

Chapter - VII

DEFAMATION AND DEEPPAKES IN THE DIGITAL ERA: LEGAL ACCOUNTABILITY, ETHICAL ISSUES AND PUBLIC GENERAL AWARENESS

BY

Shanmugapriya.V & V. Srividhya

Abstract

The evolution of online communication has significantly changed the origins and distribution of digital information. The recent advancement in artificial intelligence plays a key role in generating text, images, and videos on digital platforms. The capability of creating highly realistic artificial intelligence-generated content that can be misused by users and result in serious damage to individuals, institutions and organizations. The distribution of deepfake or manipulated content has raised the chances of reputational damage, which requires digital literacy and awareness among the public. The existing legal frameworks primarily deal with the cases of human statements and provide limited guidelines on automated or manipulated content. In the era of artificial intelligence, the usage of chatbots such as large language models has led to many problems including defamation under the Indian Laws. The determination of misleading or malicious words became difficult because these chatbots have completely changed the connectivity between automated systems and the Internet. Even these automated systems can interact with the users which makes it more difficult to find the authenticity of the content. This chapter outlines the instructions and responsibilities for both developers and implementer. In this paper, the legal consideration of artificial intelligence-based defamation in India, technical harm with preventive measures, ethical implications, and developer responsibility are presented, to act as a guide to develop the artificial intelligence content. Legal responsibility and preservation of people's digital information will be achieved through these changes.

Keywords: Artificial Intelligence, Defamation, Deepfakes, Indian Law, Legal Accountability

Introduction

The Indian law system identifies protection of reputation as part of human dignity since reputation forms the image of a person in society¹. This social identity affects interpersonal relationships, professional status and individual agency. As a result, the damage to reputation can lead to social rejection, economic and psychological suffering. In Indian constitutional jurisprudence, injury to reputation is considered to be a grave violation of the law because it has a direct impact on the dignity of the individual. The Supreme Court has therefore valued reputation as an inherent part of a right to life under Article 21 of the Constitution of India². The ability of the person to live with dignity will thus be destroyed by any reputational damage. In order to protect this constitutional interest, the Indian law of defamation operates on a criminal and civil basis. Criminal defamation aims to discourage purposeful or willful criminal assault on reputation through punishing them³. Civil defamation is an extension of this aim, which grants damages and injunction to regulate reputational damages to victims⁴. Indian courts are also discovering the importance of freedom of speech as provided in Article 19(1)(a) even though reputation protection is being practised extensively. It creates a delineation of reputational harm and conflicts of interest.

The conventional law of defamation is formulated only for human defamatory contents which includes environments, authorship, intent and publication. For this reason, reputational harm is commonly defined as apology.

Defamation Law in India

This traditional defamation in India was reconstructed as the reputational protection law. Therefore, the organization is primarily structured around criminal law, civil remedies, information technology law, and judicial interpretation which deals with the reputational harm brought about by the traditional communication procedures⁵.

i) Defamation by the Indian Penal Code, 1860

In this context, the existing Sections 499 and 500 of the Indian Penal Code (IPC) 1860 regulated criminal defamation. Section 499 included a definition of defamation and exceptions of truth, public

¹ Yadav, S. C. *AN ANALYSIS of INDIAN CRIMINAL JUSTICE SYSTEM from the PERSPECTIVE of HUMAN RIGHTS* (Socio-Legal Perspective On Law And Changing Dimensions In The Modern Era, 2024)

² Antia D, 'Analysing The Impact Of Judicial Activism On Human Rights Protection: A Comparative Review Of Nigerian And Indian Jurisprudence' [2024] SSRN Electronic Journal

³ Dhakad, M., & Meena, R., 'A Comparative Study of Law Relating to Defamation in India and Commonwealth Countries.' (2024) 37(1) *Sciences of Conservation and Archaeology* 185-194

⁴ Ward, L., & Day, J. *Defamation and Social Media Law. In Research Handbook on Social Media and the Law* (Edward Elgar Publishing 2025) 8-37

⁵ Jaiswal A and Singh S, 'EVOLVING PARADIGMS of CRIMINAL LAW in CONTEMPORARY INDIA: CHALLENGES, REFORMS, and the QUEST for JUSTICE in a DIGITAL AGE' (2025) 3 *LawFoyer International Journal of Doctrinal Legal Research*. 839

good and fair comment in order to balance competing interests. Section 500 supported this equilibrium by directing punishment for intentionally or knowingly about reputational harm.

ii) Bharatiya Nyaya Sanhita Transition, 2023

This fault-based approach to defamation remained in the transition to the Bharatiya Nyaya Sanhita, 2023⁶. Section 356 is mostly a copy of the structure and language of the IPC provisions. As a result, the shift can be viewed as continuity and not actual reform in defamation law.

iii) Role of the Information Technology Act, 2000

Despite this, the rapid growth of digital communication did not directly change the definition of defamation with the Information Technology Act, 2000. Instead, it indirectly acts to counter the harm by controlling intermediary liability and content moderation. This led to the creation of defamatory content between the users, digital platforms, and the service providers. There is no such Indian law that covers the problems due to the AI-generated reputational content. The current legislation only defines the following:

- The accountability of independent AI content.
- Standards of breach or responsibility of AI developers.
- The responsibility for harm caused without the action of humans.

