



Jus Corpus Law Journal

Open Access Law Journal – Copyright © 2026 – ISSN 2582-7820

Editor-in-Chief – Prof. (Dr.) Rhishikesh Dave; Publisher – Ayush Pandey

This is an Open Access article distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike 4.0 International (CC-BY-NC-SA 4.0) License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium provided the original work is properly cited.

Deepfakes, Democracy and the 3-Hour Takedown Rule

Akshata Shetty^a

^aBharati Vidyapeeth New Law College, Pune, India

Received 04 June 2026; Accepted 04 July 2026; Published 07 July 2026

*On February 10, 2026, India's Ministry of Electronics and Information Technology rolled out sweeping new amendments to its IT Rules, fundamentally redrawing the country's digital speech landscape.¹ Platforms now face a sharply reduced three-hour window, down from the previous thirty-six, to take down deepfakes or synthetic content once ordered by the government. There's also a new requirement: all AI-generated content must be clearly labelled. Miss these obligations, or even stumble, and the consequences are severe: platforms lose their "safe harbour" protection, becoming as legally exposed as the users who post the content in question. The rationale for this change is obvious and deeply felt. Synthetic media has been weaponised in India: think election deepfakes, non-consensual intimate imagery, and state-led information warfare across borders. The harms are not abstract. They're immediate, sometimes devastating, and threaten the very foundation of trust in public discourse. Yet, the government's cure comes loaded with costs of its own. By bypassing judicial checks and procedural safeguards, the system places extraordinary power in the hands of the executive, delegating censorship to private intermediaries, and, in doing so, operating outside the judicial oversight and proportionality that India's Constitution demands. This article contends that the 2026 amendment directly undermines Article 19 of the Constitution and the standards set in *Shreya Singhal v Union of India*,² converting once-neutral intermediaries into de facto enforcement arms for the state. By comparing India's model to*

¹ 'Govt Sets 3-Hour Deadline for Social Media Platforms to Take Down AI-Generated, Deepfake Content' News on AIR (10 February 2026) <<https://www.newsonair.gov.in/government-sets-3-hour-deadline-for-social-media-platforms-to-take-down-ai-generated-deepfake-content>> accessed 02 June 2026; Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules 2026

² *Shreya Singhal v Union of India* (2015) 5 SCC 1

regulatory frameworks in the EU and the United States, this article argues that New Delhi's solution, intended to defend democracy, places democracy itself in peril by making truth and lies equally disposable at the stroke of an executive order.

Keywords: *safe harbour, peril, democracy, deepfake.*

INTRODUCTION

On March 18, 2026, stand-up comedian Pulkit Mani uploaded a satirical Instagram reel that skyrocketed to over 16.5 million views.³ Then, almost as suddenly as its rise, the reel was blocked nationwide. A legal notice referencing Section 79(3)(b) of the IT Act⁴ was all anyone got by way of explanation. There was no public reasoning, no chance to contest, no due process. That same night, several parody and satire accounts on X disappeared, gone in a blink, with no transparency. This wasn't some deepfake conspiracy or dirty trick: the reel was legitimate satire. It simply vanished, snuffed out by new government rules. That incident cuts to the heart of the matter: a real crisis has met a regulatory sledgehammer so broad that it doesn't just target the disease; it starts consuming healthy flesh with it.

India's new rules mark an unprecedented clampdown on synthetic media. For the first time, 'synthetically generated information' gets a legal definition. Platforms must react within three hours, not days, when the government flags content. And now, 'safe harbour' is precarious: one missed deadline, and online giants can face crippling liability. The deepfake threat is painfully real. In 2024, election deepfakes of political leaders circulated with ease. Videos showing Bollywood celebrities appearing to criticise Prime Minister Modi went viral, racking up hundreds of thousands of views before exposure.⁵ In 2023, manipulated, sexually explicit videos of prominent women, including Rashmika Mandanna and Katrina Kaif, spread like wildfire.⁶ State actors have also weaponised synthetic media in cross-border

³ Kavita Chowdhury, 'Indian Government Cracks Down on Video Reels Lampooning PM Modi' *The Diplomat* (31 March 2026) <<https://thediplomat.com/2026/03/indian-government-cracks-down-on-video-reels-lampooning-pm-modi/>> accessed 02 June 2026

⁴ Information Technology Act 2000, s 79(3)(b)

⁵ Aditya Kalra et al., 'Deepfakes of Bollywood stars spark worries of AI meddling in India election' *Reuters* (22 April 2024) <<https://www.reuters.com/world/india/deepfakes-bollywood-stars-spark-worries-ai-meddling-india-election-2024-04-22/>> accessed 02 June 2026; 'Ranveer Singh files FIR over deepfake video where he is seen criticising PM Modi' *The News Minute* (23 April 2024) <<https://www.thenewsminute.com/news/ranveer-singh-files-fir-over-deepfake-video-where-he-is-seen-criticising-pm-modi>> accessed 02 June 2026

⁶ 'Rashmika Mandanna: India actress urges women to speak up on deepfake videos' *BBC* (28 November 2023) <<https://www.bbc.com/news/world-asia-india-67449421>> accessed 02 June 2026; Noor Nanji and Shruti

information disputes in the region. Yes, the danger is immediate and undeniable. But these new rules demand a more probing question: has the remedy become more dangerous than the disease?

This article proceeds in eight further parts. Part II traces the evolution of India's deepfake crisis from its origins in celebrity exploitation to its current entanglement with electoral politics and cross-border information warfare. Part III sets out the architecture of the 2026 amendment in granular detail, including its provisions on traceability, data retention, and the enhanced obligations imposed on platforms that enable the creation of synthetic content in the first place. Part IV examines the constitutional fault lines the amendment exposes, reading the new regime against *Shreya Singhal*, *Puttaswamy*, the proportionality doctrine articulated in *Modern Dental College* and refined in *Anuradha Bhasin*, and the early judicial responses the rules have already provoked. Part V interrogates the over-removal problem in greater empirical depth. Part VI widens the lens to a comparative survey of Singapore, South Korea, China, the European Union, and the United States. Part VII documents the institutional backlash from the press, civil society, and the platforms themselves. Part VIII proposes a set of concrete reforms capable of preserving the legitimate aims of the amendment while restoring the procedural safeguards it presently lacks, before Part IX concludes.

