INDIA'S DEEPFAKE BLIND SPOT: THE CASE AGAINST A DOWNSTREAM-ONLY CRIMINAL FRAMEWORK

Akanksha Singh, LL.M., Dharmashastra National Law University, Jabalpur, M.P.

ABSTRACT

The rapid proliferation of deepfake technology has emerged as one of the most formidable challenges confronting contemporary legal systems. Characterised by an unprecedented degree of realism, near-zero creation costs, and exponential scalability, deepfakes function not merely as a novel category of digital harm but as a multiplier technology, amplifying traditional offences such as fraud, defamation, impersonation, and non-consensual pornography to a qualitatively new level of severity. India's existing statutory framework, comprising the Bharatiya Nyaya Sanhita, 2023, the Information Technology Act, 2000, and the Information Technology Rules, 2021, addresses deepfake-related harms only incidentally, responding to their downstream consequences such as defamation, obscenity, and identity theft, while remaining conspicuously silent on the upstream conduct of unlawful creation and dissemination. Even the landmark 2026 amendments to the IT Rules, which introduced definitional clarity for synthetic media and strengthened intermediary obligations, operate within a predominantly reactive and harm-remedial paradigm. This downstream-centric approach is structurally inadequate, as deepfakes inflict irreversible damage within moments of dissemination, leaving little room for preventive intervention.

This article argues that deepfakes possess sufficiently distinct technological characteristics such as realism, virality, scalability, and persuasiveness, warranting an independent statutory-penal recognition under Indian law. This article advocates for the insertion of an upstream-focused criminal provision that penalises the intentional creation and distribution of harmful deepfake content as an independent offence, alongside targeted regulatory reforms to expand intermediary accountability.

I. INTRODUCTION TO DEEPFAKE TECHNOLOGY, TYPES AND ITS MULTIPLIER NATURE.

Deepfake refers to synthetic or manipulated media which includes images, audio, video or multimodal combinations thereof, which uses a Machine Learning model to create realistic portrayals of events, actions, expressions or speech that did not actually occur.[1] Such portrayals can be through visual deepfakes, audio deepfakes or multimodal deepfakes, which is nothing but various type of deepfake manifestations.

In visual deepfakes, the technology provides features like face swapping, lip synchronization, facial re-enactment or expression transfer.[2] In audio deepfake, the tool enables one to clone voice, synthesis text-to-speech or audio driven mouth animation. Lastly, the multimodal deepfakes provide a combination of audio-visual features thereby increasing the believability or persuasiveness in a synthetic content. With such varied modals, deepfake technology helps users to make a hyper-realistic content simply by putting audio-imagery dataset in the deepfake software, which is easily available on an open-source software. As a result, the technology as a content creation tool proves to be a double-edged sword, where on one hand it can be used as a tool for innovation and creativity and on the other hand a tool to commit criminal offences such as defamation, impersonation, fraud, non-consensual sexual images etc.

However, if we further dig into deepfake features, it displays distinct technological characteristic, which demands a differential treatment for solving this menace. Firstly, deepfakes possess a high degree of realism, making content significantly more persuasive than traditional falsehood. Studies indicate that approximately 68% of deepfakes are now "nearly indistinguishable from genuine media", and nearly 70% of individuals lack confidence in distinguishing real from synthetic media[3], highlighting the deceptive capacity of deepfakes. Secondly, Deepfakes are highly scalable and easily reproducible, causing widespread harm with minimal resources. The cost of creating a deepfake can be as low as a few dollars where the technology requires only basic tools and short audio video samples to generate a convincing output.[4] As a result, the volume of deepfake content has surged dramatically from 500,000 in

[1] Fakhar Abbas & Araz Taeihagh, *Unmasking Deepfakes: A Systematic Review of Deepfake Detection and Generation Techniques Using Artificial Intelligence*, SCIENCEDIRECT (2024).
[2] Justus Thies et al., *Face2Face: Real-time Face Capture and Reenactment of RGB Videos*, ARXIV EPRINT ARCHIVE (2016), https://arxiv.org/abs/1703.08971 (Last visited 11 January 2026).
[3] *Deepfakes statistics & trends 2026: Growth, risks and future insights*, KEEPNET, (March 14th, 2026), https://keepnetlabs.com/blog/deepfake-statistics-and-trends, (Last visited on 18th April 2026).
[4] Shannon Bond, *It takes a few dollars and 8 minutes to create a deepfake. And that's only the start.*, NPR, (March

2023 to over 8 million in 2025, reflecting exponential scalability.[5] Thirdly, the speed and scale of dissemination in digital platforms render rapid and irreversible harm. With such characteristics, deepfake as a tool amplifies the existing offences by providing the uniqueness of scalability, virality and realism to such offences. Therefore, deepfakes operate as a multiplier technology, enhancing both the scale and severity of traditional criminal acts.