The laws' absence instead of giving protection, creates an irregularity in fulfillment. There is an incomplete coverage of issues of defamation and redress of the contents generated by AI in the Indian laws. It is necessary to achieve accountability without interrupting the developments in technology for which we need judicial regulations and legal adjustments.

Digital and AI Defamation

Transition to digital communication has changed the limited interpersonal damage to a greater social damage. Location and time constraints won't be an obstacle to the digital media to provide the accessibility to online content. The reputation damaging content causes harm by broadcasting it to a large public audience⁷. Through messaging services, social media, and online news sites a large number of defamations have been reported. It is very difficult to find the person who submitted the defamatory content due to the Lack of personal attribution and use of pseudonyms⁸. The hosting or sharing of materials internationally raises jurisdiction problems which makes the enforcement of conventional methods ineffectual. Such elements delay the process of getting answers and cause the

⁶South, S. N., 'Crafting a Liability Regime for AI Systems in India' (2024).

⁷Sharma, A., 'The Evolving Landscape of Online Content Regulation in India: Problems, Challenges, and Legal Perspectives.' [2024] International Journal of Law Management & Humanities 2258

⁸Metzler H and Garcia D, 'Social Drivers and Algorithmic Mechanisms on Digital Media' (2023) 19 Perspectives on Psychological Science 735 <<https://journals.sagepub.com/doi/full/10.1177/17456916231185057>>

worst damage to the organization structure of internet platforms. Instead of accuracy, the core areas of automated systems content are engagement, popularity, and virality⁹. Because of the automated amplification method, it repeats the malicious contents without any analysis. Such challenges are increasing along with growing usage of artificial intelligence on digital media. AI systems have become directly involved in content creation and no longer assume the mediation of human speech. It radically changes the terms in which the laws of defamation will be executed and reveals the weak points of the existing legal regulations.

Concept of AI-Generated Defamation

AI-generated defamation is reputational damage that is the result of autonomously produced content in artificial intelligence systems¹⁰. The systems can produce text, images, audio, or video with the help of statistical models trained on large datasets. The results are obtained based on pattern recognition and not factual validation, and this increases the chances of errors and misrepresentation. The process of human authorship dilution is the main characteristic of AI-sourced defamation. Developers create it, and it is prompted by users, yet the outcome is developed by automated mechanisms that are not predictable or controllable. This dilutes the classical connection between speaker, intent and responsibility. The harm is also enhanced by the style of presentation of AI-generated content. The fluency of outputs, well-structured and unemotional in tone, creates an impression of reliability. In case of this type of delivery of incorrect information, there is a higher possibility of this being believed and spread. The damage to reputation may occur even without intentional harm. The Indian law has no present differentiation between human speech and machine-generated speech¹¹. This theoretical void gives ambiguity to the courts, which makes use of fault-based teachings. It is important to acknowledge that AI-generated defamation is a unique phenomenon to create effective reactions to it.

Defamation Risks Arising from Conversational AI Systems

Conversational AI systems like ChatGPT use large language models that are trained on massive amounts of text data collected through books, articles, websites, and other sources of text made

⁹ Meti, M. V., 'Balancing Free Speech with Hate Speech and Defamation: An Indian Perspective' 2 International Journal of Law and Research

¹⁰ Moreno FR, 'Generative AI and Deepfakes: A Human Rights Approach to Tackling Harmful Content' (2024) 38 International review of law computers & technology 1

¹¹ Hindelang S and Moberg A, *YSEC Yearbook of Socio-Economic Constitutions 2021 : Triangulating Freedom of Speech* (Springer 2022)

publicly available¹². Such systems do not scan databases or verify facts in real time. They instead produce responses by forecasting the most likely order of words on the basis of patterns that they have acquired in training. This leads to the generation of language that seems to be coherent, confident and authoritative even when the underlying information is false or missing. This manner of operation presents a particular risk of libel. It is usually achieved in a phenomenon that is referred to as hallucination, wherein the system produces information that lacks factual support. In some other situations, timely sensitivity is involved where minor differences in the queries of the users can prompt the system to establish wrongdoing or relationships that may not be proven. Notably, these defamatory contents can be produced without the bad intention of either the user or the developer.