THE DEEFAKE CRISIS: FROM RASHMIKA TO RAJNITI

The evolution of India's deepfake problem is almost textbook. It started with targeting celebrities: clickbait, outrage, and sexual humiliation. In November 2023, a deepfake of actor Rashmika Mandanna superimposed her face onto a video originally posted by another woman, in a context many viewers found sexually exploitative. The video was removed within hours, but its digital footprints, screenshots, clones, and fragments proliferated anyway, demonstrating the eerie permanence of these fakes. By 2024, deepfakes had become election weapons. Videos featuring stars such as Ranveer Singh and Aamir Khan, doctored to appear to take shots at the Prime Minister and to endorse the opposition, circulated widely before fact-checkers and the actors themselves intervened. A more disturbing case followed soon after: a deepfake of Union Home Minister Amit Shah, doctored to make it appear he

Menon, 'Why are there so many deepfakes of Bollywood actresses?' *BBC* (23 December 2023) <<https://bbc.com/news/entertainment-arts-67726019>> accessed 02 June 2026

was advocating the removal of reservation rights for Scheduled Castes, Scheduled Tribes, and Other Backwards Classes, circulated widely before fact-checkers set the record straight.⁷ And as border tensions flared in 2025, synthetic media stirred further division and mistrust. With each incident, pressure mounted on regulators to do something, anything.

The speed at which deepfake incidents have surged is hard to ignore. Industry reports logged a 257% jump in deepfake cases worldwide through 2024, and in just the first months of 2025, incidents rose a further 19% compared to all of the previous year.⁸ India follows this same trajectory, but in its own way. Where once deepfake creation demanded specialised technical skills, now anyone with a smartphone can use consumer apps or browser tools, no coding required. A teenager, given twenty minutes and the right app, can make a fake video that, ten years ago, would have needed an entire visual-effects studio.

This new ease of faking things isn't without consequences. It has triggered grassroots backlash, too. In 2025, before the 2026 amendment's three-hour takedown rule, even genuine satire, unaltered and humorous, got swept up and removed by automated filters on big social platforms. Smaller, regional platforms struggled even more. Several South Indian social media apps admitted in 2024 that they simply couldn't afford the round-the-clock content moderation needed to catch synthetic media, a clear warning about what happens if a strict national rule is enforced without considering each platform's resources.

Even with the three-hour takedown rule in place, the bigger problem hasn't gone away. Just two months after the law took effect, a public interest petition in the Gujarat High Court pointed to the persistent spread of AI-generated videos online, once again framing the issue as a threat to public order and democracy itself.⁹ The continued damage, despite the world's harshest takedown policy, is telling. Speedy removal doesn't seem to stop deepfakes from

⁷ 'FIR registered after doctored video of Amit Shah goes viral' *The Week* (29 April 2024) <<https://www.theweek.in/news/india/2024/04/29/fir-registered-after-doctored-video-of-amit-shah-talking-about-reservation-goes-viral-on-social-media.html>> accessed 02 June 2026; Abhijeet Kumar, 'Delhi Police files FIR over fake video of Amit Shah's remark on reservation' *Business Standard* (29 April 2024) <https://www.business-standard.com/india-news/delhi-police-files-fir-over-fake-video-of-amit-shah-s-remark-on-reservation-124042900203_1.html> accessed 02 June 2026

⁸ Maitrayee Dey and Aruna Madrekar, 'Deepfake Statistics By Types, Fraud, Crime, Scams and Facts (2026)' (*Electro IQ*, 22 December 2025) <<https://electroi.com/stats/deepfake-statistics/>> accessed 02 June 2026

⁹ 'Gujarat High Court Seeks Response From Meta, Google, X, Reddit & Scribd On PIL Over Deepfake, AI Content; Asks Them To Onboard SAHYOG Portal' *Live Law* (15 April 2026) <<https://www.livelaw.in/amp/high-court/gujarat-high-court/gujarat-high-court-pil-deepfake-ai-generated-content-intermediaries-530410>> accessed 02 June 2026

causing harm. By the time a platform deletes the content, three hours or not, it has usually been screenshotted, reshared on encrypted apps, and worked its way into the digital fabric the takedowns were supposed to protect. So, by the time India's government stepped in with legislation, the debate had shifted from whether synthetic media needed regulation to how best to do it. The 2026 amendment offered a single answer: more speed. But does speed alone really meet the constitutional requirements of necessity and proportionality? That is the question this article turns to next.

THE 2026 AMENDMENT: WHAT THE LAW SAYS

The February 2026 amendment, effective February 20, introduced four defining changes to the IT Rules.

Defining Synthetic Content: For the first time, Indian law defines 'synthetically generated information' (SGI) as information that is artificially or algorithmically created, generated, modified, or altered using a computer resource, in a manner that appears reasonably authentic or true.¹⁰ This definition is broad by design, mirroring the technology-agnostic approach of the EU's AI Act. The result: both harmless and harmful synthetic content are swept into regulatory focus, with no distinction.

Tiered Takedown Windows: The headline change: platforms now have just three hours to remove flagged content on government or court order. For the most sensitive categories, non-consensual deepfake nudity and child sexual abuse material, the window shrinks further, to just two hours. The rationale is simple: misinformation travels at light speed in the digital era. But is such haste compatible with careful review and due process?

