VOLUME 6 | ISSUE 3

INTERNATIONAL JOURNAL OF ADVANCED LEGAL RESEARCH

WHEN ALGORITHMS UNDRESS REALITY: DEEPFAKES AND THE CASE FOR DEDICATED AI LIABILITY LAW IN INDIA

- Navni Krishna & Ahalya Reghunath*

Abstract

With the recent rise in generative AI solutions to create hyper-realistic synthetic media or "deepfakes" have been associated with a set of legal and regulatory challenges due to their potential for misuse in impersonating someone else's identity (creating fake news), sexual exploitation (creating child pornography), fraud (scamming people) and misinformation (political or social manipulation). The most recent controversy of the alleged use of Grok AI is an illustration of the potential to aid in amplifying the spread of harmful content by creating this type of content on a large scale. Furthermore, these events have raised questions regarding accountability in a distributed technology ecosystem.

This article discusses the limitations of the existing Indian law mechanisms to provide remedies for those who have suffered harm as a result of deepfakes, and to provide explanations for the harmful effects of deepfakes. Although there are many provisions in various pieces of legislation in India including Article 21 of the Constitution of India, the Information Technology Act, 2000, the Intermediary Guidelines, the Digital Personal Data Protection Act, the Bharatiya Nyaya Sanhita and the Copyright Act. However, these pieces of legislation offer fragmented relief, primarily in products (i.e., a reactive response to providing relief) and content (only related to the creation and distribution of harmful content) only. Deepfakes, due to their unique and dynamic qualities as compared to traditional forms of harm are composed of autonomous systems, scalable creations and probabilistic distributions making them not fit into the conventional scheme to determine and decide (or who would be liable for) who is responsible for creating the problem in the first place (i.e., who are liable in the chain of responsibility among developers, implementers/platform owners and consumers).

This article suggests that fault-based models that depend on human intention should not be used to evaluate distributed AI systems as they are not adequate. Instead, it proposes a tiered framework of Shared and Tiered Liability, incorporating a duty of care created by statute, duties of transparency and a risk-based classification of AI systems. Therefore, this paper calls for the creation of a stand-

*Students at SLS, CUSAT

alone AI liability regime in India that will utilise a shift from the punitive approach of post-harm liability to a more effective regime of preventive governance to ensure accountability and promote responsible innovation.

Keywords

Deepfakes; Artificial Intelligence, Legal Liability; AI Governance; Intermediate Liability; Digital Regulation; Synthetic Media; Shared Liability Framework

INTRODUCTION

AI-generated artificial media (e.g., video, audio, photo), called deepfakes, have made it possible for people to create new forms of entertainment and communication that look so real that they cannot be distinguished from actual videos and recordings. Because they are created using generative AI technology which blurs the line between reality and fake, they can be made quickly.

Grok AI has become the focus of controversy and has highlighted the risk of using generative AI technology to create misleading and manipulated outputs and how they can increase the scale of deception/harm by creating content that appears to be real. In addition, it raised significant questions about the responsibility of platforms to control misleading and manipulated content being used on their platforms, creating algorithmic safeguards to prevent synthetically created misleading and manipulated content, and how quickly because created synthetically created misleading and manipulated content can spread to a widespread audience prior to being stopped. More importantly, this highlighted uncertainty as to who, if anybody, is liable if synthetically generated misleading and manipulated content creates reputational harm, psychological harm, or social harm.

India lacks a comprehensive legal framework that specifically addresses the use of artificial intelligence to create misleading and manipulated media and provide clear liability standards. All existing laws address harmful content after the content has been disseminated and do not address the autonomous and scalable creation of synthetically created media. This article posits that the creation of deepfakes has revealed a structural gap within the traditional liability doctrine and asserts that India needs to implement a specific legal framework that will allow for the addressing of Artificial Intelligence-related harms in a coherent and responsible manner.

WHY DEEPFAKES BREAK TRADITIONAL LIABILITY RULES

The traditional concept of legal liability for crimes or civil wrongs relies on a clear understanding of who is culpable for criminal conduct; that is, the human actor who committed the action which

caused harm must be identified. Liability is usually based upon a finding that the defendant acted either intentionally, negligently, or in a way that could reasonably have been expected to cause harm to an identifiable person or group of people.