For example: “A customer requests a chatbot to give a summary of the controversies of a local business proprietor. The chatbot answers that the person was committing financial fraud, although nothing of the kind has been claimed. The user, relying on the answer, adds it to a report that he/she shares online, which imposes reputational damage on the person”.

This case illustrates the harm that can be caused to the reputation in the case of indirect dependence on the AI-generated content. Chatbots use the conversational format, which builds trust when responses are presented in the form of direct answers and not suggestions. The users can hence cite the information as true and replicate it in their scholarly literature, work papers or on the internet. The fact that the traditional law of defamation finds it difficult to cope with this kind of harm is that the laws of fault require intent or knowledge. The chatbots is not legally a person, and the developers are not in direct control of specific outputs in real-time. This issue is complicated by the fact that liability is attributed on the basis of automated predictions and not intentional publication. The popular use of conversational AI systems adds further risk of secondary dispensation¹³. The results of chatbots are easy to copy, share and spread through the internet, prolonging reputational damage even after the initial exchange.

Risks of Deepfakes and AI Misinformation

The falsely produced or manipulated digital content that looks, sounds, or appears very similar to real persons is called deepfakes. Contrary to previous media manipulation, deepfakes depend on artificial intelligence which produces outcomes with more accuracy, so it becomes difficult to differentiate authentic ones from the deep fake content. Because it makes the fraudulent content as authentic proof

¹² Bansal G and others, ‘Transforming Conversations with AI—a Comprehensive Study of ChatGPT’ (2024) 16 Cognitive Computation

¹³ Preininger AM and others, ‘Artificial Intelligence-Based Conversational Agent to Support Medication Prescribing’ (2020) 3 JAMIA Open 225

this has a direct impact on reputation. It represents an example of artificial media, which was generated by machine learning algorithms trained using extensive datasets¹⁴. The Generative models are considered as technological foundations because their outcomes are gradually approaching authenticity such that they not only reproduce the physical appearance but also the observable behaviors such as tone accents and language style.

Consequently, synthetic media can convincingly imitate real people in various formats. This possibility heightens the chances of defamation since deceitful statements can be looked at as valid without question. Deepfakes are presented in several types of forms that pose certain threats to the reputation and confidence of the population. The formats are different, but their commonality lies in the fact that they mislead the viewers by being realistically simulated. They are listed as follows:

i) Video Deepfakes

These are the manipulation of the face expressions or movements of a person in video recordings. Such videos can misrepresent people as saying or doing things they never did. Due to the popular opinion of strong evidence of video, video deepfakes are especially harmful in cases of defamation¹⁵. After being posted on online sources, doctored videos can easily go viral through sharing and algorithmic promotions. Reputational damage can be quite persistent even after the verification because of recurrent circulation and recollection. Video deepfakes are visual, and therefore denial or correction is not as effective as when dealing with text-based allegations.

ii) Audio Deepfakes

It refers to the artificial creation of voice recordings to resemble the speech style of a person. Voice synthesis provides artificial intelligence with the ability to reproduce tone, accent and mannerism accuracy. This type of recording can be used in professional, political or personal conflict. Due to the perceived credibility of voice recordings, reputational damage can be inflicted with no visual verification that it was an audio deepfake¹⁶.

iii) Image Deepfakes

This involves images that consist of altered photographs or fake images of people. Artificial intelligence systems may produce life-like images that represent individuals in a scenario that has never existed. These pictures can be easily shared, but are difficult to trace after being shared. The reputational effects of image-based deepfakes are major, especially when the images depict illegal,

¹⁴ Farid H, 'Creating, Using, Misusing, and Detecting Deep Fakes' (2022) 1 Journal of Online Trust and Safet

¹⁵ Harris, K. R., 'Video on Demand: What Deepfakes Do and How They Harm' (2021) 199 Synthese 13373

¹⁶ Hery, A., Joseph, O., Femi, O., & Luz, H., 'Audio Deepfakes: Threats to Voice Assistants and Voice-Activated Systems.'

offensive or compromising situations of individuals¹⁷. Even temporary circulation may cause long-term damage through repeated copying, reposting and archiving across multiple platforms.

iv) Textual Deepfakes

Through the artificial intelligence language models it includes the development of fake textual content written in authoritative formats which comprises fake news stories, fabricated interviews and deceitful reports. It can developed into realistic or credible texts. The content has a professional structure that is considered to be credible by readers¹⁸. These main four ddeepfakes generate text, audio, and visuals synthetically are real by faking the real. So it can categorised as deception and disinformation. The permit to microblog and network by the social media networks and corporations is the major reason for the rapid spreading of AI generated disinformation. Video-sharing websites, blogs, forums, and other content-hosting platforms play the role of distributors by hosting and recommending audio-visual content, including artificial intelligence-generated videos. There is a greater chance of sharing defamatory information through both individual and group messages. Even the online news aggregators also select and regenerate the digital content automatically.