Such unique features of the technology demand for a specific criminal-penal provision to deter the very unlawful creation and dissemination of harmful content. But before that analysis must be done regarding the deepfake harms at the individual, societal and national levels, how the government of India is tackling the same through legislative and regulatory steps, what are the legal gaps in such measures and lastly the required amendments in criminal framework and suggestions.

II. DEEPFAKES AND ASSOCIATED CRIMINAL HARMS

Deepfake manipulated contents blur the lines between authentic and fabricated content, challenging the very reliability of such information and content. Such realism enhances the capacity to manipulate perceptions, making fabricated contents indistinguishable from genuine records. Also, the growing accessibility of open-source generative AI tools has accelerated the proliferation and democratization of deepfakes technology across digital platforms, has expanded the potential misuse, enabling anyone to generate realistic yet fabricated content at minimal cost with limited technical knowledge. Such misuses are giving rise to criminal harms not only at individual level but also at societal and national level. Thereby fulfilling the very foundational requirement in criminal jurisprudence behind making an act or omission an offence is its harmful impact on society.

1. HARM AT INDIVIDUAL LEVEL

In 2018, an anonymous programmer on an online forum released a series of synthetic pornographic video superimposing the faces of well-known actresses onto explicit content without their consent.[6] The incident marked a turning point in the weaponization of deepfake

23rd, 2023), https://www.npr.org/2023/03/23/1165146797/it-takes-a-few-dollars-and-8-minutes-to-create-a-deepfake-and-thats-only-the-sta (Last visited on 18th April 2026).

[5] *See Deepfakes statistics & trends 2026, Supra* note 3.

[6] Samantha Cole, *AI-Assisted Fake Porn Is Here and We're All Fucked*, VICE, (Dec. 11th, 2017), https://www.vice.com/en/article/gal-gadot-fake-ai-porn/ (Last visited on 4th February 2026).

tool against individual dignity, targeted humiliation and abuse. Another incident, in 2023, where scammers were using AI voice cloning tools to impersonate family members in distress to extract emergency funds from victims[7], resulting in financial loss to victims of impersonation deepfakes. Beyond financial losses, this deepfakes involving impersonation creates reputational risks by attributing fabricated statements to the victim. For instance, a manipulated video purported to showing executives making discriminatory remarks which was circulated online causing reputational crisis before any verification can occur.[8] Therefore deepfake tools facilitate perpetrators to impersonate, commit fraud, extortion, defame, creating non-consensual sexual contents, causing harm at the individual level.

2. HARM AT SOCIETAL LEVEL

Deepfake has introduced a new vector for social harm that is collective in scope; synthetic media possess the power to deceive public at large and hamper social trust. It has the capacity to break the ability of citizens to trust what they see and hear, thereby undermining public confidence in audio-visual evidence. These audio-visual disinformative contents can go viral quickly due to enhanced persuasiveness and believability, as compared to the text only disinformation.

If such deepfakes are deployed in political contexts, it can magnify the pre-existing vulnerabilities like communal disharmony, inflammatory campaigns, hate speeches, deepening polarization, risk of social unrest and erosion of civic trust. For instance, in April 2024, during election season in India, a widely circulated audio clip allegedly featuring a senior political leader making inflammatory remarks, which was later suspected to be synthetically manipulated, incited public debate on the disruptive role of deepfakes.[9]

3. HARM AT NATIONAL LEVEL

Deepfake content can be a threat to a national security as they are capable of destabilizing the

[7] Alvaro Puig, *Scammers Use AI to enhance their family emergency schemes* (2023), FTC, https://consumer.ftc.gov/consumer-alerts/2023/03/scammers-use-ai-enhance-their-family-emergency-schemes (Last visited on 4 February 2026).

[8] Olina Banerji, *AI's Latest Hazard for School Employees: Deepfakes*, CENTRE FOR DIGITAL EDUCATION GOVERNMENT TECHNOLOGY, (May 8th, 2024), https://www.govtech.com/education/k-12/ais-latest-hazard-for-school-employees-deepfakes (Last visited on 5th February 2026).

