 DOI: https://10.5281/zenodo.22912311

Deepfakes, Criminal Responsibility and the Limits of India's Cybercrime Framework: Rethinking Liability in the Age of Generative AI

Srishti Punj[1]

© The Author(s), under exclusive licence to Oxford Law Review 2026.

Abstract

Generative artificial intelligence has transformed the production of synthetic audio-visual content, making it increasingly difficult to distinguish authentic recordings from artificially generated or manipulated representations. Deepfakes present a distinctive challenge for criminal law because the technology can facilitate fraud, impersonation, extortion, non-consensual sexual imagery, reputational harm, and other forms of digital wrongdoing while simultaneously complicating the attribution of criminal conduct to a particular human actor. This article examines whether India's existing criminal-law framework is capable of assigning responsibility for unlawful conduct involving deepfakes, with particular attention to the Bharatiya Nyaya Sanhita 2023, the Information Technology Act 2000, the Bharatiya Sakshya Adhiniyam 2023 and the evolving intermediary framework governing synthetically generated information. The article adopts a doctrinal and comparative methodology, supplemented by an analysis of the emerging relationship between artificial intelligence, mens rea, attribution and criminal responsibility. It argues that the principal difficulty is not simply the absence of a dedicated deepfake offence, but the fragmentation of existing legal responses and the uncertainty surrounding the attribution of culpability where human conduct is mediated by generative AI. The article develops a differentiated framework based on human agency, intention, knowledge, control and resulting harm, distinguishing the criminal responsibility of creators, users, disseminators and intermediaries. It further examines comparative approaches in the European Union, United Kingdom and United States to identify principles that may inform Indian law without mechanically transplanting foreign regulatory models. The article concludes that India's response should preserve the foundational principle of individual criminal responsibility while developing technologically informed rules for attribution, proof and differentiated liability. Such an approach can address serious deepfake-enabled harms without converting every act involving synthetic media into a criminal offence.

Keywords: Deepfakes; Generative Artificial Intelligence; Criminal Responsibility; Mens Rea; Cybercrime; Synthetic Media

1. Introduction

[1] **Department of Laws, Panjab University, Chandigarh, India**

 DOI: https://10.5281/zenodo.22912311

"Code is law." — Lawrence Lessig (1999, 6)

When Lawrence Lessig famously described code as a form of regulation capable of shaping human behaviour in cyberspace, artificial intelligence had not yet reached the generative capabilities that now define the digital environment. His proposition, however, acquires renewed significance in the age of generative artificial intelligence ('AI'). Contemporary AI systems do not merely regulate access to digital spaces or facilitate the transmission of information; they can generate highly realistic images, audio, video and other forms of synthetic media capable of depicting persons saying, doing or appearing to do things that never occurred. Lessig's observation therefore provides an appropriate conceptual starting point for examining deepfakes: when technological systems increasingly determine what can be created, altered, circulated and believed, criminal law must confront not only the conduct of human actors but also the technological processes through which that conduct is mediated (Lessig 1999, 6).

Deepfakes represent one of the most consequential manifestations of this transformation. Although manipulated digital content predates generative AI, recent developments in machine learning have substantially lowered the technical barriers to producing convincing synthetic media. Images can be generated or altered, voices can be cloned, faces can be superimposed and videos can be produced in which individuals appear to make statements or engage in conduct that never occurred. The resulting problem is not merely one of technological authenticity. It is increasingly a problem of criminal responsibility: when synthetic content causes legally cognisable harm, who should be regarded as responsible for its creation, publication or dissemination?

The question becomes particularly difficult because a deepfake may involve multiple actors and stages. An individual may provide the source material, employ a generative AI system to create the synthetic content, edit or refine the output, upload it anonymously, and subsequently distribute it through multiple online platforms. The AI system may have generated the final output through processes that are technically complex and not entirely predictable to the user. At the same time, intermediaries may facilitate the storage, recommendation or dissemination of that content. The conventional criminal-law model, which ordinarily seeks to connect a prohibited act and a culpable mental state to an identifiable human actor, therefore encounters a technologically mediated chain of conduct.

The legal significance of deepfakes extends across several categories of harm. Synthetic media may be employed for impersonation, fraud, extortion, harassment, reputational attacks, non-consensual intimate imagery and other unlawful purposes. Importantly, however, not every deepfake is inherently unlawful or criminal. Synthetic media may also be created for satire, artistic expression, education, entertainment, research and other legitimate purposes. A criminal-law response must therefore distinguish between technological capability and legally culpable conduct.

The central question is not whether deepfake technology is inherently dangerous, but under what circumstances its use satisfies the established requirements of criminal liability.

This distinction has become particularly important in India because the regulatory landscape has changed considerably during 2025 and 2026. In October 2025, the Ministry of Electronics and Information Technology ('MeitY') proposed amendments to the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, specifically addressing the growing misuse of synthetically generated information, including deepfakes. The proposal identified risks associated with impersonation, non-consensual intimate imagery, fabricated content, fraud and the misleading portrayal of individuals or real-world events. (MeitY)

A significant development followed on 10 February 2026, when the Information Technology Rules were amended to introduce specific due-diligence obligations concerning synthetically generated information ('SGI'). The amended framework requires relevant intermediaries that facilitate the creation or dissemination of such information to deploy reasonable and appropriate technical measures against specified unlawful forms of synthetic content. It also introduces requirements concerning prominent labelling and technological provenance mechanisms for permissible synthetically generated information. The amended provisions expressly contemplate content that falsely depicts a natural person or real-world event in a manner likely to deceive, alongside categories such as non-consensual intimate imagery and false electronic records. (MeitY)

These developments are significant, but they do not by themselves resolve the underlying question of individual criminal responsibility. Intermediary obligations operate principally at the level of platform governance and content regulation. Criminal liability, by contrast, requires a determination of whether a particular person committed an offence with the requisite mental state. The existence of a regulatory duty upon an intermediary does not necessarily answer whether the person who generated or disseminated a deepfake possessed the intention or knowledge necessary for criminal liability. This distinction exposes an important analytical gap between content regulation and criminal attribution.

The problem is further complicated by India's transition to a new substantive criminal-law framework. The Bharatiya Nyaya Sanhita, 2023 ('BNS') has replaced the Indian Penal Code, 1860 as the principal penal statute, while the Bharatiya Sakshya Adhiniyam, 2023 ('BSA') has replaced the Indian Evidence Act, 1872. Consequently, an examination of deepfake-related criminality can no longer be based exclusively on the pre-2024 legal framework. It requires an assessment of how existing offences concerning deception, personation, forgery, reputation and related forms of unlawful conduct interact with AI-generated synthetic content, as well as how such conduct can be proved through electronic evidence.

To what extent does India's existing criminal-law framework adequately attribute criminal responsibility for the creation and dissemination of AI-generated deepfakes, and what legal principles

 DOI: https://10.5281/zenodo.22912311

are necessary to address the distinctive problems of mens rea, attribution and harm created by generative AI?

The article advances the argument that the principal difficulty is not simply that India lacks a single, dedicated offence labelled 'deepfake'. Existing criminal-law provisions may already address several forms of conduct associated with harmful synthetic media. The more fundamental problem lies in the fragmentation of legal responses and the uncertainty surrounding attribution of culpability across the deepfake lifecycle. A person who intentionally creates and disseminates a fraudulent deepfake is situated differently from a person who unknowingly forwards manipulated material, an intermediary that merely hosts user-generated content, or a developer whose general-purpose AI system is misused by an independent actor.

