DEEPFAKES AND THE DEATH OF EVIDENTIARY CERTAINTY: RETHINKING DIGITAL PROOF IN INDIAN CRIMINAL TRIALS

Jiya Jhaveri & Ishya Ramchandani, NMIMS, Kirit P Mehta School of Law

ABSTRACT:

The increasing use of digital evidence in criminal trials has transformed the process of fact-finding by offering seemingly objective representations of events through videos, audio recordings, and electronic data. However, recent advancements in artificial intelligence, particularly deepfake technology, have begun to undermine this perceived reliability. Deepfakes, which involve the creation of highly realistic synthetic media using machine learning techniques, blur the distinction between authentic and fabricated content, raising serious concerns for evidentiary integrity.

This paper examines the implications of deepfakes for Indian criminal trials, focusing on their impact on evidentiary standards and judicial decision-making. It argues that the challenge posed by deepfakes is not limited to the risk of false evidence being introduced in court, but extends to a broader erosion of trust in digital proof itself. The concept of the "liar's dividend" is central to this analysis, highlighting how the mere possibility of manipulation enables individuals to deny even genuine evidence, thereby weakening its probative value.

The study analyses the existing legal framework governing electronic evidence in India, including the provisions of the Indian Evidence Act, 1872 and the Bharatiya Sakshya Adhiniyam, 2023, along with relevant judicial precedents. It demonstrates that while these laws emphasise procedural safeguards such as certification and admissibility, they do not adequately address questions of substantive authenticity in the context of AI-generated media. A comparative perspective is also undertaken, examining approaches adopted in other jurisdictions to regulate and authenticate synthetic content.

The paper concludes that deepfakes expose a structural limitation in current evidentiary doctrine, necessitating a shift towards more robust methods of verification and evaluation. It highlights the need to reconsider how authenticity is established, how doubt is managed, and how the legal system can maintain confidence in its fact-finding function in an era where digital representations can no longer be presumed to reflect reality.

Introduction:

Criminal trials are fundamentally premised on the discovery of truth. Courts depend on evidence to reconstruct events, and over time, digital material especially videos, audio recordings, and electronic data has come to play a central role in this process. Such evidence has often been treated as inherently reliable because it appears to capture reality as it happened. However, recent advances in artificial intelligence have begun to challenge this assumption in a serious way. When technology can convincingly fabricate events that never occurred, the line between truth and falsehood becomes increasingly difficult to draw.

Deepfakes, a form of AI-generated synthetic media, illustrate this problem clearly. Using machine learning techniques such as generative adversarial networks (GANs), it is now possible to create highly realistic videos, images, and voice recordings that closely mimic real individuals. What makes this development particularly troubling is not only the quality of these outputs but also their accessibility tools for creating such content are no longer limited to experts. As noted in recent scholarship, deepfakes have evolved from experimental technology into a widely available tool capable of producing convincing false evidence with minimal effort.

The concern is particularly significant in the Indian context, where digital evidence has become increasingly central to criminal investigations and prosecutions. CCTV footage, mobile phone recordings, and social media data frequently form the backbone of evidentiary claims. The legal framework governing such material earlier under Section 65B of the Indian Evidence Act, 1872 and now under Section 63 of the Bharatiya Sakshya Adhiniyam, 2023 primarily focuses on procedural aspects such as certification and admissibility. However, these provisions were developed in a technological environment where the authenticity of digital records was less contested and do not fully address the risks posed by AI-driven manipulation (IJFMR, 2025).

In this context, the rise of deepfake technology does not merely introduce a new evidentiary issue; it raises deeper questions about the reliability of digital proof itself. It forces a reconsideration of how courts determine authenticity, how burdens of proof should be allocated, and whether existing legal safeguards are sufficient in an era of synthetic media. This paper examines these challenges and argues that the current Indian legal framework is not adequately equipped to deal with the evidentiary uncertainties created by deepfakes, making reform both necessary and urgent.