[9] PTI, *Amit Shah deepfake video: case registered against Maharashtra youth congress's social media handle,* (April 30th, 2024), https://www.thehindu.com/news/national/maharashtra/amit-shah-deepfake-video-case-registered-against-maharashtra-youth-congresss-social-media-handle/article68124147.ece#google_vignette (Last visited on 27th February, 2026).

very information ecosystem of a nation. Deepfake can also be exploited by the adversarial states or non-state actors to destabilise rival democracies, invoke secession demands, fabricate battlefield evidence, undermine troop morale etc. For instance, during Operation Sindoor, deepfake were circulated where senior Indian defence officials such as chief of the army, air force, and navy as well as top Indian ministers were targeted by using their AI generated audiovisuals to spin narratives about India's supposed loss of combat equipment and personnel to Pakistan, or India conceding to Pakistan.[10]

The gravity of these collective harms conclusively frames the normative stakes of criminalization debate, while individual victims experience immediate injury, society as a whole, bears the risk of structural destabilization.

III. LEGISLATIVE AND REGULATORY STEPS TO SOLVE DEEPFAKE

Deepfake technology has posed a complex challenge for legal systems worldwide, and India is no exception to it. In India, the misuse of such technology has evoked increasing judicial attention and public debate, particularly as AI generated contents have begun circulating widely on social media platforms. To solve this, the enforcement authorities rely upon a combination of existing criminal and cyber law provisions, primarily the Bharatiya Nyaya Sanhita, 2023 (hereinafter referred as BNS), the Information Technology Act, 2000 (hereinafter referred as IT Act), the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 (hereinafter referred as IT Rules), to address offences which are results from unlawful deepfake content. Further the government of India also advocates that India is well equipped in tackling deepfakes by relying on existing provisions of BNS, IT Act, IT Rules.[11]

Government of India maintains the position that, even if the criminal framework doesn't explicitly criminalize deepfake, but we are equipped with provisions under BNS like section 294, 318, 319, 356 etc; or under IT Act, section 66C, 66D, 67, 67A, 79 etc, or under IT Rules, regulation like Rule 3 and 4. The application of these legislations and regulations are discussed below in detail.

[10] DAU secretariat, *AI generated misinformation during operation sindoor,* DEEPFAKE ANALYSIS UNIT, (Dec. 24, 2025) https://www.dau.mcaindia.in/blog/a-i--generated-misinformation-during-operation-sindoor (Last visited on 28th February, 2026).

[11] MeitY, *India well-equipped to tackle evolving online harms and cyber crimes; Government to Parliament*, PIB (Aug. 8th 2025) https://www.pib.gov.in/PressReleasePage.aspx?PRID=2154268&reg=3&lang=2 (Last visited on 10th March 2026).

1. APPLICATION OF BNS AND IT ACT

One of the most immediate points of intersection lies in offences relating to cheating and impersonation. Section 318 of BNS defines cheating as deception coupled with dishonest inducement, while Section 319 of BNS criminalises cheating by personation.[12] These provisions have relevance in deepfake induced audio-visual cloning to extort money, illustrating how traditional deception offences are being technologically augmented, as evident from the data showing that the deepfake fraud attempts spiked by 3000% in 2023.[13]

Further if we investigate the offence of defamation under section 356 of BNS, it provides another avenue for deepfake harms to be addressed. Deepfake videos that falsely depicts individuals engaging in unlawful conduct can significantly damage reputation, thereby satisfying the elements of defamation.

Offences relating to obscenity and sexual misconduct are particularly relevant in the revenge porn deepfake scenario. Section 294 of BNS penalises sale, distribution or public exhibition of obscene material.[14] However, obscenity laws were developed with physical or photographic representations in mind and not for images created through deepfakes. Deepfake helps in exacerbating harm by fabricating explicit material involving individuals who never participated in its creation. This not only violates privacy and dignity but also leads to enduring reputational and psychological harm, often with limited avenues for effective redress. The injury caused is not only immediate but also irreversible and continuously proliferating, as content may persist indefinitely on digital space.

The BNS also contain provisions to regulate public order and incitements, including Section 196 and Section 353, which penalise acts promoting enmity or spreading false statement likely to cause public mischief.[15] Deepfake videos that fabricate speeches can incite communal tensions or political unrest, may fall within these provisions.

