DEEPFAKES, DUE DILIGENCE AND THE GOOD SAMARITAN PARADOX

How India's 2026 IT Amendment Rules Resolve the Global Platform Liability Debate

Authored by Gagan Verma (Partner – Kochhar & Co.) and Mahima Wahi (Principal Associate at Kochhar & Co.)

Published in LiveLaw

The Global Problem Nobody Has Solved

There is a question that haunts every regulator trying to govern AI-generated content. When a platform voluntarily takes down a deepfake, does that editorial judgment cost it the very intermediary immunity that allows it to function? In the US, the debate has raged under Section 230 of the Communications Decency Act, 1996 (47 U.S.C. § 230), which shields platforms from liability for user content. The catch? Courts remain split on whether active content moderation amounts to "publishing", which would void the shield. Two jury verdicts against Meta in March 2026 (K.G.M. v. Meta in Los Angeles and New Mexico v. Meta) held that algorithmic curation was not protected, and a bipartisan Sunset Section 230 Act (S.3546) was introduced in December 2025 seeking outright repeal. Congress remains deadlocked. Europe took a different route with the Digital Services Act's "Good Samaritan" carve-out, but its overlay with the EU AI Act transparency mandates is still being worked out by practitioners. China's Cyberspace Administration issued its Deep Synthesis Provisions in November 2022 (effective January 2023), mandating labelling, real-name registration, and takedown obligations. But China's regulatory apparatus is state-directed in a way that does not translate to market economies with independent judiciaries.

India has cut through this knot. The IT (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026 (G.S.R. 120(E), 10.02.2026, effective 20.02.2026) tackle the problem head-on through a new Rule 2(1B). The provision says, plainly, that if an intermediary removes content (including SGI) while following these Rules, that act of removal will not be held against it under Section 79(2)(a) or (b) of the IT Act, 2000. It does not matter whether the removal was triggered by a human reviewer or by an automated classifier. The drafters were clearly alive to the Section 230 trap. They chose to close it off legislatively rather than leave it to courts to sort out. A platform that invests in building a deepfake detection engine will not find that investment cited against it in safe harbour litigation. No other large jurisdiction, to our knowledge, has said this with equivalent directness.

What Exactly Changed On February 20, 2026

The amendment introduces "synthetically generated information" or SGI (Rule 2(1)(wa)) as a defined term. It covers audio, visual, and audio-visual content that has been created or altered through AI or algorithmic processes and that looks convincingly real. Think deepfakes, voice

clones, AI-fabricated video. The definition is intentionally broad. But a proviso protects routine, legitimate work. Colour correction, noise reduction, file compression, and similar everyday edits are excluded, provided they do not distort the substance of the original. Similarly, using AI tools to prepare presentations, training decks, or research papers is fine, as long as no false document or false electronic record is produced in the process. And technology used purely for accessibility, translation, or searchability is carved out as well. These exclusions matter enormously for IT companies whose engineering teams live inside generative tools every day.

Intermediaries now face three distinct sets of obligations, each with teeth. Under the new Rule 3(3), any platform or tool that lets users create or share SGI must put up guardrails. Specifically, it must deploy reasonable technical safeguards to block four categories of unlawful synthetic content before they ever reach an audience. Those four categories are worth spelling out. Child sexual exploitative and abuse material, and non-consensual intimate imagery. Fabricated documents or fabricated electronic records. Content connected to preparation or procurement of explosives, arms, or ammunition. And finally, synthetic depictions that put words in a real person's mouth, fabricate their actions, or manufacture events that never took place, all with the intent to deceive. For SGI that is lawful, the obligation is different but equally demanding. It must carry a visible label, and it must embed metadata containing a unique identifier that traces back to the platform that produced it. Rule 3(3)(b) goes one step further by prohibiting the intermediary from offering any tool or feature that lets users strip out that label or metadata. **A separate** set of obligations applies to Significant Social Media Intermediaries under a newly introduced Rule 4(1A), and it changes the compliance posture fundamentally. These platforms can no longer wait for complaints. The uploader has to declare, upfront, whether AI was used. The platform verifies. If confirmed as SGI, a label goes on before publication. This is gatekeeping at the front door, not cleanup after the damage is done. **Finally, and** this is where platforms will feel the pinch most immediately, the response timelines have been slashed across the board.