Accordingly, this article proposes that criminal responsibility for deepfake-related conduct should be analysed through a combination of human agency, intention, knowledge, control and harm. Such an approach preserves the foundational principle that criminal punishment should ordinarily correspond to individual culpability while recognising that generative AI creates new forms of technologically mediated conduct. The analysis therefore avoids treating all synthetic media as inherently criminal and instead examines how different forms of conduct and different mental states should influence the attribution of responsibility.

Methodologically, the article adopts a doctrinal and comparative legal approach. It examines the relevant Indian statutory framework, recent regulatory developments, emerging judicial responses and principles of criminal responsibility, while drawing comparatively upon developments in the European Union, United Kingdom and United States. The comparative analysis is not intended to advocate wholesale transplantation of foreign legislation; rather, it identifies regulatory principles that may assist in evaluating possible approaches within the Indian constitutional and criminal-law context.

The article makes three principal contributions. First, it shifts the analytical focus from the general regulation of deepfake technology to the more precise question of criminal attribution. Secondly, it examines the relationship between generative AI and traditional requirements of actus reus, mens rea, causation and proof, particularly where the precise output of an AI system may not have been anticipated by its user. Thirdly, it develops a differentiated framework for evaluating the responsibility of creators, users, disseminators and intermediaries according to their respective levels of agency, knowledge, control and contribution to legally cognisable harm.

2. Deepfakes And The Transformation Of Criminal Harm

The emergence of deepfake technology represents more than an incremental development in conventional digital manipulation. It marks a transformation in the manner in which false

 DOI: https://10.5281/zenodo.22912311

representations can be created, personalised and disseminated. Traditional manipulation ordinarily requires some alteration of pre-existing material; generative artificial intelligence, by contrast, can produce apparently authentic representations of persons, voices and events that may never have existed in the form depicted. The significance of this transformation for criminal law lies not merely in the falsity of the resulting content, but in the capacity of synthetic media to reproduce indicia of identity and authenticity at a scale and level of accessibility that complicate conventional assumptions concerning deception, authorship, consent and responsibility.

2.1 Conceptualising Deepfakes

The term 'deepfake' originally developed from the combination of 'deep learning' and 'fake', but its contemporary legal meaning extends beyond any single technical method of production. Earlier forms of deepfake generation were strongly associated with deep-learning architectures and techniques such as face-swapping. Contemporary generative AI, however, permits realistic synthetic content to be created through a wider range of models and processes. A legal definition tied too closely to a particular technological architecture therefore risks becoming obsolete as generative techniques evolve.

Recent regulatory approaches increasingly adopt technologically broader definitions. Under the European Union's Artificial Intelligence Act, a deepfake is understood in terms of AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and falsely appears authentic or truthful (European Union 2024). The European Commission's 2026 guidance further explains the concept through the cumulative ideas of resemblance, an existing or plausibly existing subject, and the false appearance of authenticity or truthfulness (European Commission 2026).

India has now moved toward a similarly functional approach, although it employs the broader expression 'synthetically generated information' (SGI). Following the 2026 amendments to the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, SGI encompasses specified audio, visual and audio-visual information that is artificially or algorithmically created, generated, modified or altered through a computer resource so as to appear real, authentic or true and to depict an individual or event in a manner likely to be perceived as indistinguishable from a natural person or real-world event (MeitY 2026). The official explanatory material identifies deepfakes, voice cloning and realistic AI-generated or AI-altered images and videos as examples capable of falling within the definition (MeitY 2026).

This development is important for two reasons. First, it indicates a movement away from regulating 'deepfakes' merely as a colloquial technological category and towards regulating the characteristics and effects of synthetic content. Secondly, it demonstrates that deception or realistic misrepresentation, rather than the use of one particular AI architecture, is becoming central to legal classification.

 DOI: https://10.5281/zenodo.22912311

The distinction is important for criminal law. If liability were attached merely to the fact that content was generated by AI, legitimate forms of synthetic media could unnecessarily fall within the criminal-law net. Conversely, a definition confined to a particular technical process could allow substantially identical harmful conduct to escape regulation merely because a different generative method was employed. A technologically neutral approach should therefore focus on the relationship between synthetic creation, representation, deception and legally cognisable harm.

2.2 From Digital Manipulation to Synthetic Reality

Digital manipulation is not a novel phenomenon. Photographs, recordings and videos have long been capable of alteration. What distinguishes contemporary generative AI is the combination of realism, accessibility, personalisation and scalability.

The development of generative systems has significantly reduced the technical effort required to create convincing synthetic media. Recent scholarship observes that forms of manipulated content that previously demanded substantial quantities of source material and specialised expertise can increasingly be generated from very limited source material (Diakopoulos 2025). This change alters the risk structure of digital falsification: a person's ordinary digital presence may supply material capable of being incorporated into synthetic representations without that person's participation or consent.

Deepfakes may take several forms. A person's face may be superimposed upon another body; an existing photograph may be synthetically altered; clothing may be digitally removed or replaced; a person's voice may be cloned; or an entirely artificial audio-visual sequence may be generated in which the targeted person appears to speak or act. Scholarship on non-consensual sexual deepfakes, for example, distinguishes between techniques involving face-swapping, synthetic undressing and the generation of new synthetic imagery (Gieseke et al. 2025).

The conceptual transition is therefore from the manipulation of evidence of reality to the generation of plausible synthetic reality. Criminal law has traditionally encountered forged documents, fabricated evidence, impersonation and fraudulent representations. Deepfakes do not necessarily create entirely unprecedented categories of wrongdoing. Their disruptive quality instead lies in their capacity to combine several familiar forms of wrongdoing through a technology capable of producing persuasive representations of identity and conduct.

This observation is important because technological novelty should not automatically be equated with legal novelty. The fact that an offence is committed using generative AI does not necessarily require an entirely new criminal offence. The relevant inquiry is whether the technology changes the constituent elements, scale, evidential characteristics or attribution of the underlying harm sufficiently to expose a gap in existing criminal law.

2.3 The Dual-Use Character of Synthetic Media

A coherent criminal-law analysis must begin by rejecting the proposition that deepfakes are inherently criminal. Synthetic media is a dual-use technology. The same generative capacities capable of facilitating deception can be employed for filmmaking, artistic expression, education, accessibility, satire, historical reconstruction and other legitimate purposes.

Contemporary regulatory frameworks recognise this distinction. The 2026 Indian framework expressly excludes certain routine or good-faith forms of editing and enhancement from its conception of SGI where they do not materially alter, distort or misrepresent the substance, context or meaning of the underlying content (MeitY 2026). Similarly, the EU AI Act recognises the relevance of artistic, creative, satirical and fictional contexts when applying transparency obligations to deepfakes (European Union 2024).

Accordingly, the relevant inquiry should not be:

Was artificial intelligence used to create the content?

Rather, it should ask:

What did the human actor do with the technology, with what mental state, with what degree of control, and with what legally relevant consequence?

This distinction provides the conceptual foundation for the differentiated model of criminal responsibility developed later in this article.

2.4 Deepfakes as Instruments of Criminal Harm

The criminal significance of deepfakes becomes clearer when synthetic media is considered as an instrument through which different forms of harm may be inflicted, rather than as an autonomous category of wrongdoing.