Deepfakes challenge the long-standing legal premise of identifiable human agency. Since AI-generated content can be created by the use of algorithms without requiring direct human instruction aimed at creating a specific victim, the responsibility of the actor who caused the harm to the victim cannot usually be established by proof of the actor's intent (mens rea) or established negligence on the part of the actor (failure to act as a reasonable person would have acted in similar circumstances).

Deepfakes have the added challenge of being produced at a level of volume unprecedented in history. Generative AI systems can generate hundreds or thousands of synthetic images, videos, or audio files in a matter of minutes. After being published through a digital platform, those files can be reproduced and disseminated indefinitely. Most tort and criminal laws were created with an expectation of harm resulting from discrete events; deepfakes can generate harm to multiple victims worldwide through an automated process of replicating and sharing.

Additionally, determining who is responsible for causing harm from a deepfake is complicated. A deepfake can be the product of multiple parties' actions: the entity that developed the algorithm, the entity that deployed the algorithm, the platform delivering the output of the algorithm, and the person who created the prompt to generate the deepfake. Identifying the causal relationship for liability within this layered causative relationship is difficult. The opacity of the algorithmic process also complicates the ability to trace responsibility back to the actor responsible for creating the deepfake.

THE GROK AI CONTROVERSY AS A LIABILITY STRESS TEST

Newly released Grok AI demonstrates generative systems exposing an archaic liability paradigm's shortcomings when confronted with artificial misinformation[1]. Users are reported to have requested the first AI-driven chatbot to generate sexually exploitive and non-consensual deepfakes, presenting a unique danger to the physical safety of women and children. Users prompted the chatbot to create images depicting the act of undressing others digitally or in sexually explicit poses, creating a new

[1]Danielle Keats Citron and Robert Chesney, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' (2019) 107 *California Law Review* 1753 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954 accessed 15 February 2026

type of digital sexual harassment[2] to target women in entertainment and influencers, but also everyday women.

For minors touched by synthetic sexual exploitation, digital sexual exploitation of children through synthetically created images may come under the definition of Child Sexual Abuse Material (CSAM)[3], creating compounded psychological trauma for children and creating opportunities for the secondary victimisation of children through cyberbullying and the online sharing of their images. The velocity and quantity of this type of image creation and distribution results in an unprecedented amount of harm before any regulatory or platform-based intervention.

Beyond the recognisable social and ethical implications, the controversy demonstrates an even bigger, systemic dilemma around how responsibility is deliberated in AI systems: the distributed nature of responsibility (i.e., multiple parties having a stake in a given outcome). Although a user provides text via a prompt to produce the result, the model architecture, content moderation, implementation, and platform host are the responsibility of the developers and providers. Consequently, traditional legal regimes have not adapted to the unique scenario presented when assessing liability for the end user, AI developer, platform host, or any combination thereof.

Thus, the Grok AI controversy can be seen as a demonstration not only of an isolated incident, but also as a regulatory stress test of legacy, intent-driven liability approaches and their ability to address harms associated with AI capability.

WHO IS LIABLE FOR DEEPFAKE HARM?

A key issue in deepfake regulation is accountability; specifically identifying who is liable for the harm caused by AI-generated content. While traditional forms of liability focus on identifying a clear human actor (e.g., someone whose actions, intent, or negligence can be evaluated), deepfakes disrupt this model through its distribution of agency amongst a number of actors in the technological ecosystem.

One area of possible liability is with the developer. AI companies develop and train generative models, create system architecture, and provide – or fail to provide safeguards (e.g., watermarking, content filters, or prompt restrictions) to prevent misuse of their systems. From this standpoint, developers are creating an inherently foreseeable risk of harm when providing a system with lack of safeguard(s). However, holding devs liable for all harmful uses of their systems discourages

[2]Mary Anne Franks, 'Criminalizing Revenge Porn' (2015) 49 *Wake Forest Law Review* 345 https://papers.ssrn.com accessed 15 February 2026.

[3]Europol, *Facing Reality? Law Enforcement and the Challenge of Deepfakes* (2022) https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes accessed 15 February 2026.

developers from innovating and fails to recognise the utility of many generative systems which serve a positive function in our society.

The focus then moves to platforms that host or implement these systems. Platforms control access to the systems, implement moderation policies and distribute content. Issues arise concerning the intermediary liability of platforms, and the due diligence obligations of platforms, when deep fake content is created or disseminated through platforms systems. Most current safe harbour laws will protect platforms from liability unless the platform has actual knowledge or fails to take appropriate action; therefore, a gap exists between the technological capabilities associated with these systems and the legal accountability of the platforms.