Neither deadline depends on how a takedown order reaches a platform. Whether a notice arrives through the government's Sahyog portal, a court order, or a direct communication from the Ministry, the clock starts the moment the platform receives it. The amendment also compresses the grievance side of the process: the previous fifteen-day window for acknowledging user complaints has been cut to seven days, with certain categories of

¹⁰ Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules 2026, r 2(1)(wa); Chandni Joshi, 'India's 2026 IT Rules Amendment: The World's First Binding Synthetic Content Provenance Mandate' (*Bhatt & Joshi Associates*, 18 February 2026) <<https://bhattandjoshiassociates.com/indias-2026-it-rules-amendment-the-worlds-first-binding-synthetic-content-provenance-mandate/>> accessed 02 June 2026

complaint requiring action the same day. Commentators tracking the rule's practical effect have observed that a change of this magnitude 'fundamentally alters compliance architecture for large platforms such as Meta, Google, and X,' requiring continuous, round-the-clock monitoring rather than scheduled review. A three-hour window is not simply a faster version of the old regime; it is a different regime altogether, one in which no human reviewer can plausibly assess every flagged item before the deadline expires, particularly once a post has begun to trend.

Civil society reacted within hours of the notification. The Internet Freedom Foundation argued that timelines this compressed eliminate any meaningful human review and described the scheme as creating a 'prior restraint regime incompatible with Article 19(1)(a)' of the Constitution.¹¹ Whether or not one accepts that characterisation in full, the underlying worry, that a platform facing a three-hour deadline will default to automated, error-prone removal rather than considered judgment, runs through much of the constitutional and empirical criticism examined in Parts IV and V below.

Mandatory Labelling and Metadata: Platforms must now clearly label synthesised or AI-generated content and preserve associated metadata.¹² The rules require that these labels remain attached and unaltered, and that metadata is retained, potentially indefinitely. Yet no technical standard for tamper-proof metadata is specified, making compliance a murky, moving target.

Safe Harbour Becomes Conditional: Previously, intermediaries enjoyed limited legal immunity ('safe harbour') under Section 79(3) of the IT Act if they followed the guidelines. The amendment turns this on its head.¹³ If a platform fails to label AI-generated content or misses a removal deadline, it loses that protection and faces full liability, treated, post hoc, as the publisher of illegal material. There is no proportionality, no margin for error.

¹¹ 'New IT Rules Violate Digital Rights, Undermine Constitutional Protection for Users: IFF' *News Click* (11 February 2026) <<https://www.newsclick.in/new-it-rules-violate-digital-rights-undermine-constitutional-protection-users-iff>> accessed 02 June 2026; 'Centre fixes three-hour deadline for social media platforms to take down flagged AI content' *Scroll* (10 February 2026) <<https://scroll.in/latest/1090639/centre-fixes-three-hour-deadline-for-social-media-platforms-to-take-down-flagged-ai-content>> accessed 02 June 2026

¹² Jolwarhring Hrangbung, 'IT Rules 2026 Deepfake Regulation: Three Hour Takedowns and AI Labelling Obligations' (*Mondaq*, 19 March 2026) <<https://www.mondaq.com/india/new-technology/1760554/it-rules-2026-deepfake-regulation-three-hour-takedowns-and-ai-labelling-obligations>> accessed 02 June 2026

¹³ 'AI & Deepfake Law in India under IT Amendment Rules 2026' (*Prime Infoserv*, 11 February 2026) <<https://primeinfoserv.com/blog-ai-deepfake-law-india-it-rules-2026-amendment/>> accessed 02 June 2026

Special Burdens on Big Platforms: Significant Social Media Intermediaries (those with more than five million users in India) must add user declaration checks before uploads and deploy technical systems to identify synthetic content before it goes public¹⁴. Non-compliance is not an option: one missed step means lost safe harbour. The investment and technical hurdles are daunting even for the largest players.

Traceability, Identity Disclosure, and Data Retention: The 2026 amendment quietly hands the government considerably more control over investigating synthetic content. The Ministry of Electronics and Information Technology no longer needs a public complaint to act: it can demand information from platforms whenever it deems necessary. Meanwhile, the Ministry of Information and Broadcasting can refer any case directly to the Inter-Departmental Committee, which can decide to block content on the spot. The government can therefore launch investigations, collect evidence, and order takedowns of deepfakes or other synthetic content without any outside complaint or independent review. It is all handled internally.

And there is more. If law enforcement believes a deepfake is tied to a crime, platforms must reveal who made it. Anonymity for deepfake creators is, in practice, gone. This traceability closes a loophole that long let people create harmful content without being caught, something victims of non-consensual fakes have been fighting for. The rules also make platforms retain data, including metadata, user logs, and records, for far longer than the old 180-day standard. In practice, this creates round-the-clock surveillance and turns private services into ready-made data vaults for the government. Civil liberties advocates say it is not far removed from direct state surveillance, just using private companies as intermediaries. With all these new powers, tracing users, retaining data, and easy government access, it is hard not to feel uneasy about privacy. It revives many of the concerns first raised in *Puttaswamy*, except that now private companies do most of the data stockpiling.

The Compliance Burden on Platforms That Enable Synthetic Content Creation: Nobody is exempt here. The amendment does not just target the big social networks: it pulls in every service dealing with synthetic content, including generative AI models, video and voice

¹⁴ Vedant Choudhary and Ankit Chugh, 'India's 2026 Amendment to IT Rules: Regulation of Deepfakes, AI Content and the Three-Hour Takedown Regime' (*SKV Law Offices*, 19 February 2026) <<https://skvlawoffices.com/indias-2026-amendment-to-it-rules-regulation-of-deepfakes-ai-content-and-the-three-hour-takedown-regime/>> accessed 02 June 2026

editors, and image-cloning apps. Every such platform must warn users about legal risks, update its terms of service, and ban certain content outright.