One category involves fraud and impersonation. Synthetic audio or visual representations may be used to create the appearance that another person has authorised a transaction, requested the transfer of money or communicated information that the person never provided. In such cases, the deepfake functions as the deceptive mechanism through which an established economic or identity-related harm is pursued.

A second category concerns non-consensual sexual synthetic media. This is particularly significant because the origins and development of deepfake technology have been closely associated with the creation of non-consensual sexual imagery. Recent scholarship treats such material as a form of image-based sexual abuse and emphasises that the harm can arise not only from dissemination but also from the non-consensual act of synthetic creation itself (McGlynn and Johnson 2025; Gieseke et al. 2025). The victim is represented as participating in intimate or sexual conduct that did not occur, producing harms relating to dignity, autonomy, privacy and reputation.

A third category involves extortion, coercion and harassment. Synthetic sexual or compromising material may be accompanied by threats of publication or used to compel a person to provide money, further imagery or other benefits. Here, the deepfake may function simultaneously as fabricated content and as an instrument of coercion.

A fourth category concerns reputational and identity-related harm. A synthetic recording may falsely portray a person making statements, endorsing products, engaging in unlawful conduct or participating in events. The harm may extend beyond conventional reputational injury where the synthetic representation appropriates a person's voice, face or other identifying characteristics for deceptive purposes.

Deepfakes may additionally be deployed to fabricate events or communications capable of producing wider social harms. Yet the existence of serious potential harm does not eliminate the need for careful differentiation. The criminal law should not treat a deliberately fabricated representation used for fraud, a non-consensual sexual deepfake, a disclosed satirical video and an unknowingly forwarded synthetic clip as equivalent merely because each involves AI-generated content.

2.5 Creation and Dissemination as Distinct Forms of Conduct

One of the most important distinctions for criminal responsibility is between creating harmful synthetic content and disseminating it.

Much legal attention has historically concentrated on publication or distribution because dissemination exposes the fabricated content to third parties and can substantially magnify harm. Recent scholarship, however, increasingly questions whether distribution should be regarded as the only legally significant event. McGlynn and Johnson (2025), examining sexually explicit deepfakes, argue that non-consensual creation itself can constitute a distinct form of harm and therefore warrants independent legal attention.

The distinction can be illustrated through three scenarios. In the first, A intentionally creates an intimate deepfake depicting B but never distributes it. In the second, A creates the content and sends it to thousands of users. In the third, C did not create the content but knowingly redistributes it for the purpose of humiliating B. Although each scenario concerns the same synthetic representation, the conduct, mental state and causal contribution of the actors differ.

A further distinction arises where a person disseminates content without knowing that it is synthetic. Criminalising such conduct merely because the content ultimately proves false would risk imposing liability without sufficient culpability. The problem becomes even more complex where automated recommendation systems contribute to dissemination without any individual human decision to transmit a particular item to a particular recipient.

The deepfake lifecycle should therefore be disaggregated into legally relevant stages:

creation → modification → possession → publication → dissemination → amplification → exploitation.

Responsibility should not automatically attach equally at every stage. Instead, the significance of each stage depends upon the applicable offence, the actor's mental state, the degree of control exercised and the resulting or intended harm. The following section therefore considers whether the existing Indian legal framework is capable of translating these technologically mediated harms into identifiable offences and legally sustainable forms of criminal responsibility.

3. The Attribution Problem: Who Committed The Crime?

The central difficulty presented by deepfakes is not necessarily the identification of a legally prohibited harm. Existing criminal law may already recognise conduct such as cheating, personation, identity theft, extortion, sexual abuse, defamation or unlawful dissemination. The more difficult question is attribution. Where artificial intelligence participates in producing the harmful representation, several persons may occupy different positions in the causal chain. One person may provide the source image, another may operate the generative system, a third may upload the resulting material, and an intermediary may facilitate its circulation. Criminal law must therefore determine which human contribution is sufficiently connected to the prohibited result to justify criminal responsibility.

3.1 Deepfake Lifecycle and Differentiated Human Responsibility

A deepfake may pass through several stages: **source acquisition → generation → modification → possession → publication → dissemination → amplification → exploitation**. Different actors may become involved at different stages. The person whose likeness is appropriated may have no involvement in its creation; the individual who generates the content may not be the person who distributes it; and a platform may facilitate circulation without participating in creation. Similarly, an AI provider may supply general-purpose technological infrastructure without knowing how a particular user will employ it.

The principal categories of actors may therefore be distinguished as follows:

Actor	Possible conduct	Central legal question
Subject/victim	Identity or likeness appropriated	What legal interest has been violated?

 DOI: https://10.5281/zenodo.22912311

AI user/creator	Generates synthetic content	What was the user's intention or knowledge?
Editor	Modifies or combines material	Did the person materially contribute to the unlawful representation?
Uploader/disseminator	Publishes or forwards content	Was dissemination intentional or knowing?
Platform/intermediary	Hosts, recommends or distributes content	What statutory duties and knowledge thresholds apply?
AI provider/developer	Supplies generative technology	Is there a sufficient causal and culpable connection to the specific harm?

The clearest case of attribution arises where an individual intentionally uses an AI system to create unlawful synthetic content. If A obtains B's photograph, generates an intimate deepfake of B and uses it to threaten B with publication, the AI system may have performed the computational generation, but the relevant human conduct includes acquiring the material, initiating the generation, selecting the target and employing the output for coercion. AI in such circumstances may function as an instrument through which the accused acts.

The distinction between creation and dissemination is equally important. A person who creates a deepfake but does not distribute it may occupy a different legal position from someone who knowingly uploads it to a public platform. The United Kingdom's approach to intimate-image abuse illustrates the value of focusing on the harmful conduct rather than the precise technical process through which the representation was created. Its statutory framework encompasses material that shows or appears to show a person, including manufactured or altered material commonly described as deepfakes (UK Government 2023). For India, this suggests that criminal provisions should, where appropriate, be technologically neutral and focus on the underlying harm rather than requiring proof that a particular AI technology generated the material.

3.2 Attribution, Mens Rea and Evidentiary Challenges

The use of AI should not, by itself, establish criminal liability. The prosecution must still prove the mental element required by the relevant offence. This becomes particularly significant because generative systems can produce outputs that are not entirely predictable. A user may enter a prompt without intending a particular harmful characteristic, or may receive an unexpected output and

immediately delete it without dissemination. These circumstances are materially different from deliberately instructing an AI system to produce a fabricated sexual video and then publishing it anonymously.

The relevant inquiry should therefore remain: what did the accused know, intend or foresee when undertaking the legally relevant conduct? The technological complexity of the process should not replace the ordinary inquiry into culpability. Criminal law must continue to distinguish intention, knowledge, recklessness and negligence rather than treating AI use itself as evidence of a sufficient mental state.

Attribution also presents a distinct evidentiary problem. Where a perpetrator uses an anonymous or pseudonymous account, establishing that a video is synthetic does not necessarily establish who created or disseminated it. The prosecution may need to establish a chain running from:

synthetic content → source account → technical identifier → device/user → human conduct → relevant mental state.

Each link may be contested. Multiple persons may have access to an account or device, while cloud-based AI services may hold relevant logs outside India. Cross-border access to data, preservation of digital evidence and cooperation between service providers and law-enforcement authorities may consequently become important to effective prosecution.

