TRANSITIONING FROM PASSIVE SAFE HARBOURS TO ACTIVE SENTINELS: A CRITICAL INFORMATION TECHNOLOGY (INTERMEDIARY GUIDELINES AND DIGITAL MEDIA ETHICS CODE) AMENDMENT RULES, 2026.

By Padmini Majhi

VOLUME I | ISSUE I | ARTICLE IV

APRIL 2026

The Legalis IP Quarterly

Abstract

The proliferation of high-fidelity synthetic media in India presents unprecedented social, economic, and psychological risks to its 850 million internet users. Initially, an artistic medium, the malicious application of generative AI has facilitated identity manipulation, misinformation, and fraud, challenging existing regulatory frameworks spreading misinformation, and facilitating fraudulent activities, which have exposed serious regulatory and legal challenges. Currently, the Indian judicial system still relies on "technologically neutral" provisions, specifically Section 66C (Identity theft) and Section 66D (Cheating by personation) of the Information Technology Act 2000, and Sections 319 and 353 of the Bharatiya Nyaya Sanhita, 2023. While these laws are bailable and cognisable offences, they fail to address the technical aspects and forensic challenges of synthetic media. This paper focuses on the doctrinal and comparative methodology where a strategic realignment in Indian jurisprudence is introduced by the IT (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules 2026, where for the first time deepfake has been statutorily recognised as Synthetically Generated Information (SGI), which mandates watermarking and labelling mandates and Rule 4(1A), which requires platforms to verify user declarations through appropriate technical measures, while comparatively examining the IT Amendment

Rules 2026 against the EU AI Act's value-chain transparency model and the United States TAKE IT DOWN Act's harm-specific targeting approach. This research paper argues whether the current legislative pivot imposes an undue burden onto the digital intermediaries by transitioning them from "Passive Safe Harbours to Active Sentinels". Furthermore, forensic targeting of individual creators is often technically infeasible and shifts the liability to platforms, creating a de facto censorship regime. Such a shift risks undermining the legal protections and procedural due process established in the Shreya Singhal v. Union of India judgment. This interrogation evaluates the 2026 Rules through the lens of constitutional proportionality.

Keywords: *Deepfakes, Synthetically Generated Information (SGI), IT Amendment Rules 2026, Intermediary Liability, Safe Harbour, Shreya Singhal, EU AI Act, TAKE IT DOWN Act, Article 19, Digital Censorship, Generative AI Regulation*

I. INTRODUCTION

The Election Commission of India issued an advisory directing political parties to label AI-generated or synthetic content used in their social media campaigns.[1] The urgency of such measures became evident when, in late November 2025, a deepfake video circulated widely depicting Chief Election Commissioner Gyanesh Kumar bowing before Home Minister Amit Shah.[2] A similar incident had emerged in late November 2023, when actress Rashmika Mandanna's face was superimposed onto a video of British-Indian influencer Zara Patel. These incidents illustrate that, although India's existing legal framework addresses certain forms of cybercrime, it still lacks a targeted statutory regime capable of addressing the specific harms posed by synthetic media.

[1] *EC Directs All AI-Generated Poll Ads to Be Labelled as Such*, *Times of India* (Oct. 25, 2025), https://timesofindia.indiatimes.com/india/ec-directs-all-ai-generated-poll-ads-to-be-labelled-as-such/articleshow/124798284.cms

[2] *'Potential to Mislead': FIR Registered Against X for AI Video on PM Modi, ECI Chief Gyanesh Kumar*, *Times of India*(Mar. 26, 2026), https://timesofindia.indiatimes.com/india/potential-to-mislead-fir-registered-against-x-for-ai-video-on-pm-modi-eci-chief-gyanesh-kumar/articleshow/129817472.cms

The Indian judicial system still relies on "technologically neutral" provisions, specifically Section 66C (Identity theft) and Section 66D (Cheating by personation) of the Information Technology (IT) Act 2000, and Sections 319 and 353 of the Bharatiya Nyaya Sanhita (BNS) 2023.[3] Although these laws are non-bailable and cognisable offences, they fail to address the technical aspects and forensic challenges of synthetic media. Despite the significance of this regulatory shift, no comprehensive scholarly analysis has examined the constitutional validity of the 2026 Rules against the guardrails established in *Shreya Singhal v. Union of India* or evaluated India's approach with EU and US models. The Government of India notified the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026 (the "2026 Rules"), mandating a watermarking and labelling mandate and Rule 4(1A) mandating platforms to verify user declaration through technical measures burdening the social Media Intermediaries while transitioning them from passive safe harbours to active sentinels. This transition challenges the constitutional guardrails established in *Shreya Singhal*, which protects citizens freedom of speech and expression under Article 19(1)(a).

This paper argues that the 2026 Rules, while pursuing a legitimate aim, are constitutionally suspect, failing the proportionality test under Article 19(1)(a), replicating none of Section 69A's procedural safeguards, and structurally producing a de facto regime through safe harbour conditionality.

Part II examines the anatomy of deepfakes and the failure of existing Indian laws to address them. Part III analyses the pre-amendment intermediary liability regime and the constitutional guardrails of *Shreya Singhal*. Part IV examines the architecture of the 2026 Rules. Part V undertakes a comparative analysis of the EU and US frameworks. Part VI critically interrogates the constitutional validity of the 2026 Rules. Part VII proposes recommendations. Part VIII concludes.

[3] Information Technology Act, No. 21 of 2000, §§ 66C–66D (India); Bharatiya Nyaya Sanhita, No. 45 of 2023, §§ 319, 353 (India).

