FROM SAFE HARBOUR TO SYNTHETIC TRUTH: INDIA'S 2026 IT RULES AMENDMENT AND THE CONSTITUTIONAL FUTURE OF AI-GENERATED CONTENT

Saransh Patwal, Rajiv Gandhi National University of Law, Patiala

I.INTRODUCTION

In October 2023, a digitally manipulated video of actress Rashmika Mandanna circulated across Indian social media platforms, which triggered an emergency advisory from the Ministry of Electronics and Information Technology.[1] This incident exposed a regulatory flaw in the intermediary liability jurisprudence: what happens when harmful content is not merely shared by an intermediary, but it is fabricated with?

India's regulatory response was seen by the 2026 amendment to the Information Technology (Intermediary Guidelines and Digital Ethics Code) Rules, imposing affirmative obligations on the social media platforms to detect, disclose, and remove "synthetically generated information" (SGI).[2]

This article argues that while the regulatory action is constitutionally legitimate, but the structure of the 2026 amendment is defective. By transforming the intermediaries from spectators to active arbiters with no requirement that the content actually causes harm or not, no proportionate limits, and no one to keep a check on them, they would end up silencing a lot of legitimate speech just to avoid getting in trouble. Through legal analysis and comparative examination, this work proposes an alternative to preserve the regulatory objective while protecting the freedom of speech which is threatened.

[1] 'From Rashmika deepfake to draft rules: India's 2-year fight against AI misuse' *Moneycontrol* (23 October 2024) https://www.moneycontrol.com/artificial-intelligence/from-rashmika-deepfake-to-draft-rules-india-s-2-year-fight-against-ai-misuse-article-13628828.html accessed 30 March 2026.

[2] Ministry of Electronics and Information Technology , 'The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021' (updated 10 February 2026) https://www.meity.gov.in/static/uploads/2026/02/550681ab908f8afb135b0ad42816a1c9.pdf accessed 30 March 2026.

II. THE SAFE HARBOUR REGIME AND ITS FRACTURE

The core premise of India's intermediary liability regime is that the platforms are not the publishers. Section 79 of the Information Technology Act, 2000 grants intermediaries' immunity from third party content liability,[3] if they make sure that they don't modify or tamper with the offending content. The Supreme Court in *Shreya Singhal V. Union of India* supported this structure, stating that intermediaries could only be required to take down the content when either the court or the government orders them to, and not upon private complaint alone.[4] The order was well thought as the court realized that proactive monitoring obligations on the platforms would produce excess removal of content than the law strictly requires.

But the 2021 Intermediary Guidelines began eroding this foundation, Rules 3 and 4 introduced due diligence obligations that required platforms to monitor content categories, respond to grievances within designated deadlines, and most importantly, enable traceability of encrypted messages.[5] This traceability mandate still remains uncertain , with WhatsApp contending before the Delhi High Court that mandatory traceability is in violation to the right of informational privacy recognized in Justice K.S. Puttaswamy (Retd) V. Union of India.[6]

By defining SGI broadly and imposing affirmative detection obligations, including watermarking mandates, disclosure requirements, and time bound takedown obligations, the amendment fundamentally changes the intermediary's role. Platforms are asked to pre-emptively identify and manage a category of content defined by its technical origin rather than its actual impact. The safe harbour, once a structural guarantee of an open internet, has become contingent on a platform's capacity to perform epistemological functions the legal system has never previously demanded.

III. SYNTHETIC TRUTH AS A CONSTITUTIONAL PROBLEM

A. The Over-Breadth Problem

Article 19(1)(a) of the Constitution guarantees freedom of speech and expression, while

[3] Information Technology Act 2000, s 79; see also 'Section 79 of IT Act, 2000: Understanding the Safe Harbour Rule' (2025) LaEx.
[4] *Shreya Singhal v Union of India* (2015) 5 SCC 1.
[5] *Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules 2021* rr 3–4 ('IT Rules 2021'); *WhatsApp LLC v Union of India* (Delhi HC, WP(C) 6222/2021, pending).
[6] *Justice KS Puttaswamy (Retd) v Union of India* (2017) 10 SCC 1; *WhatsApp LLC v Union of India* (n 5).