This distinction between authenticity and attribution is fundamental. Authenticity concerns whether the digital representation is genuine or manipulated; attribution concerns who created, controlled or disseminated it; and *mens rea* concerns whether that person's state of mind satisfies the relevant offence. These are separate legal questions and should not be collapsed into one technological inquiry.

3.3 Intermediaries, AI Providers and Technological Facilitation

Intermediaries present a different attribution problem because a platform may neither create the deepfake nor intend the resulting harm. Nevertheless, it may facilitate distribution, recommend the material or continue hosting it after receiving legally relevant notice. India's 2026 regulatory amendments are significant because they impose more specific obligations concerning synthetically generated information, thereby recognising the regulatory role of intermediaries in the detection, labelling and management of synthetic content (MeitY 2026).

Regulatory responsibility, however, is not synonymous with criminal responsibility. An intermediary's failure to comply with a statutory due-diligence obligation may attract regulatory consequences without establishing that the intermediary committed the underlying substantive offence represented by the deepfake. Criminal liability should therefore depend upon factors such as

actual knowledge, statutory notice, ability to control the material, failure to comply with a specific legal obligation, intentional facilitation, substantial contribution and the causal relationship between the intermediary's conduct and the resulting harm.

A related but more difficult question concerns AI developers and providers. Suppose a company develops a general-purpose image-generation model and an unrelated user employs it to create an unlawful deepfake. The mere fact that the technology was a condition in the causal chain should not ordinarily establish criminal culpability. Relevant questions include whether the system was specifically designed to facilitate prohibited conduct, whether the provider knew of substantial and specific misuse, whether misuse was intentionally facilitated, whether safeguards were deliberately disabled or ignored, and whether the provider materially contributed to the particular unlawful act.

The UK's Crime and Policing Act 2026 provides a useful comparative example by creating offences concerning the making, adapting, supplying or offering to supply tools capable of generating purported intimate images, while recognising a defence based on reasonable steps to prevent misuse (UK Parliament 2026). The approach distinguishes general technological provision from more culpable forms of technological facilitation. For India, this supports a distinction between general-purpose AI provision and intentional provision or operation of a tool for a prohibited purpose, rather than imposing automatic criminal liability on technology providers.

3.5 Human Agency–Intention–Knowledge–Control–Harm Framework

The preceding analysis suggests that an actor's criminal responsibility should not be determined merely by their position within the technological chain. A more precise framework can be constructed around five variables:

Human agency — Did the individual initiate, direct or materially participate in the conduct?

Intention — Was the relevant outcome intended?

Knowledge — Did the person know the nature of the content or the relevant circumstances?

Control — What degree of control did the person exercise over creation, publication or dissemination?

Harm — What legally cognisable harm was intended, foreseeable or actually produced?

These variables help differentiate the principal actors. An intentional creator will generally exhibit high agency, knowledge and control; a knowing disseminator may exercise less control over creation but significant control over circulation; an unknowing recipient ordinarily lacks the requisite knowledge and agency; and a general-purpose AI provider ordinarily has an indirect relationship with a particular unlawful output. Conversely, a provider of a tool specifically designed

 DOI: https://10.5281/zenodo.22912311

and supplied for prohibited use may possess a substantially stronger connection to the resulting offence.

The framework does not replace the constituent elements of individual offences. Rather, it provides a structured method for determining whether a person's relationship to synthetic content is sufficiently culpable to justify criminal attribution. It also demonstrates why a single technology-centred offence may be inadequate: the same deepfake may involve several actors whose contributions differ significantly in agency, knowledge, control and harm.

The central attribution principle is therefore that criminal responsibility should follow culpable human agency rather than technological proximity. A person should not become criminally responsible merely because an AI system they developed was subsequently misused by another individual. Equally, a person should not escape responsibility merely because an AI system performed the immediate computational act of generating unlawful content. The law must identify the point at which human conduct, culpable mental state and legally cognisable harm converge.

For Indian law, this principle has two implications. First, existing offences should be applied in a technologically neutral manner where their statutory ingredients are satisfied. Secondly, where existing offences fail to capture a distinct form of harm—particularly where creation itself constitutes the relevant wrong—legislative intervention may be justified on a harm-specific and *mens rea*-sensitive basis rather than through an indiscriminate prohibition of deepfakes.

4. Mens Rea In The Age Of Generative AI

4.1 Mens Rea and Technologically Mediated Conduct

The use of AI should not itself constitute evidence of criminal intention. AI may be used for artistic, educational, professional or entertainment purposes, and an output may sometimes contain characteristics not contemplated in the original prompt. The criminal-law inquiry must therefore remain focused on the mental state accompanying the legally significant conduct.

This distinction can be illustrated through three situations. First, a person deliberately instructs an AI system to create a non-consensual sexual representation of another individual and publishes it. Here, the use of AI is directly connected with the intended unlawful conduct. Secondly, a person creates a fictional character and the output unexpectedly resembles a real individual, without publication or knowledge of the resemblance. Similarity alone would provide a weak basis for criminal responsibility. Thirdly, a person knowingly receives a fabricated video depicting another individual committing a crime and deliberately circulates it to cause reputational harm. Here, the relevant mental state attaches principally to dissemination.

 DOI: https://10.5281/zenodo.22912311

4.2 Intention, Knowledge and the Unpredictability of AI

Intention represents the clearest form of culpability in deepfake-related conduct. Where a person deliberately employs generative AI to produce unlawful synthetic content for a particular purpose, the fact that the final pixels or audio were computationally generated does not necessarily weaken the inference of human agency. For example, deliberately uploading an individual's photograph into a system with the purpose of producing a non-consensual intimate representation demonstrates intentional initiation of the process, even though the precise characteristics of the final image are computationally generated.

Nevertheless, intending to initiate a generation process is not necessarily equivalent to intending every characteristic of its output. Generative systems may produce variations that the user did not anticipate. Criminal responsibility should therefore depend upon the mental element required by the particular offence rather than an assumption that the accused intended every computational consequence.

Knowledge assumes particular importance where the accused did not create the deepfake but subsequently disseminated it. A person who genuinely believes apparently authentic material to be genuine is situated differently from a person who knows it is fabricated and deliberately distributes it. The distinction may therefore be drawn between creation with knowledge and unlawful purpose, receipt without knowledge, and knowledge of falsity followed by deliberate dissemination.

Generative AI also complicates the relationship between intention and outcome. Users can exercise substantial control through prompts, parameters, selection and repeated generation, but the precise characteristics of an output may not always be anticipated. AI unpredictability should therefore neither automatically eliminate responsibility nor automatically establish it. If a person deliberately attempts to generate harmful material and the output differs only in minor details, that difference may have little relevance. Conversely, if a person intends a harmless fictional representation and the system unexpectedly produces a harmful representation of an identifiable individual, the absence of relevant knowledge or intention may become significant.

4.3 Recklessness, Negligence and the Risk of Overcriminalisation

Whether recklessness should constitute sufficient *mens rea* for particular deepfake offences presents a more difficult question. An individual may recognise a substantial risk that material has been manipulated and nevertheless deliberately circulate it. A narrowly defined recklessness standard could address such conduct where the relevant offence permits it.

However, deepfake detection is increasingly difficult. Extending recklessness too broadly could expose ordinary users to criminal liability for failing to identify sophisticated synthetic

 DOI: https://10.5281/zenodo.22912311

material. This concern becomes particularly significant where content is circulated rapidly through social-media and messaging platforms and ordinary users lack reliable technical means of authentication.