In cases of cyber offences, one of the frequently invoked provisions in cases involving digital impersonation is Section 66C of IT Act, which criminalises identity theft through the fraudulent

[12] Bharatiya Nyaya Sanhita, 2023, § 318-319, No. 45, Act of Parliament, 2023 (India).
[13] *See Deepfakes statistics & trends 2026, Supra* note 3.
[14] *See* Bharatiya Nyaya Sanhita, 2023, *Supra* note 12, § 294
[15] *Id.* § 196, 353.

use of electronic signatures, unique passwords or any unique identifications.[16] Similarly, Section 66D of IT Act which penalises cheating by online impersonation through computer resources, may theoretically be applied to deepfakes enabled scams where audio-visual cloning is used to cheat and solicit financial transfers. Another significant area is the dissemination of obscene or sexually explicit material online. Section 67 of IT Act criminalises the publication or transmission of obscene material in electronic form, while Section 67A addresses the circulation of sexually explicit content.[17]

2. APPLICATION OF IT RULES, 2021

The Indian governmental response to deepfake related harms has been largely mediated through the intermediary liability under IT Act, 2000 and operationalised through IT Rules, 2021. This framework doesn't explicitly recognise deepfake as a distinct legal category; however, it establishes a system of conditional "safe harbour" under Section 79 of the IT Act whereby intermediaries are exempted from liability only upon compliance with prescribed due diligence obligations.[18] This conditional immunity thereby positions intermediaries as key enforcement actors in addressing unlawful online content.

Central to this framework is Rule 3 of the IT Rules, 2021, which prescribes the due diligence obligations that intermediaries must observe to retain safe harbour protection.[19] Rule 3(1)(b) requires intermediaries to inform users not to host, display, upload or share content that is defamatory, obscene, or otherwise unlawful, categories under which deepfake content is typically subsumed in practice.[20] Further, Rule 3(2)(a) mandates the appointment of a grievance redressal officer and requires intermediaries to acknowledge user complaints within 24 hours and dispose within 15 days, thereby institutionalising a notice based enforcement mechanism.[21] This notice and takedown model reflects a reactive regulatory approach, wherein the burden of identifying unlawful content is placed on affected individuals rather than on the platform itself. The reactive nature of this regime is further reinforced by Rule 3(1)(d), which requires intermediaries to remove or disable access to unlawful content within 36 hours upon receiving 'actual knowledge' through court order or government notification.[22] This provision

[16] Information Technology Act, 2000, No. 21, Act of Parliament, 2000 (India), § 66C.
[17] *See* Information Technology Act, 2000, *Supra* note 16, § 67–67A.
[18] *Id.*, § 79.
[19] Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, R. 3.
[20] *Id*., R. 3(1)(b).
[21] *Id*., R. 3(2)(a).
[22] *Id.*, R. 3(1)(d).

is read with the Hon'ble Supreme Court's interpretation in Shreya Singhal v. Union of India, that 'actual knowledge' is understood only upon formal legal notice such as court orders, thereby limiting the scope of proactive monitoring by intermediaries.[23] While this interpretation was made to safeguard freedom of speech by preventing arbitrary censorship, it also constrains the ability to respond swiftly to the rapid virality characteristics of deepfake content. Despite such provisions, the IT Rules lacked deepfake specific definition and did not account for unique characteristics of deepfake content that is realism, persuasiveness and virality, where the 36 hours of notice takedown mechanisms will also prove insufficient as victims face irreversible harms within minutes of sharing of content on digital platforms.

The limitations inherent in the pre-existing intermediary liability framework, particularly the lack of deepfake specific recognition, led the government of India to introduce a more targeted response through the 2026 amendment to the IT Rules, 2021.

3. 2026 DEEPFAKE SPECIFIC AMENDMENT IN IT RULES, 2021

The government of India, in 2026, brought amendments in IT Rules, 2021 which is a deepfake specific amendments to the Rules. This amendment explicitly incorporated a definition under Rule 2(1)(wa), for 'Synthetically Generated Information', as any audio-visual information generated artificially or algorithmically, in a manner which appear to be real, authentic or true and depicts any individual or event in a manner which is perceived as indistinguishable from a natural person or real-world event. Thereby including the very nature of deepfakes into the definition.