Article 19(2) enumerates narrow exceptions. The Supreme Court in *Modern Dental College and Research Centre v State of Madhya Pradesh* adopted a structured proportionality framework requiring that any restriction on fundamental rights be suitable, necessary, and strictly proportionate to the objective pursued.[7] The 2026 amendment fails this test at both the necessity and proportionality stages.

AI-generated content is not inherently harmful, political satire using synthetic voice modulation, investigative journalism deploying AI reconstruction of suppressed evidence, documentary filmmaking using deepfake technology for archival restoration, and counter-speech campaigns using synthetic media to expose disinformation all fall within the amendment's definitional scope without causing any of the harms it targets.[8] An obligation to detect and disclose all SGI, regardless of demonstrated harm is baseless and an overkill. In *S Rangarajan v P Jagjivan Ram*, the court said that restrictions on speech must be proximately and directly, not remotely or speculatively connected to the evil they seek to prevent.[9] The law is already unclear, and it's made worse by the fact that the people can be punished even in absence of any mens rea. It treats accidental, satirical, and malicious synthetic content as same, denying courts the contextual analysis that has historically distinguished protected expression from actionable speech.

B. The Chilling Effect on Dissent

The amendment interferes with the freedom of expression rights of political activists. In *Anuradha Bhasin v Union of India*, the Supreme Court held that restrictions on speech-enabling infrastructure (for eg.- internet) must satisfy tests of legality, necessity, and proportionality, and must be subject to judicial scrutiny.[10] When platforms face open-ended liability for hosting synthetic content that is satirical, clearly labelled, or non-harmful, the rational response is the removal of it. This amendment makes the people fear exercising their right to expression which the Constitution most wants to protect.

A political satire page using AI-generated audio to mock a minister's rhetorical style, or an activist organisation using AI videos to represent communities that are marginalized and too

[7] *Modern Dental College and Research Centre v State of Madhya Pradesh* (2016) 7 SCC 353 [138].
[8] Internet Freedom Foundation, 'Deep Fakes and Democracy: Regulatory Challenges for India' (Working Paper, 2024) https://internetfreedom.in accessed 20 March 2026.
[9] *S Rangarajan v P Jagjivan Ram* (1989) 2 SCC 574 [45].
[10] *Anuradha Bhasin v Union of India* (2020) 3 SCC 637 [96].

vulnerable to speak publicly or a journalist using AI reconstruction to show censored evidence. These all are harmless and legitimate use of freedom of expression via AI, but still are all subject to takedown under the 2026 amendment without any requirement, if they have caused actual harm or not. *Foundation for Media Professionals v Union Territory of Jammu and Kashmir* reinforced that access to the internet is itself a very important aspect of freedom of speech and the government should have a solid reasoning to tamper with it.[11]

C. The Presumption Inversion

The core problem with the amendment is that it treats the content as 'guilty until proven innocent'. *Virendra v State of Punjab* established that stopping speech before it causes harm (prior restraint), demands a higher constitutional justification than punishing speech after it has been published.[12] The 2026 amendment's detection-and-removal obligations are, in practice, exactly that, platforms end up suppressing synthetic content before anyone is hurt, purely because of how it was made (using AI).

Current AI detection tools make this worse, as they miscalculate constantly, as they flag content that isn't AI-generated and miss out on content that actually is AI-generated.[13] When platforms face penalties for those misses, accuracy stops being the goal. The rational response is to remove anything that looks AI generated and ask questions later. That is a system where speech is treated as suspect by default which is the opposite of what Article 19 actually says.