Consider a small Indian generative AI start-up. Competing with giants like Meta or Google was hard enough already. Now it must watermark every output, build detection tools, and issue continuous legal notices, obligations that used to be optional but are now mandatory. A single slip means losing safe harbour protections. Regulators describe the rules as aimed at big tech, but it is the smaller start-ups that are squeezed hardest. They must meet the same standards as billion-dollar players with a fraction of the resources, which makes it easier for international giants to consolidate the market while local teams falter or drop out entirely. A rule meant to level the field risks tipping it further.

CONSTITUTIONAL FAULT LINES: SHREYA SINGHAL TO SAHYOG

Indian constitutional law gives strong protection to speech and expression under Article 19(1)(a), constrained only by specific, ‘reasonable’ restrictions under Article 19(2).¹⁵ Key judicial precedents shape how these restrictions are evaluated.

Shreya Singhal v Union of India: The Supreme Court struck down Section 66A of the IT Act in 2015, ruling it vague, overbroad, and dangerously chilling to lawful speech.¹⁶ Crucially, the Court held that intermediaries should act only on government or court orders that follow due process. The new 2026 rules not only extend this power; they demand reflexive, near-automatic compliance with executive orders, transforming platforms into agents of the state.

Justice KS Puttaswamy v Union of India: Puttaswamy (2017) established privacy as a core constitutional right.¹⁷ It supports the need to regulate deepfakes that attack private life. Yet the amendment’s blanket requirements to retain metadata and labels could breach informational privacy, imposing heavy, perhaps unconstitutional, surveillance burdens on platforms and users.

The Proportionality Test: The *Modern Dental College* case locks courts into a strict four-part test.¹⁸ Restrictions on fundamental rights must be: legal; in service of a legitimate aim;

¹⁵ Constitution of India 1950, arts 14, 19(1)(a) and 19(2)

¹⁶ *Shreya Singhal v Union of India* (2015) 5 SCC 1

¹⁷ *Justice KS Puttaswamy (Retd) and Anr v Union of India and Ors* (2017) 10 SCC 1

¹⁸ *Modern Dental College and Research Centre and Ors v State of Madhya Pradesh and Ors* (2016) 7 SCC 353

necessary; and proportionate. The three-hour rule, drastic and absolute, with no margin, invites scepticism. Is such extreme speed truly required, or does it simply overwhelm intermediaries and incentivise over-removal?

X Corp v Union of India and the Sahyog Portal: The Karnataka High Court's 2025 review of the Sahyog Portal addressed a system in which platforms received flagged content with limited transparency around the reasons behind any given order, and in which users had no direct path to explain or object before removal.¹⁹ The Court ultimately upheld the Sahyog framework as a lawful due-diligence mechanism, while critics maintain that the absence of individualised reasoning or an appeal route before takedown amounts to what some commentators have called 'self-censorship by proxy': platforms comply pre-emptively rather than risk losing safe harbour.²⁰ The 2026 amendment's approach to synthetic content follows a similar structural logic.

Anuradha Bhasin v Union of India and the Doctrine of Proportionate Restraint: Commentators often invoke the four-part proportionality test from *Modern Dental College*, but it was *Anuradha Bhasin v Union of India* that pulled the concept squarely into the digital age.²¹ When the government shut down the internet in Jammu and Kashmir for months, the Supreme Court intervened in early 2020 and held that the state could not simply suspend digital communication without justification. Any restriction, the Court insisted, had to be temporary, regularly reviewed, and the least invasive option available. The judges also called for transparency: shutdown orders had to be published so that they could be challenged. To prevent the familiar device of calling something 'temporary' and then letting it persist indefinitely, the Court built a review process into its guidelines.

The 2026 amendment proceeds as though those safeguards never existed. The three-hour takedown rule is not temporary at all: as soon as the government flags something, the platform has three hours to remove it, with no discussion and no appeal. There is no

¹⁹ *X Corp v Union of India* (2025) SCC OnLine Kar 19584; Mustafa Plumber, 'X Corp Moves Karnataka High Court in Appeal against Ruling Upholding Centre's Blocking Powers through Sahyog Portal' (*Live Law*, 15 November 2025) <<https://www.livelaw.in/high-court/karnataka-high-court/karnataka-high-court-hearing-x-corp-appeal-against-central-government-blocking-orders-310049>> accessed 02 June 2026

²⁰ Harsh Gour, 'An exhaustive takedown of Karnataka HC's decision upholding Sahyog portal in X Corp v. Union of India' *The Leaflet* (06 November 2025) <<https://theleaflet.in/digital-rights/law-and-technology/an-exhaustive-takedown-of-karnataka-hcs-decision-upholding-sahyog-portal-in-x-corp-v-union-of-india>> accessed 02 June 2026

²¹ *Anuradha Bhasin v Union of India* AIR 2020 SC 1308

mechanism to test whether the underlying order even makes sense. Once a takedown notice arrives, platforms face a stark choice: comply, or risk losing safe harbour and facing litigation. The right to challenge is absent. An independent review is not even contemplated. And transparency is nowhere to be found: takedown orders are processed through the Sahyog Portal, out of public view. So, when someone like satirist Pulkit Mani finds his content missing, he has no way of learning why, and no real chance to contest it. The process remains opaque, echoing precisely the kind of unaccountable power the Karnataka High Court was asked to scrutinise in the X Corp litigation, even if that Court ultimately declined to strike the framework down.²²

Of course, Anuradha Bhasin, for all its strong language, did not restore internet access for everyone, and even its ‘independent’ review committee was staffed by government officials, the same officials who had approved the shutdowns. Still, the 2026 amendment does not even attempt that pretence: there are no checks left at all. If a system riddled with loopholes could not prevent government overreach, what hope does a system with no oversight offer? The risk of abuse is plain to see. The basic protections that Anuradha Bhasin sought to entrench are largely absent here.