Understanding deepfakes, AI and synthetic media:

Deepfakes refer to a category of synthetic media which is used to create or manipulate visual and audio content making it appear authentic. The term deepfake was first used in 2017 to refer to a case of a celebrity video clip being combined with a deep learning method in order to deceive audiences. Unlike traditional forms of digital editing which often leaves tangible traces of manipulation, deepfakes are designed to replicate facial expressions, voice and body movements with a very high degree of precision which makes the detection difficult.

At the core, deepfake technology are machine learning models, particularly Generative Adversarial Networks (GAN). These systems operate using a dual process mechanism. One network generates fake content while the second network evaluates its authenticity. This process repeatedly continues and the generated output becomes progressively realistic, often to a point where it is indistinguishable from genuine recordings.

Deepfakes can take several forms, the most commonly discussed are face swapping where an individual's face is super imposed onto another's body.

Another growing form is voice cloning, where AI models clone a person's speech patterns, pace and tone to produce fabricated audio recordings. Additionally, Lipsyncing manipulation allows existing videos to be altered so that individuals appear to say things they never actually said. These variations display that deepfakes are not limited to visual manipulation but extend across sensory dimensions making them particularly convincing as forms of evidence.

As noted in the International Journal for Multidisciplinary Research (IJFMR) Study[1], this technological advancement has significantly lowered the barrier to creating convincing/ realistic synthetic media allowing even non- experts to produce manipulated content with readily available tools.

The growing concern regarding deepfakes is their increasing realism. Earlier types of manipulated media could be identified through inconsistencies such as distorted edges, mismatched features and unnatural glitches. However, developments have significantly reduced these flaws.

[1] International Journal of Future Multidisciplinary Research. (2025). *Deepfake evidence and its legal challenges in India*.

Studies have shown that humans struggle to distinguish between real and AI generated content.

Chesney and Citron, 2019[2]: Chesney and Citron (2019) argue that deepfakes not only enable fabrication but also erode trust in genuine evidence. They highlight that this loss of trust allows individuals to deny authentic material, thereby weakening the reliability of digital proof.

At the same time, Deepfakes aren't necessarily harmful, they have legitimate uses in areas such as film making, education and accessibility.

For example, it can be used to recreate historical figures for academic purposes or generate dubbing in multiple languages.

However, their misuse has gained greater attention in context involving misinformation, fraud, harassment, and more.

In the legal context, the problem is not that deepfakes exist but that they reduce the reliability of digital evidence as a whole when any audio- visual material can be fabricated. Courts can no longer trust such evidence. Hence, the transition from a system where digital evidence was presumed reliable to one where its authenticity must be actively and strictly established.

Role of deepfake in criminal cases & Legal Frameworks existing in India:

Electronic records such as CCTV footage, call recordings, emails, whatsapp messages, social media content are frequently relied upon to establish facts and record testimonies. This depicts digitisation of everyday life where human activity increasingly leaves behind electronic traces that can be presented in court.

The law has gradually adapted to the reality of digital footprint. Under the Indian Evidence Act, 1872, electronic records were formally recognised as admissible evidence under Section 65(b) which laid down conditions for their admissibility. This framework has been carried forward under Section 63 of Bharatiya Sakshya Adhiniyam, 2023.

These laws were designed to ensure digital evidence presented in court by requiring proof of

[2] Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, *107*(6), 1753–1820. https://doi.org/10.15779/Z38RV0D15J

its integrity. Judicial interpretation has reinforced the importance of these safeguards.

In the case Anvar P.V. v. P.K. Basheer, the Supreme Court clarified that electronic evidence is admissible only if it complies with the requirements of Section 65B of the Indian Evidence Act. The Court rejected the earlier approach of treating electronic records like primary evidence. It emphasised that certification is a mandatory condition, not a procedural formality. This decision firmly established the importance of procedural safeguards in ensuring the reliability of digital evidence.

In Arjun Panditrao Khotkar v. Kailash Kushanrao Gorantyal, the Supreme Court reaffirmed and strengthened the ruling in Anvar P.V., holding that Section 65B certification is indispensable for admissibility of electronic records. It clarified that such certification can be produced even at a later stage of proceedings if initially unavailable. The judgment recognised practical difficulties but maintained strict compliance with procedural requirements. It highlights the judiciary's continued reliance on formal authenticity as a guarantee of evidentiary reliability.