II. THE ANATOMY OF DEEPFAKES: TECHNOLOGY, HARM, AND THE LIMITS OF EXISTING INDIAN LAW

A. How Deepfakes Work: A Legal-Technical Primer

Deepfakes are synthetic media content produced through artificial intelligence techniques primarily Generative Adversarial Networks and diffusion models, capable of creating realistic audio, video and audio-visual content depicting real life individuals or events in a situation that never occurred. Unlike earlier forms of digital manipulation, deepfakes are capable of generating content indistinguishable from original media even to trained observers. Current AI-based detection tools maintain an accuracy rate of 65% to 70%, rendering them insufficient for the conclusive forensic identification required by judicial standards.

B. Harm Taxonomy: Identity Fraud, NCII, Electoral Manipulation, and Reputational Injury

The range of harms caused by deepfake technology breaks down into five different types, each requiring a specific regulatory response.

In the domain of identity fraud and financial crime, AI-generated videos of N.R. Narayana Murthy served as a bait to trick victims into a fake trading platform operated by Lakhani, while similar deepfake videos of Mukesh Ambani and Murthy cheated several victims of around ₹95 lakh in late 2024.[4]

In the sphere of non-consensual porn, for instance, on Oct 13th 2023, the face of Rashmika Mandanna was taken as a deepfake and digitally imposed on the body of British Indian

[4] HT News Desk, *Two Bengaluru people fell prey to Narayana Murthy and Mukesh Ambani deep fake videos, loses close to ₹90L, Hindustan Times (Nov 4, 2024),* https://www.hindustantimes.com/cities/bengaluru-news/two-bengaluru-people-fell-prey-to-narayana-murthy-and-mukesh-ambani-deep-fake-videos-loses-close-to-rs-90l-report-101730688063694.html *[https://perma.cc/2UG5-8Q3J].*

Influencer Zara Patel in the case below, clearly shows that synthetic media can destroy an individual's dignity with the help of a proper legislative framework.[5]

The electoral application of deepfake mischief was pushed to the foreground by none other than India during general elections of 2024 with an important observation: last year there was a surge in deepfakes, which include fake videos of political figures and celebrities that endorse a candidate who they never even met, therefore, harm impact democratic society directly.

Judicial interventions have been the consequence of harm to the reputation of public figures. Towards the end of 2025, Salman Khan, Kumar Sanu, Nagarjuna, Aishwarya Rai Bachchan and other celebrities approached the Delhi High Court against unauthorised use of their personas in AI-generated works.[6] Similarly, in *Anil Kapoor v. Simply Life India & Ors.*, the Delhi High Court granted an injunction which cements protection for personality rights from AI misuse.[7]

Apart from the politicians and other public figures, ordinary individuals too become victims of deep fakes in psychological and social fronts, an injury that has no sufficient legal redress since it is still less reported due to ignorance and stigma.

C. Why Technologically Neutral Provisions Fail: Section 66C, 66D, and the BNS 2023

Section 66C and Section 66D of the IT Act, governing identity theft and cheating by personation respectively, were designed to address password theft and impersonation fraud in financial transactions, not for the forensic and epistemic complexities of synthetically generated media.[8] Both provisions require proof of dishonest intent and an actual fraudulent

[5] Arvind Ojha, *Rashmika Mandanna deepfake video: Delhi Police registers case, India Today* (Nov 10, 2023), https://www.indiatoday.in/india/story/rashmika-mandanna-deepfake-video-delhi-police-case-registered-2461547-2023-11-10 [https://perma.cc/RN5R-4Y56].
[6] TOI Entertainment Desk, *After Aishwarya Rai Bachchan, Kumar Sanu and Nagarjuna, Salman Khan moves to Delhi High Court seeking protection of personality, publicity rights, Times of India (Dec 13, 2025), https://timesofindia.indiatimes.com/entertainment/hindi/bollywood/news/after-aishwarya-rai-bachchan-kumar-sanu-and-nagarjuna-salman-khan-moves-delhi-high-court-seeking-protection-of-personality-publicity-rights/articleshow/125897674.cms [https://perma.cc/F5WJ-BQCL].*
[7] *Anil Kapoor v. Simply Life India*, CS(COMM) 652/2023 (High Court of Delhi 2023).
[8] Information Technology Act, 2000, §§ 66C, 66D (India).

transaction, elements that are structurally inapplicable in the case of AI, synthetically generated content and digital forgery as there are no such fraudulent transactions in the deepfakes crime.[9]

Section 319 (Cheating by Personation) and Section 353 (Public Mischief) were designed to combat misinformation and deception, and suffers from the same structural inadequacy to deal with deepfakes.[10] Neither provision mentions forensic evidence required to prove the involvement of deepfakes or AI based crimes. Furthermore, a deepfake electoral misinformation video does not fit into this provision.[11]

Both IT Act provisions and BNS provisions are bailable offences, meaning the accused person can secure bail easily, that is inadequate to a category of harm causing an irreversible and psychological damage to the victims within hours of publication. Moreover, neither framework set down any forensic evidentiary standard to address the technical and forensic aspects in respect to synthetic media, leaving courts without any legal basis to evaluate deepfake evidence.[12] The legislative vacuum left by both acts necessitated the 2026 Rules, where the architecture and constitutional implications of which are examined in Parts III and IV.[13]

III. THE PRE-AMENDMENT INTERMEDIARY LIABILITY REGIME

A. Section 79 and the Safe Harbour Architecture

Section 79 of the IT Act establishes the foundational safe harbour architecture for the "Social Media Intermediaries" (such as Google, Instagram, Facebook) from liability for third party content hosted on their platforms.[14] The architecture ensures the platforms provide a neutral ground of communication rather than being an active participant in content creation or

[9] *Id.*
[10] Bharatiya Nyaya Sanhita, 2023, §§ 319, 353 (India).
[11] *Id.*
[12] Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026, rule 2(1)(wa) [hereinafter 2026 Rules].
[13] Information Technology Act, 2000, § 79.
[14] *Id.*

dissemination. For the intermediaries to be exempted from the liabilities, Section 79(2) states[15] that the intermediaries' role must be limited to provide a communication system, where third-party information is sent, hosted, and received. The intermediaries must not initiate or select the receiver of the transmission, or select or modify the information, it must also follow guidelines prescribed by the Central Government. Intermediaries lose the safe harbour protection upon proof of conspiracy, abetment, or inducement of an unlawful act, abetted, aided or induced for committing the unlawful act. Furthermore, if it fails to expeditiously remove or disable access to the information after receiving the actual knowledge or being notified by the government.