Lastly, user liability seems to be the most logical choice, at least in the case when an individual initiates the creation of harmful assigned content. User liability alone cannot cover the complexity of these accounts/processes. Anonymous accounts; worldwide access; and automated dissemination make enforcement of user liability burdensome. Moreover, a liability model based on fault assumes a party's reasonable intent or negligence can be identified, this is likely impossible when the harm caused by user input occurs through an algorithm-based process.

The complexities presented show just how deficient purely fault-based methods will be in addressing the problem created by deepfakes. Essentially, deepfake harm is usually scalable, automated, and spread across many different actors; therefore, causation and intent may be hard to identify independently. A more coherent response may involve creating a calibrated structure of shared or tiered liability combined with developer/platform due diligence obligations, coupled with direct liability for malicious users. Depending on the context in which the actions occur, some aspects of strict liability would likely also be reasonable to ensure victim compensation as well as specific deterrents.

Without some sort of recalibration, deepfakes will essentially continue to identify legal doctrine structural weaknesses, ultimately leaving victims with no clear path to resolution, and broken pathways of accountability throughout the entire AI value chain.

CURRENT INDIAN LEGAL FRAMEWORK

India, at present, deals with harmful digital content through scattered provisions across the Constitution of India, Information Technology Law, Criminal Statutes and other intermediary provisions but none of them specifically target deepfakes and suggest remdies for the aggrieved.

Article 21 of the Constitution of India

Right to privacy has been recognized as a fundamental right under Article 21 of the Indian Constitution, as affirmed by Justice K.S. Puttaswamy (Retd.) v. Union of India[4]. According to this judgment, an individual's right to privacy encompasses their ability to exercise control over disseminating personal information, which is both necessary and imperative. [5] However, constitutional remedies are invoked on a case by case nature and it doesn't impose accountability for AI systems that enable mass-scale violations. They only come into play after a violation or harm has happened.

Information Technology Act, 2000

The Information Technology Act 2000 addresses online harms through offence and intermediary provisions. The key provisions that deals with deepfake cases are:

Section 66C - identity theft

Section 66D ' cheating by personation using computer resources

Section 66E - violation of privacy through capturing/publishing private images

Section 67 & 67A - obscene and sexually explicit content

Section 69A -blocking powers

These provisions punish misuse after publication. They do not regulate AI model design, training risk, or its media generation capability. Liability attaches to the act of misuse, not to the architecture that enables misuse.

Intermediary Rules : 2026 Amendment on Synthetic Media (SGI Framework)

The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules 2021, as amended by G.S.R. 120€ dated 10 February 2026 (effective 20 February 2026), introduce a dedicated framework for "synthetically generated information" (SGI)—defined as AI-created or AI-altered audio, visual, or audio-visual content made to appear real or indistinguishable from natural persons or events (excluding good-faith edits, accessibility uses, and research).

Key obligations now include:

I. Platforms must deploy reasonable technical measures to prevent creation and spread of prohibited SGI, such as non-consensual intimate deepfakes, fraud impersonation, and deceptive event simulations.

II. Permissible SGI must carry clear labels, embedded provenance metadata, and unique traceability identifiers.

III. Rapid takedown timelines: about 3 hours for court/government-ordered unlawful SGI and 2–3 hours for sensitive categories like non-consensual deepfake imagery.

[4] Justice K.S. Puttaswamy (Retd.) v. Union of India (2017) 10 SCC 1

[5] Vig, Shinu. "Regulating Deepfakes: An Indian perspective." Journal of Strategic Security 17, no. 3 (2024) : 70-93.

IV. Mandatory user advisories every three months warning about legal consequences under cyber and criminal laws.

V. Compliance is tied to Section 79 safe-harbour protection, while non-compliance risks loss of immunity.

Despite this progress, the model is still platform-duty driven and reactive, not a full lifecycle AI liability regime.

Digital personal data protection bill

The DPDP Act mandates explicit, informed consent for processing digital personal data (including biometrics such as facial images and voice patterns), with fines up to ₹250 crore for breaches. It regulates how data is collected and handled, but not the unpredictable harms caused later by trained AI models, leaving a gap in accountability.