The Importance of digital evidence lies in its perceived objectivity. As a commentator puts it, "what appears to show something may have never occurred." Complementing these evidentiary provisions, our substantive laws specifically address misuse of digital content. The Information Technology Act, 2000, criminalises acts such as identity theft under Section 66(c), under Section 66(d) cheating by personation using computer resources and publication or transmission of obscene material in electronic form under Section 67.

The Bharatiya Nyaya Sanhita, 2023 contains provisions relating to forgery of electronic records and manipulation, which may be invoked where altered digital content is used to cause harm. Together, these statutes attempt to regulate both, the admissibility and misuse of electronic evidence within the criminal justice system.

The deepfake evidentiary crisis is more than a procedural challenge in India, directly implicating multiple fundamental rights in the Constitution. The right to fair trial, under Article 21, is threatened when the Courts admit electronic material without their authenticity being reliably determined. The presumption of innocence is also eroded when deepfake visuals and audios provide counterproductive evidence that is hard to negate.

Article 14's guarantee against arbitrary state action is equally engaged. Relying on unverified synthetic media constitutes arbitrary action under Shayara Bano v. Union of India (2017). Deepfake detection tools provide only probabilistic outputs (e.g., "87% confidence"), not definitive determinations, and no legal standard specifies what confidence level satisfies proof beyond reasonable doubt. This opacity along with unequal access to forensic deepfake detection, marginalized and rural accused persons often lack such resources, creating indirect discrimination violating Article 14.

Article 19(1)(a) presents a tension: while evidentiary use of deepfakes falls outside free speech protection, labelling and regulation regimes may incidentally burden legitimate expression such as satire and political commentary.

Moreover, Article 20(3)'s protection against self-incrimination is less directly implicated, though compelling an accused to provide biometric samples (voice, gait, facial data) for deepfake comparison would require careful scrutiny. Collectively, these constitutional dimensions demand a synthetic-aware legal framework that balances liberty, equality, and expression while preserving fair trial rights.

Unlike oral testimony which may be affected by lapse in memory or by bias digital records are often treated as precise representations of events of a video recording can capture actions in real time. While meta data associated with electronic files can provide information about time and location. Where establishing factual accuracy is crucial,this has made digital evidence valuable in criminal cases.

However, this reliance also makes the system vulnerable. The laws governing digital evidence in India largely focus on procedural authority. Procedural authenticity making sure that the record was properly stored[3]

Evidentiary crisis: deepfakes and liars dividend:

The rise of deepfake technology has created a serious evidential crisis in criminal law by weakening the assumption that evidence affects reality that traditional courts have proceeded on the understanding that although evidence may be disputed it is possible to verify digital

[3] Record of Law. (n.d.). *Deepfakes and digital evidences: A new challenge to the Indian judiciary*. https://recordoflaw.in/deepfakes-and-digital-evidences-a-new-challenge-to-the-indian-judiciary/

evidence especially video and audio recordings have been treated reliable because of its apparent objectivity. However, over time, the process of fact finding is placed under strain and this reliability becomes unstable.

A significant aspect of this crisis is what scholars describe as "liars dividend." This term refers to the advantage gained by individuals who deny genuine evidence by claiming that it is deepfake. The highlight of this concept lies in its duality : deepfakes do not just enable creating false evidence, they also undermine trust in authentic evidence. Due to this, even accurate and credible material can be put into doubt, making it harder for courts to rely upon.

A mere suggestion that a video or audio clip could be fabricated is often sufficient to weaken its evidentiary value, particularly in a system that places high importance on proof beyond reasonable doubt. This has serious implications for the burden of proof in criminal trials. The prosecution is required to establish its case through trustworthy evidence while the defense may challenge its reliability. However, the possibility of deepfakes allows defense to introduce doubt without necessarily providing substantive proof of manipulation. This creates a risk of what may be termed "speculative scepticism" where courts are compelled to consider hypothetical possibilities of fabrication even in the absence of concrete evidence. If accepted too easily, such arguments could make it difficult for the prosecution to rely on digital material.