Section 79 imposes liabilities on intermediaries for committing an unlawful act but also exempts them from liability for the independent actions of their users.

B. The IT (Intermediary Guidelines) Rules, 2021

The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 (subsequently amended in 2022, 2023, and 2026),[16] under which the concept establishes Safe Harbour protection for intermediaries. Rule 7 is the "enforcement" clause[17] of the guidelines the Rules explicitly state the intermediary fails to observe the due diligence that is outlined, the intermediaries would be held responsible for the unlawful act and Section 79(1) shall not apply. Rule 3 requires intermediaries[18] to periodically inform users of their own rules, privacy policies and user agreements in English or in any language mentioned in the Eighth Schedule of the Indian Constitution. Rule 3(1)(b)[19] states to make reasonable efforts that users should not host content that contains obscene, pornographic, invading another's privacy, gender biased, infringing patents and property rights, misinformation, threatening the sovereignty, integrity or security of India. Once the intermediaries are notified the content must be removed or disabled no later than 36 hours.[20] Rule 3(1)(d) states[21] that

[15] *Id.* § 79(2).
[16] 2026 Rules, *supra* note 9, rule 7.
[17] 2026 Rules, *supra* note 9, rule 7.
[18] 2026 Rules, *supra* note 9, rule 3.
[19] 2026 Rules, *supra* note 9, rule 3(1)(b).
[20] 2026 Rules, *supra* note 9, rule 3(1)(b).
[21] 2026 Rules, *supra* note 9, rule 3(1)(d).

after removal of the content, the intermediaries must keep the information and associated records for 180 days for investigation purposes. Rule 4 prescribes heightened compliance obligations for Significant Social Media Intermediaries (SSMI)[22]. It requires such intermediaries to appoint a Chief Compliance Officer resident in India, who bears personal liability for any failure to observe due diligence requirements; to publish monthly compliance reports, including details of links removed through proactive monitoring; and to designate a nodal contact person for round-the-clock coordination with law enforcement agencies. For serious offences involving threats to State security or public order, Rule 4 requires SSMI to identify the first originator of the information pursuant to a judicial order. It also mandates the appointment of a grievance redressal officer under Rule 3(2), who must acknowledge complaints within 24 hours and resolve them within 15 days. A user dissatisfied with the decision of the grievance redressal officer may prefer an appeal before the Grievance Appellate Committee (GAC).[23]

These provisions not only act as an active obligation but also give a clear understanding where the Intermediaries may be liable and for what.

C. *Shreya Singhal v. Union of India*: Constitutional Guardrails on Intermediary Liability

In the landmark case of *Shreya Singhal vs Union of India (2015)*,[24] the Supreme Court delivered a foundational judgment to protect the fundamental right to free speech and expression under Article 19(1)(a),[25] and clarified how intermediaries are held liable for third-party content. Under Section 79(3)(b) of the IT Act,[26] an intermediary loses its "Safe Harbour Immunity" if it fails to remove the unlawful content after receiving the "actual knowledge" which is only triggered by a court order or a notification from the appropriate government. After striking down Section 66A[27], the Supreme Court clarified that intermediaries may, consistently with Article 19(2)[28], block or remove content only on

[22] 2026 Rules, *supra* note 9, rule 4.
[23] 2026 Rules, *supra* note 9, rule 3(2).
[24] *Shreya Singhal v. Union of India*, (2015) 5 SCC 1.
[25] India Const. art. 19, cl. 1(a).
[26] Information Technology Act, 2000, § 79(3)(b) (India).
[27] *Shreya Singhal,* (2015) 5 SCC 1, 93.
[28] India Const. art. 19, cl. (2).

constitutionally recognised grounds, including the sovereignty and integrity of India, the security of the State, public order, and incitement to an offence. The Court held that mere discussion or even advocacy of an unpopular cause remains protected under Article 19(1)(a), and that speech may be curtailed only when it crosses the threshold of incitement or otherwise threatens public order or State security. While upholding Section 69A[29], the Court stressed that blocking orders must be reasoned and remain open to judicial challenge. It further required that both the originator and the intermediary be given an opportunity to be heard, and that a review committee examine the legality of blocking directions at least once every two months. The judgment also reaffirmed the void-for-vagueness doctrine[30], holding that a law is unconstitutional where an ordinary citizen cannot reasonably understand what conduct it prohibits.

However, none of these constitutional safeguards find any equivalent in the IT Amendment Rules 2026,[31] which conspicuously fails to define with sufficient clarity what synthetic content is prohibited and on what grounds.