Bharatiya Nyaya Sanhita , 2023

The Bharatiya Nyaya Sanhita (BNS) does update criminal law, but it still tackles deepfake problems mainly as harmful content being circulated, not as technology-driven risk. Section 318 on cheating can be used when deepfakes are part of impersonation or fraud schemes, with serious jail terms, but it works only when clear deception and victim loss can be proved. Section 356 on defamation may cover reputational damage from fake videos or audio, yet it depends on showing publication and malice and it is often difficult when content is AI-generated and widely shared. Section 353 on public mischief can help where deepfakes disturb public order, such as during elections, but it offers little help for purely personal harms like privacy invasion or dignity loss.

Copyright Act, 1957

The Copyright Act, 1957 protects original works (s 13), exclusive rights (s 14), infringement (s 51), fair dealing (s 52), performers' rights (ss 38–38A), and moral rights (s 57). Deepfakes that reuse films, photos, or recordings may infringe these provisions. But fully synthetic deepfakes that copy identity without copying a protected work may fall outside copyright altogether. The statute protects works, not persona or likeness as such.

The current framework has some clear practical weaknesses. It focuses more on what happens after harmful content is shared than on how risky AI systems are built and released in the first place. Most liability rules still depend on proving intention and identifying a specific wrongdoer, which doesn't fit well with automated, probabilistic AI systems. Intermediaries often enjoy broad safe-harbour protection, while upstream AI developers have very few direct legal duties.

There are also real enforcement problems: black-box models, missing metadata, and traceability gaps make it hard to prove who created or altered a deepfake. Even though constitutional privacy under Article 21 calls for preventive safeguards, the present regime mostly reacts to harm once it is already public — leaving a continuing liability gap for deepfake damage.

WHY INDIA NEEDS A DEDICATED AI LIABILITY LAW

Sectoral tinkering like patching the IT Act, stretching the DPDP Act,expanding the scope of BNS offences and adding more intermediary rules cannot adequately address the borderless and rapidly evolving nature of the harms of generative AI such as the deepfakes. They simultaneously violate a person's right to privacy, dignity, reputation, personality rights, data protection, and freedom from defamation and harassment, while also infringing related copyright and performers' rights when protected works or performances are used without authorisation. Existing laws cannot keep up when harm arises from black box models at an exponential stage.

According to the Indian Cyber Crime Coordination Centre (I4C), Indians lost approximately ₹19,812.96 crore to fraud and cheating cases in 2025, with 21,77,524 such complaints registered on the National Cyber Crime Reporting Portal[6]. This forms part of a cumulative loss exceeding ₹52,976 crore to cyber frauds and cheating from 2020 to 2025. The government has stated that cybercrimes against women and children, including those involving impersonation and misinformation, are receiving special focus. Between 2021 and late 2025[7], the Citizen Financial Cyber Fraud Reporting and Management System (CFCFRMS) enabled the freezing/recovery of over ₹8,189 crore across more than 23.6 lakh complaints[8]. The Ministry of Home Affairs has informed Parliament that challenges persist in converting cyber fraud reports into FIRs and achieving convictions, due to difficulties in evidence preservation and investigation[9].

Therefore, a stand-alone AI Liability Act or a comprehensive chapter within a broader Digital India Act is necessary.

The Act should be enacted via a phased rollout :

Draft and Consultation Phase (First 6 Months)

The Ministry of Electronics and Information Technology should make a draft of the Artificial Intelligence Liability Bill. Talk to people from the industry non-profit groups and experts, in technology and law. This way the Artificial Intelligence Liability Bill framework will be useful, fair and based on the problems and uses of Artificial Intelligence.

[6]Ministry of Home Affairs, data from National Cyber Crime Reporting Portal (parliamentary response, January 2026).
[7]Ministry of Home Affairs, 'Cumulative Cyber Fraud Losses 2020–2025' (response to unstarred question, January 2026).
[8]Ministry of Home Affairs, 'Funds Saved through CFCFRMS 2021–2025' (Rajya Sabha response, February 2026).
[9] Ministry of Home Affairs, 'Challenges in Cyber Crime Investigation and Conviction Rates' (parliamentary reply, December 2025–February 2026).