Judicial developments show an emerging awareness of this issue. In Tomaso Bruno vs State of Uttar Pradesh, 2015, the Supreme Court emphasized the importance of electronic evidence such as CCTV Footage in criminal investigations observing that such material can provide objective insights into the facts of the case. This judgement reflects the judiciary's increasing dependence on digital records as reliable forms of proof.

In Anwar PV vs PK Basheer, the court underscored the need for strict compliance with certification requirements for electronic evidence reinforcing the idea that procedural safeguards ensure reliability.

Nonetheless, both these decisions operate on an implicit assumption that once properly admitted, electric evidence is fundamentally trustworthy. It is precisely this assumption that deepfake technology destabilises. The evidentiary crisis is further concentrated by the absence of clear judicial standards for addressing claims of manipulation when a party alleges that a piece of digital evidence may be a deep fake, courts currently lack a structured framework to

asses such cases. There is uncertainty as to whether the burden should shift to the party relying upon the evidence to prove its authenticity, rigorously or whether the party alleging fabrication should first establish a prima facie case. This lack of clarity can lead to inconsistent approaches and uncertainty in adjudication.

Another side of this crisis, lies in its systematic consequences, if courts begin to treat all digital evidence with increasing suspicion, its overall evidentiary value may decrease even when genuine. On the other hand, if courts continue to rely on such evidence without deeper scrutiny, there is a risk of wrongful convictions based on fabricated material. The legal system is therefore caught between two competing risks- over reliance and over skepticism, both of which threaten the credibility of the criminal judgement.

The liars dividend thus exposes structural weakness in the current evidentiary document. It enables doubt to be introduced at minimal cost while verifying authenticity remains complex and resource intensive. This imbalance has the potential to distort the fact finding process making it increasingly difficult for courts to make confident conclusions.

In this context, the evidentiary crisis requires a rethinking of traditional assumptions about proof, courts may need to adopt a more calibrated approach where mere assumptions of manipulation are insufficient but credible challenges to authenticity trigger deeper inspection and forensic verification.

The deepfake evidentiary crisis is more than a procedural challenge in India, directly implicating multiple fundamental rights in the Constitution. The right to fair trial, under Article 21, is threatened when the Courts admit electronic material without their authenticity being reliably determined. The presumption of innocence is also eroded when deepfake visuals and audios provide counterproductive evidence that is hard to negate.

[4]Comparative Perspective: EU and United States Approaches:

Within the European Union, the EU AI Act introduces a transparency-based model. Under Article 52, there is a requirement that artificially generated or manipulated content particularly

[4] Lawjurist. (2026, February 19). *Deepfake evidence and the law of admissibility: A critical analysis under the Indian Evidence Act, 1872*. https://lawjurist.com/index.php/2026/02/19/deepfake-evidence-and-the-law-of-admissibility-a-critical-analysis-under-the-indian-evidence-act-1872/

deepfakes must be clearly disclosed to users. The emphasis here is not on evidentiary admissibility but on preventing deception at the source. By imposing a legal obligation to label synthetic media, the framework attempts to preserve informational integrity before such content enters public circulation or institutional processes. This approach reflects a preventive logic: if the artificial nature of content is disclosed early, its misuse whether in public discourse or legal settings can be mitigated.

The United States, by contrast, addresses the issue through evidentiary doctrine rather than regulatory mandates. Rule 901(b)(9) of the Federal Rules of Evidence allows authentication through evidence describing a process or system and demonstrating that it produces accurate results. This provision becomes particularly relevant in the context of AI-generated or digitally processed material, as it shifts attention to the reliability of the underlying mechanism rather than the form of the evidence alone. Instead of presuming authenticity, courts are permitted to examine how a piece of evidence was generated, processed, or verified, often relying on technical explanations and expert input.

The contrast between these approaches is instructive. The EU model seeks to ensure that synthetic content is identifiable through mandatory disclosure, thereby reducing the likelihood of deception. The US model, on the other hand, accepts that complex digital material will enter legal proceedings and focuses on testing its credibility through demonstrable reliability of systems and processes. One operates primarily at the stage of dissemination, the other at the stage of adjudication.