IV. COMPARATIVE LENS: PRECISION VERSUS BLUNTNESS

India is not the first country to deal with this problem, as the EU's approach of combining the AI Act with the Digital Services Act, draws a line that India's amendment does not. Disclosure is mandatory for AI-generated content, but removal requires demonstrated harm in a specified category. Technical origin alone is not enough to trigger liability.[14] Article 50 of the AI Act requires transparency without treating AI generated content as guilty without investigating if its harmful or not. The accompanying Code of Practice turns this into industry-led standards

[11] *Foundation for Media Professionals v Union Territory of Jammu and Kashmir* (2020) 5 SCC 746.
[12] *Virendra v State of Punjab* AIR 1957 SC 896.
[13] Wenhao Yu and others, 'AI-Generated Content Detection: Technical Limitations and Legal Implications' (2024) 37 Harv JOLT 215.
[14] 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), art 50; Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services (Digital Services Act); European Commission, 'Code of Practice on Transparency of AI-Generated Content' (2024).

under regulatory supervision and not a regime where platforms act unilaterally under threat of statutory penalty.

China's Deep Synthesis Provisions (2022) points in a different direction, and the warning is worth sitting with. Beijing mandates disclosure labels and restricts AI content depicting real individuals without consent.[15] On paper, it looks attractive but in practice, the framework has operated as a censorship mechanism. The agencies that enforce it are the same ones that prosecute dissent, and there is no independent court anywhere to seek relief from. India's 2026 amendment shares more of this enforcement structure than it is comfortable to acknowledge. While China's rules might be strict or controversial, they make logical sense within their own system. India's rules, by contrast, are currently confusing and inconsistent.

V. TOWARD A CONSTITUTIONALLY COHERENT FRAMEWORK

Despite all the flaws in the amendment, it is still fixable and can follow the Constitution while still effectively controlling the risks of AI, if it just implements these changes:

First, replace synthetic-origin liability with a harm threshold, it means that a takedown obligation should not activate the moment content is flagged as AI-generated, but it should activate when the content falls within an exception under Article 19(2) and a real connection/intention to harm is established. *R Rajagopal v State of Tamil Nadu* is the right doctrinal anchor for this, as it asks what harm has occurred rather than drawing categorical lines around a content type and calling it done.[16]

Second, the case of *Shreya Singhal* must extend to synthetic content explicitly. It states that removing content should require a court's order or a transparent, legally authorized government decision.[17] Platforms are not courts, and they should not have the final say on what counts as harmful AI speech/content, especially when their incentive is to over-remove rather than get it right. An independent and new board or committee with genuine experts from the relevant field should be set up. This would give the regime procedural legitimacy without slowing down legitimate enforcement cases.

[15] 'Provisions on the Administration of Deep Synthesis Internet Information Services' (promulgated by the Cyberspace Administration of China, 25 November 2022, effective 10 January 2023) translated in *China Law Translate* https://www.chinalawtranslate.com/en/deep-synthesis/ accessed 28 March 2026.
[16] *R Rajagopal v State of Tamil Nadu* (1994) 6 SCC 632 [26].
[17] *Shreya Singhal* (n 4) [118].

Third, a good-faith safe harbour must be preserved. Platforms that implement technically sound, reasonable detection measures in good faith should retain Section 79 immunity even when synthetic content escapes detection. Penalising technically reasonable actors for imperfect outcomes will drive market consolidation toward only the largest, best-resourced platforms, chilling the intermediary ecosystem rather than simply the harmful content within it. Preserving good-faith immunity also incentivises voluntary investment in detection technology, producing the regulatory outcome the amendment seeks through market-compatible rather than punitive means.

VI. CONCLUSION

India faces a genuine regulatory challenge. AI-generated disinformation in a country of India's linguistic diversity, electoral scale, and documented history of violence triggered by viral falsehoods is not a hypothetical harm; it is a present constitutional emergency. The 2026 IT Rules amendment responds to a real and serious problem. But the constitutional future of AI-generated content in India will be determined not by whether we regulate synthetic truth, but by whether the architecture of that regulation respects the presumption of liberty that underlies Article 19. As currently framed, the amendment risks sacrificing constitutionally protected dissent, satire, and political speech in exchange for control over a technically defined content category. India's Constitution demands a more precise instrument.