Another aspect of expanding the understanding of prohibited content can be seen under Rule 3(3)(a)(i)(IV), expressly including synthetic media that impersonate an identifiable individual or misrepresents factual information in a manner likely to cause harm with or without the involvement of natural person. This development is particularly significant because it moves beyond the earlier reliance on broad categories such as defamation, obscenity and others, and instead recognises the technological modality and it's harm inherent to deepfakes.

In addition to definitional clarity, the 2026 amendment strengthen intermediary obligations by introducing proactive compliance requirements that depart from the traditional notice and takedown model. Rule 3(3)(a)(i), now put due diligence on intermediaries to deploy

[23] *Shreya Singhal v. Union of India*, (2015) 5 SCC 1.

"reasonable efforts, including automated tools", to identify and mitigate the dissemination of synthetic or manipulated content that violates applicable laws. This marks a critical transition from a purely reactive regime which was dependent on user complaints or governmental order, to a risk-based model, wherein platforms are expected to actively monitor and address potential harms.

A further innovation introduced by 2026 amendment is the requirement of labelling in relation to deepfake content. Intermediaries are now obligated under new inserted clause that is Rule 3(3)(a)(ii) to ensure that the synthetic media is clearly identified by a permanent metadata or other technical provenance mechanism, also with a unique identifier to identify the computer resource used to create, generate or modify such information. This approach seeks to preserve lawful use of generative AI while mitigating the risk of deepfake harms, thereby adopting a disclosure-based regulations rather than outright ban.

The amendments also introduce stricter timelines for content removal, especially in cases of prima facie sexual deepfakes. The intermediaries are required to remove such contents within 2 hours, which was earlier 24 hours, under Rule 3(2)(b). In Rule 3(1)(d) for the content affecting the public order, intermediaries shall be taken down or disabled its access within 3 hours, which was 36 hours prior to 2026 amendment. Under Rule 3(2)(a)(i) the complaints must be now resolved within 7 days instead of 15 days. This accelerated takedown requirement reflects the government's recognition of rapid virality, irreversibility and harms associated with deepfake dissemination.

Despite these layered legislative and regulatory steps to solve the deepfake harms, the interventions remain largely reactive in nature. The provisions under BNS and IT Act are applied incidentally when the harm has already manifested or focuses on the consequences of deepfake harms and remains silent on the specific criminalization of the deepfake.

IV. GAPS IN EXISTING LEGAL RESPONSES

To understand the inadequacies in the current legal framework, deepfakes may be analysed at two distinct stages: the upstream stage and the downstream stage. The upstream stage refers to the creation, manipulation, and dissemination of unlawful synthetic content through deepfake technology. The downstream stage, on the other hand, concerns the consequences arising from such content, including defamation, fraud, impersonation, identity theft, misinformation, and

non-consensual pornography.

The existing criminal legal framework in India predominantly addresses the downstream harms caused by deepfakes by fitting their consequences within existing offences under the BNS and the IT Act, 2000. However, it remains largely silent on penalising the upstream acts of creating and distributing malicious deepfake content. For instance, the mere creation of a pornographic deepfake using an open source deepfake tool does not, in itself, attract criminal liability under the BNS. Liability generally arises only when such content is published, transmitted, or circulated within the digital ecosystem, thereby attracting provisions such as Section 294 of the BNS. Consequently, the present framework lacks deterrence at the stage of origin itself.

This downstream-centric approach is inadequate because it responds only after harm has already occurred. Deepfakes possess the capacity to inflict irreversible reputational, psychological, social, and economic damage within a very short span of time, particularly due to the speed and scale of digital dissemination. Therefore, there is a pressing need for upstream criminalisation that targets the creation and intentional dissemination of malicious deepfake content at its source, rather than merely addressing its eventual consequences. A proactive legal framework is essential to prevent harm before it materialises, instead of relying solely on remedial or reactive measures.

A significant development in this context is the judicial recognition of deepfakes as a distinct technological harm that transcends the conventional categories of offences contemplated under existing criminal law. In TV Today Network Ltd. v. Union of India[24], the Hon'ble Delhi High Court expressly acknowledged the misuse of emerging technologies such as deepfakes and observed the absence of a comprehensive statutory mechanism to effectively address the harms arising from them, particularly in relation to penal consequences. The decision reflects judicial awareness that existing legal provisions may be insufficient to deal with the unique nature, scale, and technological sophistication of deepfake-related offences.

