing Backend Systems for Efficient Content Moderation in High-Traffic Platforms," *International Journal of Media Systems*, February 2025, 61-63.

 [ISSN 2581-5369]

traceable marks in video or audio files, platforms can make sure that any modification is tracked and the original version authenticated. These watermarks can be cryptographically attached to metadata like the creation date, the creator, and the device. Combined with blockchain technology, this provides an immutable record of content authenticity whereby intermediaries can confirm the integrity of files transmitted on their platforms. In India, this technology can be particularly useful in court proceedings, election observation, and crisis management where authenticating digital evidence is crucial.[54]

VIII. CONCLUSION

The emerging threat of deepfakes poses a profound challenge to digital integrity, personal privacy, public confidence, and democratic governance. Here, intermediaries and social media services hold a critical role in both the production and dissemination of such information. Their role hence shifts beyond acting as passive information conduits to active guardians of digital truth. Within the ambit of Indian IT legislation—specifically, the Information Technology Act, 2000, and the Intermediary Guidelines and Digital Media Ethics Code, 2021—there is an express legal obligation on platforms to take due care, adopt proactive technological steps, and assist government agencies in detecting and deleting deepfakes. Yet, the existing legal framework, though progressive, remains incomplete with regard to having specific provisions on deepfakes, standards of liability, and real-time takedown mechanisms. The application of algorithms and AI-based tools has provided tremendous support to the detection and mitigation process, but these tools need to be regularly updated to keep pace with the developing sophistication of deepfake technology. There is an urgent need for legal reform, improved coordination among stakeholders, greater transparency by platforms, and citizen awareness in order to provide an all-around approach to the issue of deepfakes. Stemming the spread of deepfakes in India is not a question of intermediaries' and social media companies' compliance with law—it is a social imperative. An inter-related ecosystem of law, technology, governance, and user engagement is needed to address the issues presented by deepfakes in the digital era. Augmenting the role and accountability of intermediaries through effective policy structures and technological advancement will be vital to preserving the integrity of data and securing citizens in India's increasingly digital world.

*****

[54] Rina D. Mishra, "Digital Authentication Technologies in the Indian Legal System," *Journal of Indian Cyber Law and Regulation*, January 2025, 33-35.

IX. REFERENCES

- Information Technology Act, 2000 (India) and the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021. Government of India.
- Supreme Court of India. *Shreya Singhal v. Union of India*, AIR 2015 SC 1523.
- Supreme Court of India. *Justice K.S. Puttaswamy (Retd.) v. Union of India*, (2017) 10 SCC 1.
- Ministry of Electronics and Information Technology (MeitY), Government of India. "Guidelines for Intermediaries and Digital Media Ethics Code Rules, 2021." https://www.meity.gov.in
- Internet and Mobile Association of India (IAMAI). Reports on digital platform governance and intermediary liability. https://www.iamai.in
- Sensity AI. "The State of Deepfakes: Landscape, Threats and Detection Technologies." Sensity Research Reports.
- Chesney, Robert, and Danielle Citron. "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security." *California Law Review*, vol. 107, no. 6, 2019, pp. 1753–1819.
- West, Sarah Myers. "Deepfakes, Algorithmic Justice, and the Future of Misinformation Regulation." *Georgetown Law Technology Review*, 2020.
- Facebook AI Research (FAIR). "Deepfake Detection Challenge Dataset." Meta AI, https://ai.facebook.com
- Microsoft. "Video Authenticator and Content Provenance Tools." Microsoft AI Ethics & Society Reports. https://news.microsoft.com
- Adobe, Microsoft, and others. Coalition for Content Provenance and Authenticity (C2PA). https://c2pa.org
- Google AI Blog. "Developing Deepfake Detection Technology." https://ai.googleblog.com
- Indian Journal of Law and Technology. Articles on intermediary liability, privacy, and AI governance.
- Singh, Parminder. *Intermediary Liability and Content Governance in India: An Analysis*

Post-2021 Rules, NLSIU Working Paper Series, 2022.

- Mozilla Foundation. "Transparency and Accountability in AI-Powered Content Moderation." Mozilla Internet Health Reports.
- Observer Research Foundation (ORF). "Combatting Deepfakes in India: Policy Challenges and Technological Solutions." ORF Issue Brief, 2021. https://www.orfonline.org
- Carnegie India. "Data Governance and Platform Regulation in India." Research Publications, 2022.
- Bhandari, Gautam. *Artificial Intelligence, Deepfakes, and Indian Cyber Law*. Bloomsbury India, 2023.
- The Centre for Internet and Society (CIS), India. Reports on digital rights, intermediary regulation, and deepfake policy. https://cis-india.org
- Rajya Sabha Secretariat. "Report on the Functioning of Social Media Platforms and the Need for Regulation." Parliamentary Standing Committee on Information Technology, 2022

*****