Negligence raises an even greater concern. Criminalising every failure to verify online material could transform ordinary digital carelessness into criminal conduct. A negligence-based offence should therefore be confined to circumstances involving a clearly defined legal duty, serious foreseeable harm and a substantial departure from the applicable standard of care.

This is especially important in relation to non-consensual intimate deepfakes. Recent United Kingdom legislation demonstrates that specific categories of synthetic sexual imagery may justify targeted criminalisation at the point of creation, rather than requiring proof of subsequent dissemination (UK Parliament 2025; UK Parliament 2026). The comparative development raises an important question for India: whether existing offences adequately capture the wrong involved when a person intentionally creates a realistic but entirely synthetic intimate representation of an identifiable individual.

A future Indian offence in this area should clearly identify factors such as the accused's knowledge of the person's identity, absence of consent, intention to create the representation and the intended purpose or use. The artificial nature of the representation should not obscure the potential harm to privacy, dignity and sexual autonomy.

4.4 Mens Rea, Control and Technological Facilitation

Control provides an important bridge between intention and technological uncertainty. The more control an individual exercises over the generation and dissemination of a deepfake, the stronger the potential basis for attributing the conduct to that individual. Relevant indicators may include selection of the victim or source material, specificity of prompts, repeated attempts to obtain a desired output, editing, removal of identifying markers, deliberate publication, amplification and attempts to conceal identity.

Control should not, however, become an independent basis for criminal liability. It should operate principally as circumstantial evidence relevant to intention and knowledge. A person who repeatedly modifies prompts to obtain a particular non-consensual image and then deliberately publishes it occupies a substantially different position from someone who receives an unexpected AI output.

The same principle becomes important when considering AI providers. Knowledge that an AI system can generate harmful material does not necessarily establish criminal intent concerning every unlawful output. A general-purpose AI system may have extensive legitimate uses, and criminal responsibility for a user's conduct should not automatically be transferred to the provider.

A stronger basis for provider responsibility may arise where a provider deliberately designs, adapts or supplies a tool for a prohibited purpose, knowingly facilitates a specific unlawful activity, or otherwise satisfies the elements of a recognised offence. The UK's 2026 approach to tools capable of generating purported intimate images illustrates such targeted technological facilitation, while preserving a reasonable-steps defence in specified circumstances (UK Parliament 2026).

4.5 Evidentiary Proof and a Human-Centred Model of AI Criminal Responsibility

Generative AI creates a further evidentiary challenge because synthetic material can be disseminated at extraordinary speed. A person may create a deepfake and upload it publicly, after which algorithms and users may distribute it to a much larger audience. The eventual scale of circulation should not automatically establish that the creator intended every subsequent consequence. The distinction between intended, foreseeable and unforeseeable downstream harm therefore remains important.

Mens rea will frequently have to be established through circumstantial evidence. Relevant material may include prompts and generation logs, saved files, editing history, communications, account activity, timing of creation and dissemination, financial transactions and attempts to delete or conceal evidence.

The growing availability of digital provenance information may assist investigations, but technological evidence must still be reliably connected to the accused. A prompt does not necessarily establish who entered it, and the presence of a file on a device does not necessarily establish who created it. The evidentiary challenge is therefore not merely to establish what the AI generated, but what the accused knew and intended.

Any future Indian deepfake offence should accordingly specify its mental element. It should distinguish between intentional creation, knowing dissemination, recklessness where justified by the seriousness of the harm, and inadvertent transmission. The more severe the criminal consequence, the stronger the justification for requiring a clear and culpable mental state.

The analysis ultimately supports a human-centred model of AI criminal responsibility. AI should be treated as a technological instrument within criminal-law analysis rather than as a substitute for human culpability. The decisive questions should be:

These questions preserve established principles of criminal responsibility while adapting them to technologically mediated conduct. The challenge is not to create an entirely new theory of *mens rea* for artificial intelligence, but to determine how established principles of culpability should operate where AI mediates the relationship between human action and harmful outcome.

5. Deepfakes, Harm And Differentiated Criminal Liability

 DOI: https://10.5281/zenodo.22912311

5.1 Why Harm Matters in Criminalisation

The criminal-law response to deepfakes should not be determined solely by the fact that synthetic content has been created. Criminal law traditionally distinguishes between conduct according to the nature of the harm caused, the interests protected by the offence, the accused's mental state and the causal connection between the conduct and the resulting harm. The same technological process may therefore produce substantially different legal consequences depending upon its purpose, context and effects.

Generating a fictional video for entertainment, creating a manipulated image to defraud another person and producing a non-consensual sexual deepfake may involve technically similar processes but represent fundamentally different forms of conduct. A technologically defined offence that criminalises synthetic content merely because it is capable of appearing authentic could consequently capture conduct that lacks the culpability ordinarily required for criminal responsibility.

The constitutional dimension is equally important. In *Justice K.S. Puttaswamy (Retd.) v Union of India*, the Supreme Court recognised privacy as a constitutionally protected right closely connected with dignity, autonomy and personal liberty (Justice K.S. Puttaswamy (Retd.) v Union of India, 2017). The judgment's recognition of informational and decisional privacy provides an important constitutional foundation for understanding why unauthorised synthetic representations of an identifiable individual may implicate interests extending beyond reputation alone.

At the same time, *Puttaswamy* does not establish that every unauthorised use of personal information is criminal. Its significance for deepfake regulation lies in identifying the constitutional interests that may justify carefully defined legal restrictions while retaining the requirements of legality, legitimate purpose and proportionality.

5.2 Non-Consensual Sexual Deepfakes

Non-consensual intimate imagery represents one of the clearest examples of why harm-sensitive criminalisation is necessary. AI-generated intimate imagery involving an identifiable individual may interfere with sexual autonomy, privacy, dignity and reputation even where the depicted sexual conduct never occurred.

Indian law already contains provisions capable of addressing sexually explicit electronic material through sections 67 and 67A of the Information Technology Act, 2000. Depending upon the facts, provisions relating to privacy, personation, cheating, criminal intimidation, extortion and defamation may also become relevant.

 DOI: https://10.5281/zenodo.22912311

The doctrinal difficulty arises because a synthetic image may falsely depict a person in a sexual situation without any underlying real-world sexual event. The criminal law must therefore determine whether the protected interest is limited to the authenticity of the depicted event or extends to the unauthorised creation and dissemination of the representation itself.

The broader constitutional jurisprudence on dignity and privacy supports the latter inquiry. In *Puttaswamy*, the Supreme Court recognised privacy as encompassing bodily integrity, personal autonomy and control over aspects of one's personal identity (Justice K.S. Puttaswamy (Retd.) v Union of India, 2017). A deepfake involving the unauthorised sexual representation of an identifiable person can consequently be analysed not merely as a false statement but as an interference with personal autonomy and dignity.

Comparative developments reinforce this approach. The United Kingdom's Crime and Policing Act 2026 creates offences concerning purported intimate images and also addresses the making, adaptation, supply or offering of tools capable of generating such material. The legislative approach is significant because it recognises that culpable conduct may occur at several stages of the technological lifecycle.

For India, however, such an approach should be adapted rather than mechanically transplanted. A future offence should distinguish between consensual synthetic sexual content, fictional content and synthetic representations of identifiable persons created or disseminated without consent. The accused's intention or knowledge should remain central to criminal responsibility.