V. NEEDED LEGISLATIVE AND REGULATORY REFORMS

1. LEGISLATIVE REFORM

The BNS, 2023, and various amendments in IT Act, 2000 while modernising several aspects

[24] *TV Today Network Ltd. v. Union of India*, 2025 SCC OnLine Del 4752, ¶ 4.

of Indian criminal law, does not expressly recognise or define deepfakes harms. This absence of statutory recognition leads to a patchwork solution for tackling with deepfake menace, where the harms arising from deepfake content are accommodated in the existing provisions of BNS or IT Act, such as fraud, defamation, impersonation etc, which are merely the consequences of the misuse of deepfake tool and the broader aspects of deepfake such as speed, realism and persuasiveness is often neglected. As a result, there is a need to introduce definitional clarity within the existing legal framework to ensure coherence and effective enforcement, as well as a specific provision criminalizing the very creation and distribution of unlawful deepfake content an offence. A primary reform in this regard which was highly emphasized by the policy makers is to incorporate a statutory definition of "deepfake".

A comprehensive definition for deepfake is provided under a Bill named "The Deepfake Prevention and Criminalization Bill, 2023". Section 2(c) defines deepfake as "*digitally manipulated or fabricated digital content, including but not limited to images, videos or audio recordings, created through the use of advanced digital technologies such as artificial intelligence, machine learning, or other advanced technologies, with the intent to convincingly and deceptively depict subjects or issues or represent individuals engaging in actions, making statements, or being in circumstances that did not occur or exist in reality".*[25]

Another most critical and central reform proposed in this article is criminalization of unlawful creation and distribution of deepfake content. The analysis undertaken in this article, established that Indian criminal law framework address deepfake harms reactively, focusing predominantly on the outcomes and not the origin of the same. The lacuna must be addressed through the explicit criminalization of the creation and dissemination of deepfakes with wrongful intent.

It is therefore proposed that a new provision be inserted, recognising creation and sharing of deepfake(s) as a criminal offence, **"Offence relating to the creation and dissemination of deepfake content -** *Whoever, by any means, creates, publishes, transmits, or otherwise disseminates any deepfake content, without the consent of individual involved or without digital watermark, with the intention of causing, facilitating, or committing an offence punishable under the law for the time being in force in India, shall be punishable with imprisonment for a*

[25] Deepfake Prevention and Criminalization Bill, 2023, Bill No. LXX of 2023, § 2(c).

term which may extend to five years, and shall also be liable to fine.

Explanation: *The offence under this section shall be independent of, and in addition to, any liability incurred under other provisions of law for offences arising from the use of impact of such deepfake content."*

The intent behind the addition of *explanation* in the provision is to ensure the automatic gradation of penalty. For instance, if any sexually explicit deepfake content is created and shared then the punishment for such creation would be applied that is up to 5 years and the punishment U/s 67A of the IT Act, 2000 will also be attracted with punishment up to 5 years. On the other hand, if a defamatory deepfake content is shared then punishment for such sharing would be applied that is up to 5 years and the provision U/s 356 of BNS, 2023 would also attract punishment up to 2 years.

This ensures proportionality while addressing the 'multiplier effect' of deepfakes, wherein synthetic media amplifies traditional offences by significantly enhancing their deceptive capacity and reach**.**

Further this proposed offence would have the following constituent elements. Firstly, actus reus through the act of creation, publication or distribution of synthetic content depicting an identifiable real individual without lawful justification or consent, where such content is capable of causing harm. Treating each stage of the pipeline (generation, modification, dissemination) as potentially constitutive of the offence. Secondly, mens rea can be construed through intentional misuse where an refaccused knowingly creates or disseminates deepfake content to deceive, harm, extort or defame etc; through a knowledge based liability where the person is aware that the content is synthetic and likely to cause harm but proceeds regardless of that; through reckless or negligence where the person disregards a substantial risk of harm (for example sharing unverified synthetic media). Further the intent may be inferred from the nature of the content like sexually explicit or potentially sensitive, or absence of consent. Thirdly, the resultant or potential harm to the depicted including individual's reputation, financial interest, or to public order and democratic integrity.

2. REGULATORY REFORM

There is a need to clarify and broaden the scope of the term "Actual Knowledge" and expanding

trigger mechanisms. The interpretation of actual knowledge as limited to court orders or government notification, as laid down in Shreya Singhal case, significantly constrains timely platform action. Reform is required to expand the scope of actual knowledge to include credible flagging mechanisms, such as verified user complaints through KYC-backed systems, trusted flaggers including recognised fact-checking organisations, and AI-assisted detection validated by human review.