The Chilling Effect Doctrine and the Problem of Anticipatory Self-Censorship: Indian constitutional law has long recognised that a law need not outright ban speech to violate the right to free expression: if a law’s design makes people too nervous to speak freely, that is enough. This ‘chilling effect’ doctrine was central to *Shreya Singhal*.²³ There, the Supreme Court objected not merely to Section 66A’s penalties but to the uncertainty it created: people could not tell what might land them in trouble next.

The 2026 amendment revives this problem in an even more acute form. The definition of ‘synthetically generated information’ is so broad that almost any digitally altered content could fall within it. The real sting is that if a platform gets it wrong and hosts questionable material, it risks losing its safe harbour protection altogether, not merely paying a fine. Given that asymmetry, the obvious move for any platform’s trust-and-safety team is to err on the side of caution and remove content at the slightest hint of risk.

²² *X Corp v Union of India* (2025) SCC OnLine Kar 19584

²³ *Shreya Singhal v Union of India* (2015) 5 SCC 1

This is not merely a theoretical worry. As Part V shows with more empirical detail, it is already happening. The crucial point is that the chilling effect operates at a distance: it is not the state directly silencing speech but the platforms doing the censoring on its behalf. Their caution, however, does not arise from nowhere. The state has designed the system to make platforms anxious enough to over-censor. Scholars call this phenomenon ‘censorship by proxy’: the government secures the results of direct censorship without risking a legal fight it might lose under Article 19(2), because private companies do the work under legal threat.

This raises a difficult question: Does this arrangement amount to ‘state action’ reviewable under the Constitution, even though a private party is technically making the call? Indian courts have not yet given a definitive answer. But as such cases accumulate, the Supreme Court is unlikely to be able to avoid the issue much longer.

Pending Constitutional Challenges: Kunal Kamra and the Gujarat High Court Petition:

The amendment’s broader regulatory architecture has already drawn fire from the courts. In *Kunal Kamra v Union of India*, the Bombay High Court struck down the 2023 amendment to the IT Rules, which had created a Fact Check Unit with broad authority to label any government-linked information as ‘fake.’²⁴ The judges found the rule impermissibly vague and held that it placed excessive pressure on free speech. The government appealed, and by March 2026, a Supreme Court bench had agreed to examine the matter, though without staying the High Court’s decision, so the Fact Check Unit rule remains, for now, unenforceable.²⁵ Kunal Kamra, the Editors Guild of India, and the Association of Indian Magazines have all supported the challenge, and have voiced particular concern about the new three-hour takedown rule’s implications for news and commentary.

A few weeks later, in April 2026, attention shifted to the Gujarat High Court, where a public interest petition took a different approach: rather than alleging censorship, it argued that the rules are not being enforced firmly enough, particularly against AI-generated content. The

²⁴ *Kunal Kamra v Union of India* 2024:BHC-OS:1575-DB; ‘Bombay High Court Allows Kunal Kamra’s Plea, Strikes Down 2023 Amendment To IT Rules On ‘Fact Check Units’ (*Live Law*, 26 September 2024) <<https://www.livelaw.in/top-stories/bombay-high-court-fact-check-unit-it-rules-kunal-kamra-270722>> accessed 02 June 2026

²⁵ Naman Kumar, ‘Supreme Court of India issued notice but refused any stay in Union of India’s challenge to the Bombay High Court judgement holding Fact Check Unit to be ultra vires’ (*Internet Freedom Foundation*, 11 March 2026) <<https://internetfreedom.in/supreme-court-of-india-issued-notice-but-refused-any-stay-in-union-of-indias-challenge-to-the-bombay-high-court-judgement-holding-fact-check-unit-to-be-ultra-vires/>> accessed 02 June 2026

Court issued notices to several major intermediaries and sought details on their due diligence practices. Read together, the two cases reveal a genuine tension: one camp argues the government has too much control over speech, while the other contends it is not doing enough to address harmful or misleading content online. How the Supreme Court ultimately resolves the Fact Check Unit appeal will shape how India regulates deepfakes for years to come, since both cases turn on the same underlying questions: when does regulation go too far, and are the rules even clear enough to apply consistently?

THE OVER-REMOVAL PROBLEM: WHEN THE CURE BECOMES CENSORSHIP

Commercial Logic Shapes Speech: Faced with the threat of criminal liability, platform legal teams run the numbers, and the lowest-risk option is to take down anything even suspected of being flagged. Satire, parody, and investigative journalism- anything edgy or controversial- can end up sacrificed. The better-safe-than-sorry mindset prevails.

Blurring the Line Between Real Harm and Creativity: The amendment's broad SGI definition does not distinguish between malicious deepfakes and clear, legitimate satire or artistic work. This means that well-crafted political parody, which should enjoy strong constitutional protection, can be swept away alongside genuine synthetic deception. In March 2026, several political satire accounts were removed under the new regime, even though everyone understood the content to be comedy rather than deception.

Who Judges What is 'Synthetic': Platforms are now expected to deploy 'technical verification measures' that reliably detect synthetic content. The reality is that current detection technology still struggles to distinguish high-quality deepfakes from genuine footage. The government, by default, ends up deciding what counts as 'synthetic,' but offers few clear standards. Civil-society monitors have warned that these risks are widespread, sometimes indiscriminate, scrutiny of lawful, dissenting, or controversial speech.

Empirical Evidence of Erroneous Takedowns: Over-removal is not merely a theoretical concern: it is playing out in real time. Even before the 2026 amendment and its strict three-hour removal rule, automated content filters on major video platforms were already making mistakes, banning unedited satire on the mistaken assumption that it was synthetic media. Industry observers had been warning about this as early as 2025. The new three-hour deadline does not fix that underlying problem. If anything, cutting the time available for

review, it makes such mistakes more likely. There is barely a window for thoughtful, case-by-case human review, especially when the content is not in English. Most detection systems are trained primarily on English-language data, so accuracy drops sharply for content in any of India's twenty-two official languages.