5.3 Fraud and Financial Deception

Deepfakes can transform conventional fraud by making impersonation more convincing. AI-generated voices, videos and images can be used to impersonate relatives, executives, business partners or other trusted persons, potentially facilitating unauthorised transfers and fraudulent instructions.

Indian criminal law already contains provisions capable of addressing several such situations. Section 319 of the Bharatiya Nyaya Sanhita, 2023 concerns cheating by personation, while section 318 addresses cheating more generally. Section 66D of the Information Technology Act, 2000 separately addresses cheating by personation through a communication device or computer resource. The importance of *deepfake* technology in such cases therefore lies primarily in its evidentiary and attribution consequences. The prosecution must establish not simply that a synthetic representation was used, but that the accused intentionally or knowingly employed it as part of the deceptive conduct.

 DOI: https://10.5281/zenodo.22912311

The evidentiary chain may include prompts, generation histories, account information, payment records, communications, device data, transaction records and subsequent efforts to conceal the conduct. The existence of a convincing deepfake is therefore only one component of the prosecution's case.

5.4 Identity Theft and Impersonation

Identity-based harms occupy an intermediate position between fraud and reputational injury. A deepfake may falsely represent an individual as having spoken, acted or participated in conduct that never occurred. The resulting harm may include unauthorised use of identity, loss of professional reputation, disruption of personal relationships or manipulation of third parties.

Section 66C of the Information Technology Act, 2000 addresses identity theft, while section 66D addresses cheating by personation through a communication device or computer resource. Their applicability depends upon the facts and statutory elements of the particular case.

The Delhi High Court's decision in *Dr Aniruddha Dhairyadhar Joshi through Power of Attorney Holder v John Does Ashok Kumars & Ors* is particularly significant in this context. In February 2026, the Court considered allegations concerning AI-generated deepfake videos, voice-cloned material and fabricated representations purporting to depict the plaintiff. The Court granted an ex parte interim injunction and directed platforms including Google, Meta and X to remove specified infringing content and similar material, subject to the terms of the order. It also directed disclosure of available basic subscriber information connected with identified URLs, accounts and handles, thereby demonstrating the relationship between content removal and the identification of anonymous actors. (*Dr Aniruddha Dhairyadhar Joshi through Power of Attorney Holder v John Does Ashok Kumars & Ors*, 2026).

The case is particularly valuable for the present analysis because it demonstrates that deepfake disputes can require courts to address several distinct questions simultaneously: whether the content unlawfully appropriates a person's identity; whether the platforms should remove it; how similar or derivative content should be treated; and how the underlying anonymous actors may be identified.

Importantly, however, *Joshi* was a civil/commercial proceeding concerning personality, publicity and related rights. It should not be treated as establishing criminal liability for deepfake creation. Its significance for criminal law lies instead in demonstrating the evidentiary and attribution problems that criminal investigators may encounter when the persons responsible for synthetic content are anonymous.

5.5 Defamation and Reputational Harm

Deepfakes can produce reputational injury by falsely portraying individuals as engaging in criminal, immoral, sexually explicit or professionally damaging conduct. Section 356 of the Bharatiya Nyaya Sanhita, 2023 provides the contemporary statutory framework for criminal defamation.

The constitutional status of reputation was considered by the Supreme Court in *Subramanian Swamy v Union of India*, where the Court upheld the constitutional validity of criminal defamation and recognised reputation as an interest protected within the constitutional framework (Subramanian Swamy v Union of India, 2016).

For deepfake cases, however, the existence of reputational harm and the attribution of that harm to a particular accused must remain analytically distinct. A fabricated video may unquestionably damage a person's reputation, but the prosecution must still establish the accused's participation and the mental element required by the offence.

5.6 Extortion, Coercion and Harassment

The capacity to manufacture apparently authentic evidence creates additional opportunities for extortion and coercion. An offender may threaten to publish a fabricated sexual image, falsely incriminating recording or manipulated audio unless the victim pays money, provides information or complies with a demand.

In such circumstances, the deepfake may function as an instrumentality of another offence rather than as the sole object of criminalisation. Depending upon the facts, provisions concerning criminal intimidation, extortion, cheating or related offences may become applicable.

The distinction is doctrinally important. If an accused generates a fabricated intimate image and then threatens its publication unless the victim pays money, the legal analysis should not stop with the question whether the image itself is unlawful. The prosecution should examine the entire sequence: creation, possession, threat, demand, communication and intended or actual dissemination.

5.7 Public Deception and Wider Social Harm

Deepfakes can also generate harms extending beyond an identifiable individual. Synthetic audio or video may falsely depict an event, institution or public figure and thereby contribute to widespread deception.

5.8 Proportionality and Freedom of Expression

A harm-sensitive approach must operate within constitutional limitations. Article 19(1)(a) protects freedom of speech and expression, subject to the restrictions authorised by Article 19(2). Any

criminal regulation of synthetic media must therefore satisfy requirements of legality, legitimate purpose and proportionality.

5.9 A Harm-Sensitive Model for India

The Indian framework should therefore be harm-sensitive, offence-specific and mens rea-sensitive. Rather than treating every deepfake as one legal category, criminal responsibility should be analysed through the interaction of five variables: the nature of the representation, the protected interest affected, the accused's mental state, the accused's degree of control over dissemination and the resulting or intended harm.

The emerging Indian jurisprudence demonstrates why such differentiation is necessary. *Puttaswamy* provides the constitutional foundation for privacy, autonomy and dignity; *Subramanian Swamy* demonstrates the legal significance of reputation; *Shreya Singhal* establishes the importance of precision when regulating online expression; *Arjun Panditrao Khotkar* highlights the evidentiary importance of authenticity and provenance; and *Joshi* demonstrates the practical difficulties of identifying and restraining anonymous actors responsible for AI-generated impersonation. These cases arise in different doctrinal contexts and should not be treated as establishing a single deepfake rule. Their combined significance lies in the legal principles they contribute to different components of the problem.

The deepfake lifecycle consequently requires differentiated treatment. Intentional creation of a non-consensual intimate image may justify direct criminal liability where the statutory elements are satisfied. Knowingly disseminating such material should be analysed separately from mere receipt. Using synthetic content to obtain money may primarily engage cheating or personation provisions. False synthetic representations causing reputational injury may engage defamation or other applicable offences. Providers of specialised tools should not be treated as automatically criminally liable merely because their technology can be misused; liability should depend upon a demonstrable legal connection between their conduct, knowledge, control and the prohibited use.

The central principle is therefore that criminal liability should follow culpable human agency, not technological proximity.

A coherent Indian framework should consequently avoid two opposite errors. The first is technological exceptionalism—the assumption that AI-generated conduct necessarily requires an entirely separate criminal-law architecture. The second is technological neutrality without adaptation—the assumption that existing offences can always be applied without addressing the new evidentiary and attribution problems created by generative AI.

The preferable approach lies between these extremes. Existing offences should remain available where their elements are satisfied, while narrowly defined statutory provisions may be introduced

where existing law leaves demonstrable gaps, particularly concerning non-consensual synthetic intimate imagery, AI-mediated impersonation and the intentional dissemination of materially deceptive synthetic content.

Such a framework would allow Indian criminal law to respond to genuine harms caused by deepfakes while preserving the foundational requirements of mens rea, individual responsibility, causation, evidentiary reliability and proportionality.