IV. The IT Amendment Rules 2026: Architecture and Obligations

A. Synthetically Generated Information: The New Definitional Framework

The IT Amendment Rules 2026, introduce for the first time a statutory definition of Synthetically Generated Information ("SGI").[32] According to the Rules, SGI refers to audio, visual or audio-visual information, which can be artificially or algorithmically created and later generated, modified or altered using a computer equipment in such a way where the information appears to be real, original, authentic and depict any individual or event in a manner that is likely to be perceived as indistinguishable from a natural person or a real world event.[33] The rules clarify that not all AI-generated content qualifies as SGI-exceptions

[29] *Shreya Singhal*, (2015) 5 SCC 1, 112.
[30] *Id.* 93.
[31] 2026 Rules, *supra* note 9, rule 2(1)(wa).
[32] 2026 Rules, *supra* note 9, rule 2(1)(wa).
[33] *Id.*

include noise reduction, colour correction or compression that does not misrepresent the content.[34] The Rules further exclude creation of standard educational materials such as PDFs, presentations that do not constitute false electronic records.[35] It also excludes AI tools that are used as accessibility tools for the purpose of translation, subtitles, refining or improving content material.[36]

B. The Watermarking and Labelling Mandate

Rule 3(3) imposes an "ex-ante" obligation for the intermediaries that facilitate the creation or sharing of SGI.[37] Under this rule, the intermediaries must deploy automated tools to prohibit users from creating or sharing unlawful SGI such as child sexual abuse material ("CSAM"), non-consensual intimate imagery ("NCII"), forged documents or impersonation as public figures.[38] The Rules further mandate that visual SGI content be prominently labelled, while audio content requires a prefixed audio disclosure. Intermediaries must also embed permanent metadata and a unique identifier into the SGI to track the source of the information.[39] Modification, suppression or removal of these labels or identifiers is strictly prohibited for users.[40] Rule 3(1)(d) reduces the takedown timeline from 36 hours to 3 hours for content identified by court order or government notification,[41] while Rule 3(2)(b) mandates removal within 2 hours for complaints involving morphed intimate imagery.[42]

C. Rule 4(1A): Platform Verification and the Constructive Knowledge Problem

An additional responsibility is imposed on the Significant Social Media Intermediaries under the Rule 4(1A).[43] This rule makes it mandatory for the SGI to be published with more stringent requirements where users should confirm and declare that such content is

[34] 2026 Rules, *supra* note 9, rule 2(1)(wa)(i).
[35] 2026 Rules, *supra* note 9, rule 2(1)(wa)(ii).
[36] 2026 Rules, *supra* note 9, rule 2(1)(wa)(iii).
[37] 2026 Rules, *supra* note 9, rule 3(3).
[38] *Id.*
[39] *Id.*
[40] *Id.*
[41] 2026 Rules, *supra* note 9, rule 3(1)(d).
[42] 2026 Rules, *supra* note 9, rule 3(2)(b).
[43] 2026 Rules, *supra* note 9, rule 4(1A).

synthetically generated using automated tools or technical measures as provided by the SSMIs. An SSMI which permits, promotes or fails to act upon violating SGI will be regarded as failing to exercise due diligence obligations.[44] The "deemed failure" provision effectively imposes a constructive knowledge standard that compels intermediaries to pre-screen content, in clear departure from the "actual knowledge" standard affirmed in *Shreya Singhal*. The constitutional implications of this shift are examined in Part VI.

D. Safe Harbour Conditionality and the Chilling Effect on Intermediaries

The IT Amendment Rules 2026 under Section 79 of the IT Act make safe harbour protection expressly conditional on compliance with SGI obligations.[45] The Rule 7 states that if an intermediary fails to observe the process of due diligence, they lose their immunity and become liable for punishment under the IT Act 2000 and the Bharatiya Nyaya Sanhita, 2023.[46] However, Rule 1B (Protection for Removal)[47] clarifies that if an intermediary removes SGI to comply with these IT Act guidelines (including using automated tools), this action does not constitute a violation of the Rules and preserves the intermediary's safe harbour status. The conditionality of safe harbour creates a structural incentive for platforms to over-remove borderline content rather than risk losing Section 79 protection entirely, producing precisely the chilling effect on legitimate expression that *Shreya Singhal* warned against.[48]

V. COMPARATIVE ANALYSIS: THE EUROPEAN UNION AND THE UNITED STATES

A. The European Union: Article 50, EU AI Act and the Value-Chain Transparency

[44] *Id.*
[45] 2026 Rules, *supra* note 9, rule 7.
[46] *Id.*
[47] 2026 Rules, *supra* note 9, rule 1B.
[48] *Shreya Singhal, (2015) 5 SCC 1*.

Article 50 of the EU AI Act, which comes into force in August 2026, establishes transparency obligations for AI-generated synthetic content across the AI value chain addressing deepfakes specifically through a risk-tiered, expression-preserving framework that contrasts sharply with India's broad SGI regime.[49]

Under Article 50 of the EU AI Act "intermediaries" referred to as online platforms or entities within the AI value chain have been categorised depending on whether they act as "providers" or "deployers" of the AI systems. The Article 50 defines AI Providers as entities who develop AI systems or place AI systems on the market under their name, including developers of General-Purpose AI (GPAI). Providers must ensure that AI-generated audio, image, video or text outputs in a machine-readable format which is detectable as synthetic.[50]

Providers are expected to offer free-of-charge interfaces (APIs) or public tools that allow third parties to verify the content if it is AI-generated or not. AI-deployers are entities or professionals who use AI systems as part of their business activities excluding private individuals. The responsibilities of deployers are to clearly disclose the image, audio or video content that has been artificially generated or manipulated ("deepfakes"). An AI-generated text published on matters of public interest must be disclosed unless it has gone through human editorial review. Deployers are expected to use a common AI icon standard at the point of first interaction with the user.[51]

The Draft Code of Practice adopts a multilayered technical approach combining visible disclosures with machine-readable watermarking, rejecting any single-tool solution.[52]

The Draft Code of Practice also establishes key exemptions. There are no obligations for the law enforcement if their AI is authorised by law for detection, preventing or investigating criminal offences. Exemptions have been provided for artistic, satirical or creative works which are non-intrusive in nature and do not hamper the enjoyment of work.[53] Labelling is

[49] Council Regulation 2024/1689, art. 50, 2024 O.J. (L) 1689/1 (EU AI Act).
[50] *Id.* art. 50(1).
[51] *Id.* art. 50(2).
[52] EUROPEAN COMMISION, DRAFT CODE OF PRACTICE ON GENERAL-PURPOSE AI MODELS (2024).
[53] Council Regulation 2024/1689, art. 50(4), 2024 O.J. (L) 1689/1 (EU AI Act).

not required for the content if the AI is used as "assistive functions" for standard editing which does not alter the semantics of the input.[54]

The EU model is notable for three features which are absent from India's framework, a value-chain split of responsibility between providers and deployers, exemptions for artistic and satirical expressions, and independent regulatory oversight through EU AI Office. These features offer critical lessons for Indian regulatory design which are examined in sub-section C.