Trigger	Earlier	Now	Impact
Takedown on court order or government notice	36 hours	**3 hours**	Round-the-clock operations become non-negotiable
Intimate or impersonation content	24 hours	**2 hours**	Needs automated flagging with immediate human review
General grievance resolution	15 days	**7 days**	Legal, product, and trust teams must coordinate faster
Content removal requests	72 hours	**36 hours**	Triage workflows need redesign
User awareness communications	Annually	**Quarterly**	Must be woven into the product experience itself

Why This Matters More Than You Think

Consider the dilemma a general counsel at a global platform faced until February 19, 2026. Her trust and safety team has already built a classifier. It catches most non-consensual deepfakes within seconds. Deploying it would protect victims. But her outside counsel warns her that under Indian law, voluntarily scanning and removing content could be read as the exercise of editorial discretion, potentially undermining the platform's claim to safe harbour under Section 79 of the IT Act. The rational commercial decision, perversely, was to do less rather than more.

Rule 2(1B) dissolves that dilemma overnight. By providing an express statutory carve-out, the amended Rules tell that general counsel, in unmistakable terms, that her platform will not be penalised for being a responsible actor. She can deploy the classifier, integrate watermark-detection pipelines, invest in metadata forensics, and build automated escalation protocols,

all without any risk that these measures will be weaponised against the company in a safe harbour challenge. For any technology business assessing where to allocate its next compliance budget, this provision should be the starting point of the analysis.

What The IT Sector Should Do Now

Companies offering generative AI tools or SaaS products with content creation features have no room for a wait-and-watch approach. Rule 3(3) is unambiguous. The product itself must do the compliance work. In practice, that means a classifier at the generation layer catching problematic outputs before they ship. It means provenance data (C2PA or equivalent) embedded in every file so the output carries an auditable trail. And it means the end user cannot strip out the label or scrub the metadata. Retrofitting this after launch is a recipe for trouble. Build it into the architecture from day one.

Platforms hosting user-generated content at scale face perhaps the steepest operational adjustment. Rule 4(1A) demands that they collect a user declaration before publication, run technical verification against that declaration, and apply a visible SGI tag where confirmed. Add to this the 2-hour window for taking down intimate or impersonation content and the 3-hour window for court-ordered or government-notified takedowns, and the practical implication is clear. These platforms need dedicated India compliance teams, on-call around the clock, backed by detection systems tuned to Indian legal thresholds. The proviso to Rule 4(1A) sharpens the stakes further. If the platform knows about unlawful SGI and does nothing, it is deemed to have failed its due diligence. Wilful blindness is now, in effect, a statutory offence.
IT outsourcing and services firms should recognise that the amendment creates a new category of client demand. Companies that moderate content, manage platform operations, or build compliance tooling on behalf of intermediary clients will need to recalibrate their service delivery to the compressed timelines. At the same time, there is a real commercial opportunity here. Automated moderation solutions, metadata-embedding services, and compliance monitoring dashboards are going to be in significant demand across the Indian market.

On the people side, every technology employer in India needs to revisit its staffing model. The Rules continue to require an India-resident Chief Compliance Officer, a nodal contact person, and a Resident Grievance Officer under Rule 4. With response windows now measured in single-digit hours, maintaining adequate on-call rosters and internal escalation chains is no longer a best practice recommendation. It is a condition of retaining safe harbour, because Rule 7 is explicit. An intermediary that fails to observe these Rules loses the protection of Section 79(1).

The Bottom Line

Here is the scorecard, as we see it. The US remains gridlocked on Section 230 reform after thirty years. The EU's AI Act sits alongside the DSA and DMA in a stack that even Brussels specialists find hard to apply as a coherent whole. China's model works within its own system but is not exportable. India, meanwhile, has done something pointed and practical. It identified a specific problem (synthetic content that deceives), built enforceable obligations around it, and offered a credible incentive for compliance through safe harbour protection.
The amendment does not try to govern artificial intelligence as a concept. It picks one fight. Synthetic content that deceives. And it builds around that fight a practical regime of prevention, labelling, traceability, and rapid response. Rule 2(1B) sits at the centre of it, offering a bargain that is easy to understand. Comply, and the law stands behind you. Fail to comply, and Section 79 protection falls away. Regulators elsewhere, still searching for the right balance between platform accountability and innovation, may find this worth examining closely.