And the costs of these mistakes do not fall evenly on both sides. If a platform fails to take down a genuine deepfake, it risks losing safe harbour and facing possible criminal charges. But if it wrongly censors a piece of satire, the worst likely consequence is an angry user or some unfavourable press, and little else. Platform compliance teams understand this asymmetry. The incentive is to remove content first and ask questions later, not the reverse. The three-hour rule does not merely permit over-removal: it creates precisely the incentive structure that guarantees it.

The Disproportionate Burden on Vernacular Media and Marginalised Voices: These mistakes do not affect everyone equally. English-language political commentators based in Delhi or Mumbai, with large audiences, legal representation, and media connections, can push back when their content is taken down. They can call in favours, attract attention, or contact platform representatives directly. That is rarely true for an independent satirist working in Assamese or Kannada, with no legal backing, no institutional clout, and no direct line to the platform. For them, a removed video is often the end of the road, with no practical route back.

This is doubly concerning given that tools for detecting synthetic media are built predominantly around English and Hindi. Regional-language content is left correspondingly exposed: detection models either flag genuine deepfakes too slowly or mistakenly delete innocuous material. The amendment's legal requirements claim neutrality, but in practice, they entrench these technical disadvantages, forcing platforms to apply standards that simply do not work well for much of India. There is no requirement to invest in better detection for regional languages, so the bias becomes embedded in law.

If lawmakers genuinely wanted to protect India's most at-risk voices, including Dalit creators, women journalists reporting from conflict zones, and regional satirists, they would need to acknowledge this bias and address it directly. Treating the rule as one-size-fits-all

sweeps the problem under the rug and leaves those who most need protection out in the cold.

GLOBAL COMPARISON: EU, US, AND INDIA'S REGULATORY LEAP

The European Union: The EU's AI Act and Digital Services Act (DSA) both regulate synthetic content, but with clearer safeguards: exemptions for satire, art, and fiction; a notice-and-action standard requiring removal 'without undue delay' rather than a single rigid statutory deadline; and required explanations and appeal rights for affected users.²⁶ In practice, the EU has long associated swift removal with a roughly 24-hour benchmark, originating in the 2016 voluntary Code of Conduct on hate speech and now broadly reflected in how the largest platforms report their response times under the DSA.²⁷ The Act also carves out evidently artistic, satirical, or fictional content from certain labelling duties.²⁸ The DSA is demanding, but it remains built on procedural fairness. In practice, the EU's enforcement actions, including its proceedings against X under the DSA, show how rules can address genuine harms while limiting collateral censorship.

The United States: The US approach is fragmented, resting on sweeping platform immunity under Section 230 of the Communications Decency Act.²⁹ Federal deepfake legislation has stalled amid political gridlock; only a handful of states have moved, with California notably requiring disclosure of election-related synthetic media.³⁰ But no federal law matches either the EU's or India's ambition, and no American platform faces a three-hour deadline or automatic liability for missed compliance.

India's Regulatory Leap: No OECD democracy has moved as fast or as far as India's 2026 amendment. Its three-hour removal window is unprecedented. Its liabilities are close to absolute, with little proportionality and few exceptions. Its definition of synthetic content

²⁶ EU Digital Services Act 2022, art 16

²⁷ 'The EU Code of conduct on countering illegal hate speech online' (*European Commission*) <https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combating-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en> accessed 02 June 2026; Dawn Carla Nunziato, 'The Digital Services Act and the Brussels Effect on Platform Content Moderation' (2023) 24(1) *Chicago Journal of International Law* <<https://cjl.uchicago.edu/print-archive/digital-services-act-and-brussels-effect-platform-content-moderation>> accessed 02 June 2026

²⁸ EU Artificial Intelligence Act 2024, art 50(4)

²⁹ Communications Decency Act 1996, s 230

³⁰ Stuart D Levi et al., 'California Enacts New Laws to Combat AI-Generated Deceptive Election Content' (*Skadden*, 27 September 2024) <<https://www.skadden.com/insights/publications/2024/09/california-enacts-new-laws>> accessed 02 June 2026

sweeps in everything from art to satire to protest. The government hails this as proactive. Critics call it a constitutional gamble, and in all likelihood, a losing one.

Singapore’s Calibrated Model: Targeted Criminalisation over Blanket Liability: Singapore takes a markedly different route from India when it comes to tackling synthetic media. The difference is not one of seriousness but of precision. India covers almost everything under a broad, content-neutral approach; Singapore focuses on clearly defined harms. Under the Protection from Online Falsehoods and Manipulation Act (POFMA), government ministers can demand corrections or takedowns of false deepfake content, particularly where elections or national security are at stake, with non-compliant platforms facing substantial fines.³¹ Singapore’s 2025 Criminal Law reforms went further still, expanding the statutory definition of ‘prohibited intimate image’ to capture wholly AI-generated material, closing a loophole that had let synthetic, non-consensual content slip through.³²

What distinguishes Singapore is not a lack of seriousness; its measures are just as forceful as India’s, but far more specific about what they target. POFMA’s correction and takedown orders apply only to particular threats, such as electoral manipulation and national security disinformation, rather than policing all AI-generated content regardless of context or intent. The deepfake ban introduced for Singapore’s 2025 general election, for instance, applies only during the official campaign period, covers only digitally manipulated content that is convincingly false and deceptive, and expressly leaves satire and parody untouched.³³ India’s SGI definition draws no such lines: a comedian’s sketch and a doctored video designed to incite violence are swept into the same category.