B. The United States: The TAKE IT DOWN Act and Harm-Specific Targeting

Prior to 2025, the United States lacked a proper uniform framework to address the issue of rising non-consensual intimate imagery, including AI-generated deepfakes. The existing laws did not explicitly address digitally fabricated content. In response, the United States Congress enacted the Take It Down Act, signed into law on 19 May 2025,[55] establishing a new comprehensive framework criminalising the publication of NCII including AI-generated digital forgeries.[56] Unlike India's generalised SGI framework the act targets only the most harmful category of synthetic media while leaving political satire, artistic expression, and general AI-generated content unregulated.

The Act categorises offences based on the specific harm caused and demographics of the victims. For adult victims, the government must prove that the publication actually caused harm in psychological, financial or reputational terms. For minors, if the content intends to "abuse, humiliate, harass or degrade," the offender should be penalised. The Act specifically targets "digital forgeries" of deepfakes as a unique category of harm, criminalising them even when no original authentic content exists. Publishing these AI-generated deepfakes is also a federal crime.[57]

[54] *Id.* art. 50(4)(b).
[55] TAKE IT DOWN Act, Pub. L. No. 119-10, 139 Stat.___(2025).
[56] *Id.* § 2.
[57] *Id.* § 3.

The platform must remove reported NCII within 48 hours of receiving a complaint request. It is also required to make "reasonable efforts" to identify and remove identical copies. Failure to comply with these removal requirements will be treated as deceptive or unfair under federal consumer protection law, allowing the FTC to impose sanctions.[58] Platforms are additionally granted immunity from liability for good faith removal of content that constitute NCII, even if later inclined to be lawful content, contrasting sharply with the protection of India's safe harbour under Rule 7.[59]

The Act deliberately exempts political satire, artistic expression and general AI-generated content, reflecting a conscious First Amendment restraint. However, critics including the EFF and CDT have argued the Act's vague language risks capturing lawful content, raising First Amendment concerns despite its narrow scope.[60] Sub-section C draws comparative lessons from both models for Indian regulatory design.

C. Comparative Matrix and Lessons of India

The three regulatory frameworks differ in four specific key areas: scope, expression, duties of the intermediaries, and their oversight methods.[61]

Parameter	India	EU	US
Feature	**IT Amendment Rules 2026**	**EU (AI Act & Draft Code)**	**United States (TAKE IT DOWN Act)**
Scope	**Synthetically Generated Information (SGI)**: Realistic audio, visual, or audio-visual	**Synthetically Content and Deepfakes**: Broadly covers AI-generated audio, image, video, and text.	**Digital Forgeries**: Deepfake intimate visual depictions of an identifiable individual

[58] *Id.* § 4.
[59] *Id.* § 4(b).
[60] ELECTRONIC FRONTIER FOUNDATION, COMMENTS ON THE TAKE IT DOWN ACT (2025).
[61] 2026 Rules, *supra* note 9; Council Regulation 2024/1689, art.50, 2024 O.J.(L) 1689/1 (EU AI Act); TAKE IT DOWN Act, Pub. L. No.119-10,139 Stat.___(2025).

	content likely to be perceived as indistinguishable from real persons or events.		created or altered using AI or software. Also covers authentic NCII.
Expression Protection	Excludes 'good-faith' removal, accessibility tools, and routine document creation. Protection for discussion and advocacy via *Shreya Singhal* precedent.	Explicit exemptions for evidentiary artistic, creative, satirical, or fictional works, requiring only minimal disclosure that doesn't "hamper enjoyment".	Exceptions for good-faith disclosures, lawful purposes (medical/scientific/education), and matters of "**public concern**".
Intermediary Obligations	SSMIs must verify user SHI declarations ex-ante. All intermediaries must label SGI and deploy measures to prevent "prohibited categories".	Providers must ensure machine-readable making (watermarking/metadata). Deployers must disclose AI use at the point of first interaction.	**Covered Platforms** must provide a clear complaint process and make "reasonable efforts" to remove identical copies of reported content.
Safe Harbour	Removal via **automated tools** in compliance with Rules does not violate the Section 79 legal immunity conditions.	Primarily addressed via alignment with the Digital Services Act (DSA) and existing data protection laws.	Platforms are not liable for good-faith removal of content that appears to be unlawful NCII, even if it turns out to be lawful.
Enforcement	**Executive-led**: Level III	AI Office, a Scientific Panel	**Federal Trade Commission**

	oversight by the Central Government via an Inter-Departmental Committee. Orders issued by "**Authorised Officers**".	of Independent Experts, and National Competent Authorities.	**(FTC)** enforces notice and removal as a consumer protection violation. DOJ handles criminal prosecutions.
Takedown Timeline	**Aggressive**: **2 hours** for nudity/ impersonation; **3 hours** for government orders; **36 hours** for deceptive content; **7 days** for general grievances.	Focused on transparency at first exposure. Illegal content (NCII, etc) must be removed "**promptly**" under the DSA framework.	48 hours from receipt of a valid request.

From the EU model, India should learn three lessons. First, there should be a value chain split between AI-developers and platforms instead of burdening all intermediaries equally. Second, to explicitly exempt and protect satire, parody, journalism and artistic expression in the SGI definition. Third, to establish a separate regulatory authority from the Ministry of Electronics and Information Technology ("MeitY") to have an independent oversight in SGI frameworks.