For India, this comparison highlights two distinct gaps. There is currently no structured obligation requiring disclosure of AI-generated content, nor is there a developed evidentiary practice that systematically evaluates the reliability of technological processes behind digital material. As a result, courts are left without clear tools either to identify synthetic media at its origin or to rigorously assess it during trial.

Adapting elements from both systems could offer a more balanced response. A disclosure-based obligation would help trace the origin and nature of digital content, while a process-oriented approach to authentication would allow courts to engage more meaningfully with questions of reliability. Together, these strategies suggest a shift towards recognising that, in the context of AI-generated media, authenticity cannot be assumed; it must be both signalled and demonstrated.

Proposed Reforms Addressing Deepfake Evidence:

The challenges posed by deepfakes require a shift from traditional, device-focused authentication towards a more layered and content-sensitive approach. One significant reform in this regard is the introduction of an Evidentiary Provenance Passport, which would function as a comprehensive forensic dossier for electronic records whose authenticity is disputed. Unlike Section 63 of the Bharatiya Sakshya Adhiniyam, 2023, which primarily certifies the device and chain of custody, this mechanism would incorporate multiple layers of verification, including timestamping through cryptographic hashing, editing history logs, AI-based detection reports, and biometric consistency checks. Such a model would move authentication beyond procedural compliance and towards a more substantive evaluation of content.

Closely connected to this is the need to regulate the misuse of deepfake allegations through a reverse burden of production. Where an accused claims that electronic evidence is fabricated, they should be required to provide a minimal evidentiary basis for such a claim, such as expert input or indicators of manipulation. Only upon meeting this threshold should the burden shift back for deeper verification, while preserving the prosecution's ultimate obligation to prove its case beyond reasonable doubt. This would help contain speculative challenges without undermining legitimate concerns.

Institutional support within courts must also be strengthened. The appointment of court-based technical advisors, functioning as neutral experts, would enable judges to better understand the complexities of AI-generated evidence. These experts could assist in evaluating forensic reports, clarifying technological processes, and ensuring that judicial reasoning is informed by technical accuracy rather than assumption.

At the investigative stage, reform is equally necessary. Introducing specialised training and certification for law enforcement personnel handling digital evidence would ensure that collection, preservation, and preliminary assessment are conducted with due care in cases involving potential manipulation. This would reduce the risk of evidentiary contamination and improve coordination with forensic analysis.

Procedural transparency must also be enhanced through a structured disclosure requirement for digital evidence, particularly where artificial intelligence tools have been used in its creation, enhancement, or analysis. Early disclosure of such details would allow for meaningful

challenge and prevent unfair surprise during trial.

Finally, courts must adapt to the probabilistic nature of modern forensic tools. This requires a shift in judicial reasoning to accommodate degrees of reliability rather than binary conclusions, supported by clear guidance on how such evidence should be interpreted. Together, these reforms aim to recalibrate the evidentiary process in a manner that acknowledges technological realities while preserving the integrity of criminal adjudication.

Conclusion:

Deepfake technology represents a profound disruption to the foundations of criminal adjudication. The difficulty it creates is not limited to the possibility of fabricated evidence, but lies in its ability to destabilise confidence in all digital material. When images, videos, and audio recordings can be convincingly manipulated, the assumption that such evidence reflects reality becomes increasingly fragile. This places courts in an uncertain position, where reliance on digital proof is necessary, yet trust in its authenticity is no longer secure.

The resulting uncertainty affects the very structure of criminal trials. Doubt, which traditionally operates as a safeguard for the accused, can now be invoked with minimal effort through claims of manipulation. This alters how evidence is perceived and evaluated, making it harder to distinguish between legitimate scepticism and strategic denial. In such a setting, the process of fact-finding risks becoming less about determining what actually occurred and more about navigating competing possibilities of truth.

What emerges is a deeper challenge to the idea of evidentiary certainty itself. The legal system is compelled to confront a reality where appearances can no longer be taken at face value and where verification becomes more complex than ever before. If this shift is not addressed with clarity and precision, it risks gradually weakening the credibility of judicial outcomes. In this sense, deepfakes do not merely complicate proof they test the capacity of the criminal justice system to sustain trust in its own conclusions.