Another reform which is advocated is to strengthen intermediary Due Diligence and Accountability. Although the 2026 amendments introduce proactive obligations such as "reasonable efforts and automated tools", the absence of precise standards creates potential ambiguity and inconsistent implementation.

Reforms should also aim at standardising due diligence obligation through mandatory periodic risk assessments for AI-generated content, independent algorithmic audits to assess bias and effectiveness of detection tools, transparency reports specifically addressing deepfake detection and removal. This can be ensured through provision which can require intermediaries to publish quarterly transparency reports on synthetic media moderation, including detection rates and takedown timelines.

Another reform is establishing a deepfake task force for the purpose if risk management, assisting central government to combat deepfake and development of detection technologies, as given under section 4 of the Bill, 2023.[26]

VI. CONCLUSION

Deepfake technology represents one of the most disruptive manifestations of contemporary artificial intelligence, fundamentally altering the relationship between reality and representation. As examined throughout this article, deepfakes are not merely manipulated media; they are highly sophisticated synthetic creations capable of fabricating speech, identity, and conduct with an unprecedented degree of realism. This realism, combined with accessibility and scalability, transforms deepfakes from a technological novelty into a potent instrument of harm. Unlike earlier forms of digital manipulation, deepfakes erode the evidentiary value traditionally attached to audio-visual content, thereby destabilising trust in

[26] Deepfake Prevention and Criminalization Bill, 2023, Bill No. LXX of 2023, § 4 & 5.

what individuals see and hear.

The harms caused by deepfakes operate across multiple dimensions that is individual, societal, and national, thereby justifying a distinct legal recognition. However, the most critical insight emerging from this research is that deepfakes do not always create entirely new categories of crime; rather, they transform and intensify existing offences. Deepfakes act as a *multiplier technology*, multiplying traditional crimes such as defamation, impersonation, fraud, obscenity, misinformation etc. This amplification occurs due to certain defining characteristics of deepfake technology, like realism, persuasiveness, speed, virality, scalability, low cost and rapid dissemination.

Despite these distinctive characteristics, the current Indian legal framework addresses deepfakes only in an indirect and fragmented manner. As analysed in this dissertation, provisions under the Bharatiya Nyaya Sanhita, 2023 and the Information Technology Act, 2000 are invoked to penalise the *consequences* of deepfake misuse, such as defamation, cheating, impersonation, obscenity, or public mischief. This approach reflects a downstream regulatory model, where liability arises only after harm has materialised in the form of a recognised offence. While this framework provides some degree of legal recourse, it remains fundamentally reactive and inadequate in addressing the unique nature of deepfake harms.

The central limitation of this downstream approach lies in its failure to address the upstream conduct namely, the creation and dissemination of unlawful deepfake content itself. The law, as it currently stands, does not treat the act of generating or distributing harmful synthetic media as an independent wrong. Instead, it waits for the deepfake to produce a legally cognisable consequence before intervention occurs. This creates a significant regulatory gap, as the harm caused by deepfakes often materialises instantly upon creation and dissemination, leaving little scope for preventive action.

This article therefore argues that deepfakes must be recognised as a distinct category of criminal harm, warranting a shift from a purely consequence-based model to a harm-preventive, conduct-based approach. Criminal law must move beyond punishing only the outcomes of deepfake misuse and instead address the entire lifecycle of the technology, including creation, manipulation, and dissemination of unlawful content.

VII. RECOMMENDATIONS

1. Introduce a statutory definition of 'deepfake' within BNS, 2023 or IT Act, 2000 to ensure clarity, uniform interpretation, and consistency.

2. Enact a specific criminal provision recognising the unlawful creation, generation and dissemination (upstream) of deepfake content as an independent offence, rather than relying solely on derivative harms (downstream) such as defamation, false impersonation, fraud, etc.

3. Expand the definition of 'actual knowledge' under Rule 3(1)(d) to include verified user through KYC-backed system, trusted flaggers inputs, and AI-assisted detection validated by human review.

4. Mandate periodic algorithmic audits and quarterly transparency reporting on synthetic media detection and removals, creating an institutional accountability mechanism.

5. Consider incorporating the key elements of Deepfake Prevention and Criminalization Bill, 2023, particularly the establishment of task force model, within the existing IT Act framework.