From the US model, India should learn to regulate only the most demonstrably harmful synthetic content rather than all SGI broadly. Second, replace proactive monitoring with a victim-complaint mechanism modelled on the 48-hour TAKE IT DOWN Act process.[62] Third, to establish an independent enforcement equivalent to the FTC separate from MeitY.

[62] TAKE IT DOWN Act, Pub. L. No. 119-10, § 4 (2025).

India's 2026 Rules are simultaneously too broad in scope and too concentrated in enforcement. Both the EU and US model demonstrate effective deepfake regulation without sacrificing free speech and expression. This is a balance India's 2026 Rules have conspicuously failed to achieve.[63]

VI. CONSTITUTIONAL INTERROGATION: FREE SPEECH, DUE PROCESS, AND THE VAGUENESS DOCTRINE

A. Article 19(1)(a) and the Proportionality Test

Article 19(1)(a) of the Indian Constitution guarantees every citizen the fundamental right to speech and expression,[64] a right that the Supreme Court has consistently held to be the cornerstone of the democratic governance.

Any restriction on the freedom of speech and expression must satisfy the four-part proportionality test[65]: legitimate aim, rational connection, least restrictive means, and proportionality impact. While the 2026 Rules pursue a legitimate aim through a rational connection, they fail the 'least restrictive means' and 'proportionality' prongs of the test. While the 2026 Rules pursue a legitimate aim through a rational connection, they fail the 'least restrictive means' and 'proportionality' prongs of the test.

In the IT Amendment Rule 2026, the Rules mandate the labelling of AI- generated content to prevent citizens from deepfake harm which is indeed a necessary step. However, by failing to categorise specific harms, the government mandates a broad labelling requirement for all SGI content that is generated from an AI must be labelled and user verified before publication on intermediaries platforms.[66] On the other hand the EU AI Act has a carved-out model and the US has a harm specific model take it down Act and have clearly defined definition of what a deepfake SGI content is and what actions to be taken against them preventing them from deepfake harm. India nevertheless chose the broadest possible regulatory approach.[67]

[63] 2026 Rules, *supra* note 9, rule 2(1)(wa).
[64] India Const. art. 19, cl. (1)(a).
[65] *Modern Dental College v. State of Madhya Pradesh*, (2016) 7 SCC 353.
[66] 2026 Rules, *supra* note 9, rule 4(1A).
[67] 2026 Rules, *supra* note 9, rule 2(1)(wa).

MeitY centralises regulatory authority, acting as the sole arbiter of content legality while in EU and US they have set up a different regulatory authority to handle deepfake harm, as examined in Part V.

B. Overbreadth, vagueness, and the Suppression of Legitimate Expression

The Information Technology Amendment Rules 2026, has introduced the most proactive and stringent framework across three jurisdictions, marked by overbreadth and vagueness that risk suppressing legitimate expression.

The Indian framework is notably more extensive than the US or EU models due to its ex-ante requirements.[68] The Rule mandates that SSMIs must verify the correctness of user declarations regarding their SGI generated content. Mandating intermediaries for technical verification prior to publication is a broad obligation burdening the platforms to act as active sentinels. Deploying measures to prevent a wide range of content including SGI that results in false electronic records, where there has been no definition on false electronic records in the amendment rules. The rules also prohibit platforms from modification or removing SGI labels of metadata, where this could prevent developers from offering legitimate Al editing tools which may result in stripping metadata during file conversion.[69]

The 2026 Rules attempt to mitigate over-regulation through specific exclusions. Content is only considered as SGI if it is likely to be perceived as indistinguishable from a natural or real-world event. There is a lack of clear standard for "indistinguishability" which may lead to inconsistency by different platforms. Under the *Shreya Singhal's* "void for vagueness" doctrine, a restriction on speech must be defined with sufficient clarity that an ordinary person knows exactly what conduct is prohibited: a standard the SGI definition conspicuously fails to meet.[70] Editing or colour correction may fall within the scope of good-faith activity, but only to the extent that such processes do not misrepresent, alter, or distort the original content. In the absence of clear and granular standards, intermediaries

[68] 2026 Rules, *supra* note 9, rule 4(1A).
[69] 2026 Rules, *supra* note 9, rule 3(3).
[70] *Shreya Singhal,* (2015) 5 SCC 1.

may struggle to distinguish permissible enhancement from material misrepresentation, thereby creating a risk of over-blocking legitimate creative expression.

The enforcement architecture for SGI-related content privileges speed over deliberation, with potentially serious consequences for free expression. Complaints involving nudity or impersonation must be addressed within two hours, while content flagged through court orders or government directions must be taken down within three hours.[71] Since failure to comply may entail the loss of safe harbour protection[72], intermediaries are structurally incentivised to remove content immediately, often without sufficient scrutiny of whether the material constitutes satire, parody, or speech in the public interest. The requirement that SSMIs verify SGI declarations prior to publication may further create a verification bottleneck, suppressing speech until its relevance or immediacy has diminished. Moreover, the obligation to issue quarterly notices to users regarding prohibited content and criminal penalties[73] may reinforce a chilling effect, deterring the creation and dissemination of legitimate AI-generated creative and political expression.