South Korea’s AI Basic Act: Labelling as the First Resort, Not the Last: With its AI Basic Act, which took effect in January 2026, South Korea became one of the first countries to pass

³¹ Protection from Online Falsehoods and Manipulation Act 2019 (Singapore); Elections (Integrity of Online Advertising) (Amendment) Act 2024 (Singapore)

³² Criminal Law (Miscellaneous Amendments) Bill 2025 (Singapore); Jonathan Kok, ‘Deepfakes in Singapore: Legal Recourse, Regulatory Powers, and Building Long-Term Corporate Resilience.’ (*Conventus Law*, 23 January 2026) <<https://conventuslaw.com/report/deepfakes-in-singapore-legal-recourse-regulatory-powers-and-building-long-term-corporate-resilience/>> accessed 02 June 2026

³³ Elections (Integrity of Online Advertising) (Amendment) Act 2024 (Singapore); Andy Leck et al., ‘Singapore: Ban on the publication, boosting, sharing and reposting of deepfake content depicting election candidates comes into effect’ (*Connect On Tech*, 28 July 2025) <<https://connectontech.bakermckenzie.com/singapore-ban-on-the-publication-boosting-sharing-and-reposting-of-deepfake-content-depicting-election-candidates-comes-into-effect/>> accessed 02 June 2026

a comprehensive national AI law.³⁴ Like India, South Korea requires clear labelling for AI-generated content that is difficult to distinguish from the real thing, and advance notice from companies using generative AI in their products. Korea enforces these rules differently, however: breaching the labelling requirement triggers an administrative fine of up to thirty million won, a little over twenty thousand US dollars, enough to sting but not to ruin a business.³⁵

Labelling is not where the framework ends, either. South Korea has also strengthened criminal penalties in specific harm areas. In 2023, lawmakers raised the maximum prison sentence for non-consensual deepfake pornography from five to seven years and made it possible to be prosecuted simply for possessing or viewing such material, a step most countries have not taken. The broader lesson here is structural rather than substantive. South Korea did not attempt to fold every kind of AI-related harm- electoral disinformation, personal attacks, and ordinary commercial use alike- into a single liability framework. Instead, it built a layered, harm-specific structure: broad labelling for transparency, severe penalties for serious violations, and enough flexibility to calibrate the response each time. India's 2026 approach, by contrast, channels almost every type of AI-related harm- doctored political speeches, deepfake pornography, and comic impersonation alike- through the same narrow 'safe harbour' filter, with the same three-hour deadline applied to all of them.

China's Top-Down Model: A Cautionary Comparator for a Constitutional Democracy:

China, through its Cyberspace Administration, takes a different path entirely. It requires labelling and user identification for generative AI, but does so within a framework where platforms function as arms of the state's information-control system.³⁶ Whether by design or not, India's new rules now resemble this model in important respects. Mandatory labelling, traceability back to a content creator, and takedown orders that cannot be challenged or

³⁴ AI Basic Act 2024 (South Korea); 'South Korea rolls out world's first full AI law, forcing labels on deepfakes' (*Malay Mail*, 22 January 2026) <<https://www.malaymail.com/news/world/2026/01/22/south-korea-rolls-out-worlds-first-full-ai-law-forcing-labels-on-deepfakes/206409>> accessed 02 June 2026

³⁵ Sam Jungyun Choi et al., 'Global AI Governance Law and Policy: South Korea' (*IAPP*, 27 August 2025) <<https://iapp.org/resources/article/global-ai-governance-south-korea>> accessed 02 June 2026

³⁶ 'Provisions on the Administration of Deep Synthesis Internet Information Services' (*China Law Translate*, 11 December 2025) <<https://www.chinalawtranslate.com/en/deep-synthesis/>> accessed 02 June 2026; 'Interim Measures for the Management of Generative Artificial Intelligence Services' (*China Law Translate*, 13 July 2023) <<https://www.chinalawtranslate.com/en/generative-ai-interim/>> accessed 02 June 2026

appealed are powerful tools, but they sit uneasily alongside the free speech protections the Indian Constitution is meant to guarantee.

The crucial difference is that China's constitution does not promise free speech in the way Article 19 does, and administrative orders there are not subject to the judicial review that Indian law, at least on paper, requires. This gap matters. India cannot simply borrow China's enforcement tools while ignoring the context in which they operate. If India deploys the same machinery, even for legitimate reasons, it owes the public an answer as to whether it is also imposing the same clampdown on dissent and unpopular speech. When a constitutional democracy adopts the infrastructure of a one-party state to control information, it must account for whether that structure is consistent with its constitutional obligations, not merely its stated policy goals.

Synthesis: What the Comparative Survey Reveals: Looking across these jurisdictions, a pattern emerges. Where countries act quickly while preserving public trust, as in the European Union, they confine their strongest tools to clearly defined harms, ensure art and satire are not caught in the net, and preserve some form of due process even on tight timelines. Where enforcement is broad and undifferentiated, as in China, those safeguards disappear, along with the protections at the core of the Indian Constitution. India's 2026 amendment occupies an uneasy middle ground. It carries the urgency and assertiveness of the strictest models without their targeted focus, and it lacks the constitutional flexibility to dispense with judicial oversight altogether. That is India's real problem: speed without nuance, severity without the structure the Constitution is meant to supply. The combination leaves India's current approach looking uniquely precarious compared to the rest of the world.

THE BACKLASH: OPPOSITION FROM PRESS, CIVIL SOCIETY, AND PLATFORMS

Journalists and Press Outlets: In the months following the amendment's notification, leading media organisations, including the Editors Guild of India, DIGIPUB, and major women's journalist networks, came together to demand withdrawal of the rules and

suspension of the Sahyog Portal, insisting that due process and free expression cannot be made hostage to executive fiat.³⁷

Civil Society and Rights Groups: Digital rights organisations have been similarly blunt, describing the framework as enabling broad surveillance and wholesale filtering of online speech, and warning that it sits uneasily with India's international commitments under the ICCPR, which requires legality, necessity, and proportionality in any restriction on expression.