6. Comparative International Perspectives On Deepfake Criminal Responsibility

The transnational nature of generative artificial intelligence makes comparative analysis particularly relevant to the development of an Indian framework for deepfake-related criminal responsibility. Synthetic content may be generated in one jurisdiction, uploaded through infrastructure located in another and disseminated globally within seconds. Comparative law should not, however, be used simply to transplant foreign statutes into India. Differences in constitutional structures, criminal-law principles, intermediary frameworks and institutional capacities require attention to the underlying principles governing attribution, mens rea, harm, intermediary responsibility and technological facilitation.

Three broad approaches are particularly relevant. The European Union has principally adopted a transparency and risk-governance model; the United Kingdom has moved towards targeted criminalisation of particular harmful synthetic conduct; and the United States operates through a comparatively fragmented combination of federal, state and constitutional mechanisms. India's emerging framework contains elements of each approach but has not yet consolidated them into a coherent model of criminal responsibility.

6.1 Comparative Approaches: EU, UK and US

The European Union's Artificial Intelligence Act represents a regulatory rather than conventional criminal-law approach. Article 50 establishes transparency obligations concerning specified AI-generated or manipulated content, including requirements relating to the identification of certain deepfakes (European Union 2024). The significance of this approach for India lies particularly in its emphasis on transparency and provenance. Reliable information concerning the generation or modification of synthetic material may assist investigators in reconstructing its technological history.

Provenance, however, cannot by itself establish criminal responsibility. A watermark or metadata record may indicate that content was generated through a particular system without establishing who operated it, why it was generated or whether the user possessed the requisite *mens*

 DOI: https://10.5281/zenodo.22912311

rea. Thus, provenance may establish technological origin; criminal liability requires proof of culpable human conduct. Transparency mechanisms should consequently operate as evidentiary infrastructure rather than substitutes for proof of attribution and mental state.

The European approach also illustrates the importance of distinguishing regulatory obligations from criminal responsibility. Duties imposed upon providers and deployers concerning transparency do not automatically make them criminally responsible for unlawful downstream uses. At the same time, European human-rights principles demonstrate the need to balance protection against synthetic-media harms with freedom of expression. Restrictions affecting satire, artistic expression, journalism and political communication must satisfy requirements of legality and proportionality. The reasoning in *Shreya Singhal v Union of India* similarly demonstrates the importance of precision when restricting online expression, particularly where broad concepts such as "false", "misleading" or "deceptive" could otherwise produce uncertainty (Shreya Singhal v Union of India, 2015).

The United Kingdom has adopted a more targeted criminal-law approach. The Data (Use and Access) Act 2025 introduced offences concerning the intentional creation of, or requesting the creation of, purported intimate images of adults without consent or reasonable belief in consent. The provisions are capable of covering representations that appear to depict an individual but are not authentic photographs or films of that person (Data (Use and Access) Act 2025). The provisions came into force on 6 February 2026 and are significant because they recognise that, for particular forms of serious synthetic abuse, creation itself may constitute the legally relevant harm, rather than liability arising only from subsequent dissemination.

The UK's Crime and Policing Act 2026 moves further upstream by addressing technological facilitation. Section 99 concerns the making, adapting, supplying or offering to supply things for use as generators of purported intimate images, while providing a defence where reasonable steps were taken to prevent misuse (Crime and Policing Act 2026). This does not establish automatic criminal responsibility for AI developers. Instead, it targets a narrower category of technological facilitation and introduces a safeguards-based distinction between culpable provision and ordinary technological capability.

The United Kingdom's Online Safety Act 2023 further demonstrates the distinction between substantive criminal responsibility and intermediary regulation. Platforms may have content-governance obligations without becoming the criminal authors of every unlawful item uploaded by users. This distinction is particularly relevant to India, where increasingly specific intermediary obligations concerning synthetically generated information should not be converted into automatic criminal liability for platforms.

The United States presents a substantially different model. Deepfake regulation is distributed among federal legislation, state criminal and civil laws, sector-specific regulation and

constitutional limitations arising from the First Amendment. Fraudulent impersonation, non-consensual intimate imagery and election-related synthetic media may therefore attract different legal responses. The US experience illustrates both the value and limitations of a fragmented model: legislative breadth does not necessarily produce comprehensiveness, while excessive fragmentation may create enforcement gaps. A single offence encompassing every form of "deepfake" could also create problems of vagueness and overcriminalisation.

6.2 Attribution, Mens Rea and Technological Facilitation

The comparative approaches reveal a common principle: technological involvement should not, by itself, determine criminal responsibility. The relevant inquiry remains whether the actor exercised sufficient agency and possessed the mental state required by the particular offence.

The UK approach demonstrates that targeted criminalisation can extend beyond dissemination to creation and, in narrowly defined circumstances, technological facilitation. The distinction between a person who deliberately creates a prohibited synthetic image and a general-purpose technology provider whose system is subsequently misused is therefore significant. Factors such as intention, knowledge, degree of control, specific contribution and the foreseeability of the prohibited use provide a more defensible basis for attribution than technological proximity alone.

These approaches support the Human Agency–Intention–Knowledge–Control–Harm framework proposed in this article. Human agency identifies the legally relevant actor; intention and knowledge establish the mental connection to the prohibited conduct; control assists in determining the degree of participation; and harm ensures that criminal law remains directed towards legally protected interests rather than syntheticity itself.

6.3 Indian Jurisprudence in Comparative Perspective

Recent Indian judicial developments demonstrate that Indian courts are already confronting many of the problems addressed through foreign regulatory and criminal-law models. In *Dr Aniruddha Dhairyadhar Joshi through Power of Attorney Holder v John Does Ashok Kumars & Ors* (2026), the Delhi High Court dealt with alleged AI-generated deepfake videos and voice-cloned content, granted interim protection against specified unauthorised material and directed relevant platforms to remove identified content and disclose available subscriber information. The case is particularly significant for the attribution problem because the alleged creators were unidentified John Does.

Other recent proceedings similarly demonstrate the expanding judicial encounter with synthetic-media harms. In *Raghav Chadha v Ashok Kumar John Doe & Ors* (2026), the Delhi High

Court considered AI-generated deepfake videos and manipulated images in the context of dignity and freedom of speech. *Ravi Shankar v John Doe(s)/Ashok Kumar(s) & Ors*(2025) likewise concerned alleged AI-generated content, deepfake material and voice cloning, with the Court addressing unauthorised use of personality attributes.

The significance of these cases should nevertheless be stated carefully. Much of the emerging Indian deepfake jurisprudence is civil or commercial rather than criminal. These proceedings demonstrate the reality of harm, the importance of takedown mechanisms, personality protection and the identification of anonymous actors, but they do not themselves establish a general criminal offence of deepfake creation. They instead demonstrate the extent to which courts are currently relying upon existing civil, personality-rights, copyright and intermediary-law doctrines.

7. Conclusion

Generative artificial intelligence has altered the relationship between human conduct, digital representation and criminal harm. Deepfakes demonstrate this transformation particularly clearly: fabricated images, videos and voices can facilitate fraud, impersonation, reputational injury, sexual exploitation and violations of privacy. Yet technological sophistication does not itself determine criminal responsibility. The central legal question remains whether a particular human actor engaged in conduct satisfying the elements of an offence with the requisite mental state.