C. Concentration of Executive Power and the Absence of Independent Oversight

The IT Amendment Rules 2026 concentrate regulatory authority exclusively within the executive branch and the absence of independent oversight. The Central Government maintains the authority over the digital media and Al ecosystem through a three-tier structure.[74] The directions to delete or block content are issued by an authorised officer and require a final approval of the Secretary, Ministry of Information and Broadcasting (MIB).[75] The grievances are heard by the Inter Departmental Committee (IDC), chaired by government authorised officers, composed of representatives from Home Affairs, Law Justice, MeitY. A Grievance Appellate Committee (GAC) exists to hear appeals, whose members are also appointed by the Central Government.[76]

[71] 2026 Rules, *supra* note 9, rules 3(1)(d), 3(2)(b).
[72] 2026 Rules, *supra* note 9, rule 7.
[73] 2026 Rules, *supra* note 9, rule 7.
[74] 2026 Rules, *supra* note 9, rules 1B, 7.
[75] *Id.*
[76] *Id.*

In *Shreya Singhal*[77], the Supreme Court upheld Section 69A on the ground that it was accompanied by procedural safeguards, including written reasons, a right to hearing, and oversight by a Review Committee. The 2026 Rules, by contrast, impose substantial obligations without providing any equivalent procedural protections. At the same time, MeitY operates simultaneously as rule-maker, enforcer, and adjudicator, concentrating these functions within a single executive authority without judicial oversight. Such an arrangement raises serious concerns under the principles of natural justice and separation of powers.[78]

The Secretary, MIB has the power to issue interim blocking directions in emergency cases without giving an opportunity of explanation to the publisher or the intermediary. A review committee also exists which is an executive body to meet at least once every two months to record the findings.

In contrast, the EU AI Act distributes regulatory authority through an independent institutional framework with enforcement delegated to the EU AI office and National Competent Authorities,[79] designated by individual member states, ensuring oversight remains separate from the ministry that made the rules. For the US TAKE IT DOWN Act, the primary body to enforce the act is the Federal Trade Commission (FTC), responsible for the notice and removal requirements, treating non-compliance as "deceptive or unfair". Criminal prosecutions are handled by the Department of Justice (DOJ).[80]

Taken together, these features show that the constitutional problem with the 2026 Rules lies not only in the breadth of their obligations, but also in the executive structure through which they are enforced. This concentration of power, without adequate procedural safeguards, creates a risk of suppressing lawful speech and leads to the next question: whether the Rules operate as a form of de facto censorship.

[77] *Shreya Singhal v. Union of India,* (2015) 5 SCC 1, 112.
[78] India Const. art. 50.
[79] Council Regulation 2024/1689, art. 50, 2024 O.J. (L) 1689/1 (EU AI Act),
[80] TAKE IT DOWN Act, Pub. L. No. 119-10, § 3 (2025).

D. The De-Facto Censorship Argument

The Rules do not explicitly censor, but they structurally produce censorship anyway. The introduction of “technical verification" mandate requiring Significant Social Media Intermediaries (SSMIs) to verify the accuracy of user declarations regarding SGIs and removing contents via automated tools bypasses the *Shreya Singhal* judgment under Section 79(3)(b) IT Act and Article 19(2) of the Constitution.[81]

The SGI definition in IT Amendment Rule 2026 is completely vague as it fails to identify how an SGI would be perceived as indistinguishable. In the *Shreya Singhal* case, the Court struck down Section 66A of the IT Act 2000 for the terms which were constitutionally vague.[82]

The pre-publication technical verification mandate may suppress the creative works for political satire or creative works, and leaves no ground for discussion or advocacy to hear the views of the user or intermediaries. No appellate mechanism, no judicial review, no timelines for restoration of wrongly removed content has been described.[83]

The IT Amendment rules requirement for updating users every 3 months for Criminal penalties, including imprisonment could be viewed as a chilling effect on the freedom of speech and expression.[84]

The cumulative effect of these structural pressures completes the transformation this paper identifies, digital intermediaries have been converted from passive harbours into active sentinels, at a constitutional cost that neither the legislature nor MeitY has adequately acknowledged.[85]

[81] *Shreya Singhal v. Union of India,* (2015) 5 SCC 1; Information Technology Act, 2000 § 79(3)(b) (India).
[82] *Shreya Singhal,* (2015) 5 SCC 1, 93.
[83] 2026 Rules, s*upra* note 9, rule 7.
[84] 2026 Rules, *supra* note 9, rule 7.
[85] *Id.*

VII. TOWARDS A BALANCED REGULATORY FRAMEWORK: RECOMMENDATIONS

A. Redefining SGI: narrowing the Scope to Harmful Synthetic Media

The current SGI definition of India is too broad and vague,[86] and it fails to distinguish between malicious deepfake content and creative content for entertainment. The SGI definition must be narrowed to cover only synthetic depictions capable of causing cognisable psychological, electoral, or reputational harm. India should adopt a harm-specific mechanism similar to the dual classification of US Take it down act distinguishing between authentic content and AI-generated digital forgeries.[87] Moreover, as the EU AI Act has protected such content as satire and art,[88] India should do the same. It should exempt creative works like satire, parody, journalism and creative expression from the SGI definition. This narrower definition would fix the problem we identified in Part VI and pass the Article 19(1)(a) proportionality test.[89] This would also satisfy the *Shreya Singhal's* case for void for vagueness and provide sufficient clarity among the intermediaries and the users on what the content falls under the SGI definition.[90]

B. Graduated Intermediary Obligations Based on Risk and Scale

India's current labelling mandates burden all large platforms regardless of their functions of a platform that provides AI tools, and faces the same obligations to the platforms that publish the content.[91] A risk-based, tiered obligation framework should replace the current uniform mandates, scaling requirements according to a platform's reach and technical function. A risk-based, tiered obligation framework should replace the current uniform mandates, scaling requirements according to a platform's reach and technical function. Small platforms should be required to perform basic labelling, medium platforms moderate verification, and large

[86] 2026 Rules, *supra* note 9, rule 2(1)(wa).
[87] TAKE IT DOWN Act, Pub. L. No. 119-10, § 2 (2025).
[88] Council Regulation 2024/1689, art. 50(4) (EU AI Act).
[89] India Const. art. 19, cl. 1(a).
[90] *Shreya Singhal,* (2015) 5 SCC 1.
[91] 2026 Rules, *supra* note 9, rule 4(1A).