Technical Risks and Prevention Methods

The technical risk of AI-generated defamation emerges from how conveniently the scalable and realistic AI technologies can be accessed. The mismatch between the content's production rate and its identification rate is one of the threats.

The AI generated defamatory and fraudulent information is more difficult to identify because those defamatory information lack watermarks¹⁹. The AI generated disinformation is untraceable whereas the generation is easier and can be done instantly. The preventive measures to avoid the difficulty in differentiating the authentic and fraudulent content includes content validation, tracking, and simplification.

Some of the tools that would help in the validation process would be digital watermarking, meta-tagging and content-authentication protocols. Detectors based on AI are being used more and more to detect doctored content, although they are not accurate. Indian research emphasizes that all of these need to operate together and under government supervision.

The need for technology-related solutions, on the other hand, is also a function of a governing environment. Risk assessments, security audits, and monitoring before installation, as well as

¹⁷ Shahzad HF and others, 'A Review of Image Processing Techniques for Deepfakes' (2022) 22 Sensors 4556

¹⁸ Blitz, M. J. , 'Lies, Line Drawing, and Deep Fake News' [2018] Okla 71

¹⁹ Vinolia, J. J., 'Unmasking Digital Deception: Legal Accountability of Social Media Platforms for Deep Fake Content.' (2024) 5 Jus Corpus LJ 157

monitoring afterwards, may point to possible misuse²⁰. The integration of a reporting and logging system is a way of accountability. However, technology-related solutions by themselves are inadequate.

Ethical Considerations and Developer Responsibility

The chapter outlines a distinct gap with regard to ethical responsibility pertaining to AI-generated defamatory content. The quick advancement of generative AI systems has raised the threat to reputation and personal identity. The lack of explicitly defined ethical responsibilities of the developers undermines the security against abuse. The implementation The absence of transparency on the information on AI-generated content works against awareness and trust of users²¹. The scarcity of information regarding the limits of the system promotes the dissemination of false information. These lack user instructions, which promote excessive reliance on AI outputs. The existing practice demonstrates poor focus on frequent system audits. The absence of regular inspections postpones the identification of abuse threats. There is a lack of good organisational accountability, which limits a good response once harm is caused. The problem of the imbalance of reputation protection and freedom of expression is still unaddressed. According to the chapter, preventive ethical conditions are still inadequate. The necessity of audit, transparency, and accountability is still underexploited. This lack of cohesive ethical governance brings about uncertainty for victims. Such a weakness justifies the requirement for more ethical considerations in the development and implementation of AI.

Conclusion and Recommendation

This chapter discussed the law of defamation in India concerning the field of digital communication and artificial intelligence. The discussion has started by examining defamation under the Indian Penal Code and whether it persists under the Bharatiya Nyaya Sanhita. The examination also examined the applicability of the Information Technology Act in the resolution of reputational damage in the online space. The evolution of traditional human communication up to AI-generated conversational content dealt with the development of AI-created defamation and its legal implications. It suggested the risks that can be posed by conversational AI systems that produce autonomous textual outputs. The chapter has also discussed deepfake technologies of both video and audio, image and text fabrication. This

²⁰ Javadi, S. A., Cloete, R., Cobbe, J., Lee, M. S. A., & Singh, J., 'Monitoring Misuse for Accountable' Artificial Intelligence as a Service'', *In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2020)

²¹ Prasad, N. L., & Jain, I. D., 'Technological Progress vs. Human Rights: Privacy, Free Expression, and Reputation in the Era of AI and Deepfakes' (2025) 2 *Journal of Legal Research and Polity* 89

analysis thus indicated the increased severity of reputational damage due to realistic and scalable synthetic content. The chapter then evaluated the contribution of platforms and intermediaries to the facilitation of the spread of AI-motivated defamatory content. It identifies the technical weaknesses of algorithmic amplification and automatic content generation. The ethical problem discussed in the chapter is the issue of accountability and responsibility in AI development. The general conclusion is that the current defamation regulations have systemic weaknesses in dealing with AI-assisted reputational damage.

The discussion hence suggests that legal principles should be adapted to new technological realities. The analysis also indicates that an effective enforcement of liability allocation requires clarity. The suggestions assist in the clear legalization of AI-generated defamation in the current systems. The regulating strategy must share responsibility between the developers, deployers, and digital platforms. Future governance of AI defamation should be coordinated through the integration of legal issues, technical issues, and ethical issues.

-----*****-----