Pilot Enforcement Phase (Next 12 Months)

When the law is put into effect the people in charge should start by trying it out in a group of areas that are more likely to have problems, like elections and financial services. This test period gives the regulators and companies a chance to see how well their systems for following the rules work and to make sure they are doing what they are supposed to do before everyone in the country has to follow the law. The regulators and companies will use this time to test their compliance systems their reporting duties and the safeguards that are, in place to protect the elections and financial services.

Full Implementation Phase (Within 18 Months)

The country should have a plan in place for everything and it needs to be supported by a special team and clear rules that everyone has to follow. This team should be in charge of making sure everything is done correctly. The rules should be final so that everyone knows what they are doing. The complete nationwide implementation of this plan should happen after that with the team and the clear rules, in place to help it along. Complete implementation of the plan should be the goal and it should be supported by the special team and the clear rules.

Institutional Oversight (AIRA)

We need to have an AI Regulatory Authority, which is the AI Regulatory Authority to make sure that companies are following the rules. The AI Regulatory Authority should also create standards, for the AI firms. Look into any problems that come up.

The AI Regulatory Authority can get the money it needs from the AI firms. These big AI firms can pay a part of what they make, like 0.5 percent to help the AI Regulatory Authority do its job without any interference.

Appeals and Accountability

Major regulatory orders should be appealable before designated judicial benches to guarantee due process and oversight.

The proposal should include the following core elements.

Risk-based Classification of AI

The Act should classify AI systems on the basis of its harm potential. The AI tools that come under the higher risk category should automatically trigger stricter legal duties and liabilities.

Statutory Duty of Care

Developers and deployers of the tools should be required to implement reasonable safeguards and the breach of this duty shall create civil liability.

Tiered and Shared Liability

The Act should distribute liability among the-

(i)Developers

(ii)Deployers/Platforms

(iii)Malicious users

It should be on the basis of control and contribution. This helps in reflecting the real causal chain on harms created by AIs

Transparency and Traceability

High risk AI systems should be statutorily required to add reliable watermarks and provenance tags to their outputs. Companies should also publish simple, anonymised summaries of their training data and risk safeguards, using compliance tools approved by the regulator, with watermarking first and full traceability rolled out within a year.

Centralized Victim Compensation Fund

Create a no-fault fund (operated by AIRA) with 1% levies on AI development platforms with >₹100 crore revenue & high-volume deployers. Provide fast-track compensation (in 30 days) for proven damages (₹1-5 lakhs for emotional distress, ₹10 lakhs for loss of reputation & extortion cases). The fund will provide up to ₹50 lakhs per victim annually.

These measures move AI liability from post-harm blame to early prevention, stopping damage before it spreads widely. Without a dedicated law, victims will face large-scale synthetic harms with little clear accountability. A focused AI liability statute would protect core rights and strengthen India's role in responsible AI governance.

CONCLUSION

Deepfakes are not only a technical advancement, they highlight a fundamental disconnect between the emergence of AI technology and the legal framework that was created prior to that technology. Because generative systems are able to generate vast volumes of hyper-realistic synthetic media, it presents difficulties to the foundational principles of liability that are generally tied to identifiable human intent to commit individual, distinct acts of wrongdoing. Currently emerging trends in scandals and their exploitation of AI technology demonstrate that harm is currently not solely defined as an outcome of an individual act, but rather as a result of distributing technology ecosystems that involved all stakeholders (including developers, deployer, platforms, and final users).

The Indian legal regime has constitutional protections for both privacy and dignity, and supplements those protections through a series of cyber laws, stability laws (like criminal law), and intermediary laws designed to regulate the way technology is used by individuals; however, these laws are

reactive and fragmented and only seek to address the potential for harm after it has occurred and not regulate the inherent risks associated with the design, deployment, and amplification of AI technologies. The lack of a clear framework for determining responsibility throughout the entire value chain of AI technology leads to victims navigating an environment of ambiguity and the likelihood of finding no or very limited recourse.

It is therefore essential to re-calibrate our understanding of liability. An AI liability regime which takes into account risk-based classification, statutory duties of care and transparency requirements, and calibrated shared liability will provide a means of moving from post-harm ascertainment of liability to preventative governance through regulation. This will provide legal certainty to good faith innovators and help protect individuals' privacy, reputation and dignity.

As digital fabrication becomes as real as the physical world, the legal system must adapt in order to prevent accountability from being diluted or delayed. A coherent AI liability policy will be critical not only to ensure rights are protected, but also to maintain public confidence and to facilitate technological development in India responsibly.