SSMIs full verification and proactive monitoring. India can learn from the EU's AI Act where there has been a separate obligation for the platforms who are providers and deployers,[92] and also those who build AI tools from those distributing content.

A tiered approach is more proportionate under Article 19(1)(a),[93] where obligations match capacity and risks rather than imposing uniform burdens that disproportionately harm smaller platforms and stifle innovation. Restoring meaningful safe harbour protections alongside this tiered framework is examined in sub-section C.

C. Restoring the Safe Harbour and Preserving *Shreya Singhal*

Rule 7 of the IT Amendment Rule 2026,[94] makes safe harbour conditional on proactive compliance. If platforms fail to verify SGI declarations, they may lose safe harbour protection under Section 79 of the IT Act 2000.[95] However, the Supreme Court in *Shreya Singhal* held that,[96] intermediaries cannot be required to proactively monitor content as it may restrict freedom of speech and expression.

To restore the Safe Harbour, platforms should only lose Safe Harbour after receiving the actual court order or government notification identifying specific unlawful content and not for failing to proactively detect it.

The Government should introduce a statutory notice and takedown process modelled on the US TAKE IT DOWN Act,[97] where victims report the harmful SGI directly to the platforms, which will trigger a mandatory removal obligation within a defined time frame without active monitoring.

[92] Council Regulation 2024/1689, art. 50(1)-(2), (EU AI Act).
[93] India Const. art. 19, cl. 1(a).
[94] 2026 Rules, *supra* note 9, rule 7.
[95] Information Technology Act, 2000, § 79.
[96] *Shreya Singhal,* (2015) 5 SCC 1.
[97] TAKE IT DOWN Act, Pub. L. No. 119-10, § 4 (2025).

This restoration would bring the framework back into alignment with *Shreya Singhal's* actual knowledge standard and preserve the safe harbour architecture essential to open digital communication.[98]

D. Establishing Independent Oversight and Judicial Review

The 2026 Rules currently centralise all rulemaking and enforcement within MeitY,[99] creating a significant oversight gap that lacks independent judicial review. To address this gap, India should establish an autonomous AI and Digital Media Regulator separate from the ministry, complemented by a multi-stakeholder advisory council of experts to provide a necessary check on executive power. Furthermore, all takedown orders must be subject to mandatory judicial review within a set timeframe to prevent permanent content removal without external scrutiny. By incorporating written reasons and the right to a hearing, this approach restores the vital procedural safeguards mandated by the *Shreya Singhal* ruling,[100] ensuring the rules are both fair and constitutionally sound.

E. Targeted Criminal and Civil Remedies for Victims

The current legal framework, spanning Sections 66C and 66D of the IT Act,[101] and Sections 319 and 353 of the BNS,[102] is fundamentally not suited for the harms caused by deepfakes, as these bailable offences offer no specific remedies for victims. To address this, India should introduce a dedicated non-bailable offence for malicious deepfakes with graduated penalties that distinguish between NCII (Non-Consensual Intimate Imagery), electoral interference, and identity fraud. This should be paired with a civil right of action—inspired by the proposed US DEFIANCE Act.[103] The civil right of action would allow victims to sue for statutory damages without proving exact financial loss. Finally, establishing statutory forensic standards is essential to help courts reliably navigate the "detection gap" where

[98] *Shreya Singhal,* (2015) 5 SCC 1, 93.
[99] *Supra* note 9, rules 1B, 7.
[100] *Shreya Singhal,* (2015) 5 SCC 1, 112.
[101] Information Technology Act, 2000, §§ 66C, 66D.
[102] Bharatiya Nyaya Sanhita, 2023, §§ 319, 353.
[103] *Disrupt Explicit Forged Images and Non-Consensual Edits Act of 2024*, H.R. 7569, 118th Cong. (2024) (not enacted).

automated detection accuracy remains insufficient. By focusing on the actual perpetrators rather than just the platforms, these targeted remedies protect legitimate expression while finally providing meaningful accountability for genuine harm.

VIII. CONCLUSION

The Rashmika Mandanna incident exposed a fundamental governance failure as it did not have a statutory framework to address the synthetic media harm. The IT Amendment Rules 2026 were India's first legislative response to the crisis.[104]

The Rules introduce important innovation. First, the introduction of the SGI definition, labelling mandate, and verification obligations, but the constitutional infirmities are significant. The broad SGI definition labelling every content as harmful irrespective of analysing the content whether it is satirical or creative in nature, fails *Shreya Singhal's* void for vagueness standard.[105] Burdening the intermediaries with labelling mandates and verification obligations is turning them from Safe harbour to Active Sentinels. Furthermore, the concentration of executive power with MeitY replicates none of Section 69A's procedural safeguards,[106] as the court upheld in the *Shreya Singhal* case.

Both the EU AI Act,[107] and the US TAKE IT DOWN Act,[108] demonstrates that effective deepfake regulation is achievable without sacrificing the constitutional rights of freedom of speech and expression. India’s failure to adopt a value-chain responsibility model or a harm-specific approach leaves its SGI framework constitutionally vulnerable and comparatively isolated.

The question India must now answer is not whether deepfake should be regulated, they must be, but rather what should be regulated and what to exempt. The more pressing question is whether regulation can be achieved without converting every intermediary into an instrument

[104] *Supra* note 9, rule 2(1)(wa).
[105] *Shreya Singhal*, (2015) 5 SCC 1.
[106] *Shreya Singhal*, (2015) 5 SCC 1, 112.
[107] Council Regulation 2024/1689, art. 50 (EU AI Act).
[108] TAKE IT DOWN Act, Pub. L. No. 119-10 (2025).

of state surveillance. The answer, as this paper has argued, lies not in abandoning the 2026 Rules entirely but in rebuilding them on the constitutional foundations that *Shreya Singhal* laid in 2015.[109]

[109] *Shreya Singhal,* (2015) 5 SCC 1.