Platforms Struggle to Keep Up: Technology companies have struggled to keep pace. A three-hour compliance window effectively requires round-the-clock, automated monitoring teams, or a default posture of mass takedown. Industry commentary has repeatedly framed the choice in similarly stark terms: comply nervously and over-remove, or risk existential legal exposure.

Parliamentary and Academic Scrutiny: It is not only journalists, activists, and technology companies who have raised concerns: members of Parliament have pressed the government on why there is no independent mechanism to review takedown orders. They have pointed out that, unlike internet shutdowns, which now require an independent review committee following Anuradha Bhasin, takedown decisions under the new rules face no equivalent oversight. The parliamentary standing committee examining the amendments has yet to resolve sharp disagreement over whether the Ministry meaningfully consulted industry and civil society, given that the draft rules were circulated for public comment only weeks before final notification in February 2026.

On the academic side, a growing number of legal scholars argue that the amendment misses a genuine opportunity. Rather than finding a better balance for addressing synthetic media, it simply makes the response faster, perhaps too fast. Commentators have noted the irony that the law treats a three-hour delay as a crisis while giving a green light to takedown orders that no one can review or appeal. Why, then, is there not equal concern for due process as there is for speed? With Parliament, academia, and civil society all raising similar concerns

³⁷ 'PCI passes resolution that attacks 'infrastructure of censorship' built by Modi govt' *News Laundry* (11 April 2026) <<https://www.newslaundry.com/amp/story/2026%2F04%2F11%2Fpci-passes-resolution-that-attacks-infrastructure-of-censorship-built-by-modi-govt>> accessed 02 June 2026

so soon after the rules took effect, it is increasingly clear that the amendment's deepest problem lies not in what it aims to do, but in the absence of any check on how it does it.

TOWARDS A PROPORTIONATE FRAMEWORK: RECOMMENDATIONS FOR REFORM

The deepfake problem is real and urgent; ignoring it is not an option. But the central argument here is not that nothing should be done. It is that the way the 2026 amendment is designed discards constitutional safeguards in pursuit of a speed that may not even achieve its stated goal. A better-calibrated system could protect people from real harm without abandoning basic fairness. Drawing on what works elsewhere, five reforms stand out.

An Independent Review Mechanism: The most important change would be an independent body, sitting outside the ministries that order takedowns, empowered to review disputes quickly and within the tight timeframes the law already imposes on platforms. This need not slow down obvious cases such as non-consensual intimate images or blatant election disinformation. Even a post-facto system, in which the body audits a sample of orders and gives ordinary users, satirists, and journalists a genuine opportunity to appeal after the fact, would mark a significant improvement. It could allow mistakenly removed content to be restored and errors to be made public. Without anything resembling this, the amendment rests on shaky constitutional ground.

Graduated, Harm-Specific Timelines: Not all harmful content is the same, and the law should stop treating it as though it were. India should follow South Korea's example and set different timeframes for different categories of harm rather than imposing a single rule across the board. Non-consensual intimate images genuinely become more damaging with each passing hour and should be removed quickly. But political satire, parody, or contested statements about public figures, where mistaken removal means a real loss of free expression, deserve a longer window, comparable to the EU's notice-and-action standard, together with mandatory explanations for affected users.

Statutory Carve-Outs for Satire, Parody, and Journalism: The amendment needs an explicit carve-out. Content that any reasonable person would recognise as satire, parody, or clearly labelled commentary should not be at risk of removal. The EU's Digital Services Act spells out such an exemption, and it works. Without one, cases like Pulkit Mani's reel, classic satire

caught up in a regime that defines synthetic content by technology rather than by context or intent, will keep recurring.

Proportionate Compliance Obligations Calibrated to Platform Scale: Another essential reform: a small Indian start-up should not be made to clear the same hurdles as a global giant with vastly greater resources. The EU addresses this through a tiered system for ‘Very Large Online Platforms,’ scaling obligations to size and capacity. Where every platform faces identical, blanket rules, large foreign companies tend to win, small Indian companies tend to lose, and the goal of building a strong domestic AI industry suffers.

Investment in Indian-Language Detection Infrastructure: Finally, if the government wants platforms to act quickly and accurately, it cannot ignore India’s language gap. Many detection tools still do not cover regional languages, leaving creators who work in those languages disproportionately exposed to error. The remedy is to fund detection technology for major Indian languages publicly and support start-ups building such tools, paired with public awareness campaigns that help people recognise synthetic content themselves. That approach builds lasting trust and is likely to prove far more effective than repeatedly shortening takedown deadlines.

CONCLUSION

The deepfake crisis is real; no one disputes that. But the government’s chosen instrument is blunt, hurried, and constitutionally dangerous. The three-hour takedown rule hands the executive immediate, unsupervised erasure power, sidestepping courts, process, and even the minimum requirements of notice or hearing.

Pulkit Mani’s erased satire was not the collateral damage of a technical mistake. It was the inevitable outcome of empowering the government to silence anything flagged, swiftly and without regard to its truth or value. When a system confuses the urgent need to fight lies with the urge to erase dissent, it enables precisely what a healthy democracy must guard against: the suppression of inconvenient or critical speech.

The Constitution permits restrictions on speech, but only where they are necessary, proportionate, and subject to independent review. The 2026 amendment cannot meet that test. It is reckless in its speed, deficient in its safeguards, and gives risk-averse platforms

every incentive to censor broadly and reflexively, leaving democratic debate thinner, poorer, and more fearful.

India's drive to curb deepfakes and shield individuals from digital harm is a righteous one. But the current rules go too far, too fast, risking the very freedoms they claim to protect. The Supreme Court will inevitably have to weigh in, likely soon. When it does, the stakes will be enormous: not simply whether deepfakes can be controlled, but whether a democracy can truly defend itself by giving the state the power to erase the truth along with the lies. If constitutional democracy is to endure, the answer must be no.