This article has demonstrated that India's existing criminal-law framework is capable of addressing several forms of deepfake-related harm. The Bharatiya Nyaya Sanhita 2023, Information Technology Act 2000, Bharatiya Sakshya Adhiniyam 2023 and intermediary framework provide legal mechanisms relevant to cheating, personation, identity theft, defamation, privacy violations, sexually explicit electronic material and electronic evidence. The principal difficulty is therefore not simply the absence of applicable law, but the attribution of responsibility where synthetic content is created and disseminated through multiple actors, technological systems and anonymous digital identities.

The 2026 amendments to the Information Technology Rules, which recognise synthetically generated information as a specific regulatory category, represent an important development. Their provisions concerning prevention, labelling and technical provenance can strengthen intermediary accountability and evidence preservation (Ministry of Electronics and Information Technology 2026). They do not, however, create a general criminal offence of deepfake creation. Regulatory responsibility must therefore remain analytically distinct from criminal authorship.

The comparative analysis similarly demonstrates that no single international model has emerged. The European Union has emphasised transparency, provenance and accountability; the United Kingdom has adopted targeted criminal offences concerning particular forms of purported intimate

imagery and technological facilitation; and the United States continues to operate through a more fragmented combination of federal and state mechanisms. These approaches suggest that India need not choose between complete technological neutrality and an all-encompassing deepfake offence. A more proportionate approach can combine existing criminal offences with targeted intervention where identifiable harms remain insufficiently addressed.

The principal contribution of this article is the Human Agency–Intention–Knowledge–Control–Harm framework. Its central proposition is that criminal responsibility should follow culpable human agency rather than technological proximity. Human agency identifies the legally relevant actor; intention and knowledge determine the mental connection to the prohibited conduct; control assists in assessing participation across the deepfake lifecycle; and harm ensures that criminalisation remains directed towards legally protected interests rather than syntheticity itself.

The framework also distinguishes between situations that can already be addressed through existing offences and genuine legislative gaps. Where AI is employed to commit cheating, personation, identity theft, intimidation or another established offence, the technological method should not prevent application of existing law. Where existing provisions fail to capture a clearly defined and serious harm, particularly certain forms of non-consensual synthetic intimate imagery, targeted legislative intervention may be considered. Such intervention should clearly define prohibited conduct, specify the requisite *mens rea*, distinguish creation from dissemination, preserve legitimate expression such as satire and artistic activity, and avoid automatic criminal liability for intermediaries or general-purpose AI providers.

Finally, constitutional safeguards must constrain the development of deepfake regulation. Privacy, dignity, reputation and freedom of expression may all be implicated by synthetic media. The principles reflected in *Justice K.S. Puttaswamy (Retd.) v Union of India* (2017), *Shreya Singhal v Union of India* (2015) and *Subramanian Swamy v Union of India* (2016) reinforce the importance of legality, precision and proportionality when criminal law affects protected interests and expression. Deepfake regulation should consequently address demonstrable harm rather than treat falsity or technological manipulation, standing alone, as criminal conduct.

Generative AI therefore does not make traditional criminal-law principles obsolete. It makes their careful application more important. Mens rea, individual responsibility, attribution, causation and proportionality remain capable of governing technologically mediated conduct, but must be applied with greater sensitivity to the distributed nature of digital creation and dissemination.

The governing principle can therefore be stated simply:

Criminal responsibility should follow culpable human agency, not technological proximity.

A coherent Indian response to deepfakes should consequently operate through the coordinated application of substantive criminal law, evidentiary mechanisms, intermediary regulation and

constitutional safeguards. The challenge is not that machines have replaced human responsibility, but that technology has made human responsibility more difficult to identify. Contemporary criminal law must make that responsibility legally traceable without sacrificing the principles that legitimately constrain the criminal sanction.

References

Books and Journal Articles

Chesney RM and Citron DK (2019) 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' 107 *California Law Review* 1753–1819. https://doi.org/10.15779/Z38RV0D15J

Lessig L (1999) *Code and Other Laws of Cyberspace*. New York: Basic Books. Harvard Law School – Code and Other Laws of Cyberspace

Indian Cases

Arjun Panditrao Khotkar v Kailash Kushanrao Gorantyal (2020) 7 SCC 1.
Indian Kanoon – Arjun Panditrao Khotkar v Kailash Kushanrao Gorantyal

Dr Aniruddha Dhairyadhar Joshi through Power of Attorney Holder v John Does Ashok Kumars & Ors (2026) CS(COMM) 178/2026, Delhi High Court, order dated 24 February 2026. Judgment / Order

Jaikishan Kakubhai Saraf alias Jackie Shroff v The Peppy Store & Ors (2024) 2024:DHC:4046, CS(COMM) 389/2024, Delhi High Court, order dated 15 May 2024. Judgment / Order

Justice KS Puttaswamy (Retd) v Union of India (2017) 10 SCC 1.
Supreme Court of India

Raghav Chadha v Ashok Kumar John Doe & Ors (2026) CS(OS) 466/2026, Delhi High Court. Delhi High Court Order – 21 May 2026

Ravi Shankar v John Doe(s)/Ashok Kumar(s) & Ors (2025) CS(COMM) 889/2025, Delhi High Court, order dated 26 August 2025. Judgment / Order

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

Subramanian Swamy v Union of India (2016) 7 SCC 221.
Supreme Court of India

Indian Legislation and Government Materials

Government of India (2000) *Information Technology Act 2000*. Act No 21 of 2000. India Code – Information Technology Act 2000

Government of India (2023a) *Bharatiya Nyaya Sanhita 2023*. Act No 45 of 2023. India Code – Bharatiya Nyaya Sanhita 2023

Government of India (2023b) *Bharatiya Sakshya Adhiniyam 2023*. Act No 47 of 2023. India Code – Bharatiya Sakshya Adhiniyam 2023

Government of India (2026a) *Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules 2026*. Gazette of India, GSR 120(E), 10 February 2026. MeitY – 2026 IT Rules Amendment

Government of India (2026b) *Frequently Asked Questions: Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026*. MeitY – 2026 IT Rules FAQs

European Union

European Commission (2026) *Transparency Obligations under Article 50 of the AI Act*. European Commission – Article 50 AI Act

European Union (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), OJ L 2024/1689. EUR-Lex – Artificial Intelligence Act

United Kingdom

United Kingdom (2023) *Online Safety Act 2023*, c 50. UK Government – Online Safety Act 2023

United Kingdom (2025) *Data (Use and Access) Act 2025*, c 18. UK Government – Data (Use and Access) Act 2025

United Kingdom (2026) *Crime and Policing Act 2026*, c 20. UK Government – Crime and Policing Act 2026

 DOI: https://10.5281/zenodo.22912311

International Instruments and Standards

Council of Europe (2024) *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law.* CETS No 225. Council of Europe – Framework Convention on AI

OECD (2019) *Recommendation of the Council on Artificial Intelligence.* OECD/LEGAL/0449. OECD AI Principles

OECD (2024) *Recommendation of the Council on Artificial Intelligence: 2024 Updated Version.* OECD AI Principles – Updated Version

UNESCO (2021) *Recommendation on the Ethics of Artificial Intelligence.* UNESCO – Recommendation on the Ethics of AI

Publisher's Note

Oxford Law Review remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Oxford Law Review or its licensors retain the rights to this article pursuant to the applicable publishing agreement with the author(s) or other rightsholder(s). Any self-archiving, reproduction, distribution, or reuse of the accepted manuscript or published version is subject to the terms of the applicable publishing agreement and applicable law.