WITNESS Submission on the Draft Amendments to the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 — in relation to “Synthetically Generated Information”

Submitted to:

Ministry of Electronics and Information Technology (MeitY)
Government of India

Date of submission:
6 November 2025

Submitted by:
WITNESS
An international human rights organisation supporting people to use video and technology to protect and defend human rights

Contact:
Technology, Threats and Opportunities Team

tto@witness.org

Executive Summary

WITNESS welcomes the opportunity to comment on the Ministry of Electronic and Information Technology (MeitY) Draft IT Amendment Rules. WITNESS appreciates the Indian Government's intent to address the growing challenges of AI-generated and synthetic content and the need to promote transparency and accountability. Using our experience working on the intersections of AI transparency, provenance and human rights for close to a decade, we believe, while this framework has noble intentions, it is too broad and platform-centric to achieve genuine accountability. India should pursue a dedicated, risk-based transparency framework that is aligned with emerging international best practices.

We offer the following five recommendations and revisions to these rules, while maintaining our view that a different framework would be more effective

1. We recommend narrowing the definition of "synthetically generated content" to focus on intent to deceive and exempt benign, assistive AI uses;
2. Deleting Rule 2(ii) to avoid unconstitutional overreach by moderating "synthetic content" under unlawful acts provisions;
3. Ensuring due-process safeguards for content takedowns under Rule 3;
4. Replacing the impractical "10 percent overlay" in Rule 4 with less prescriptive, but still clearly visible watermarks, such as icons, or other clear markings that point to latent disclosure such as tamper-evident, rights-respecting provenance metadata built on open standards;
5. Withdrawing Rule 5's unworkable automated verification requirement, as AI detection is not an exact science.

While these interventions represent the minimum necessary improvements for this current draft, India's long-term goal should focus on establishing the process and framework for a comprehensive and interoperable AI Transparency regulation that embeds accountability across the AI lifecycle (or pipeline) while protecting privacy, freedom of expression and innovation.

WITNESS Executive Director Sam Gregory notes: "We cannot regulate AI by only regulating intermediaries; we must build the infrastructure of trust through transparency, provenance and accountability through every stage of the AI pipeline."

Table of Contents

1. MeitY's opportunity to align and lead on global AI transparency

WITNESS is concerned that India's Ministry of Electronic and Information Technology (MeitY) has entered the domain of AI regulation through amendments to the existing Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 rather than by developing a dedicated, comprehensive AI framework. The current Rules derive their authority from the Information Technology Act, 2000, and while they have the force of law, their validity depends on remaining within the boundaries of that parent statute. They are built on a platform-centric model that defines intermediary "due diligence" and the conditions of safe-harbour under Section 79 of the IT Act. Extending this framework to encompass the governance of generative AI and synthetic media, without articulating a broader vision for pipeline responsibility or the roles of AI developers, deployers, and users, risks creating an overbroad and weak regulatory instrument. In its present form, the amendment converges on content-removal obligations and platform enforcement rather than fostering meaningful content transparency.

The dangers of an overbroad amendment like this can mean chilling legitimate expression, particularly among independent journalists, creators, artists and civil-society that rely on digital platforms to publish their work. AI-assisted tools -especially for those under-resourced -is integrated into editing, captioning, translation and other legitimate and non-manipulative ways for aiding in content creation. As the MeitY draft stands such content can be categorised as "synthetically generated information" and subject to mandatory 10% labelling or even takedowns if platforms over-comply. These measures will stigmatise legitimate use of technology and disproportionately harm voices outside of mainstream media, hampering access to information diversity.

WITNESS urges MeitY to reconsider these frameworks in order to avoid the chilling effect these blunt labelling and overbroad enforcements will have. AI transparency needs to be emphasized as a matter of *process accountability* and not *content control*, as seen by international recognised frameworks within the EU AI Act (Article 50 on Transparency) and California's AI Act (see AB853). These frameworks underscore the importance of distinguishing between harmful manipulation versus benign and assistive uses of AI. The current labelling framework of this amendment does an injustice to India's position as a global AI and technology leader. India should embrace the technical sophistication of provenance and interoperable standards such as the Coalition for Content Provenance and Authenticity (C2PA). These mechanisms ensure risk-based, context-sensitive approaches that empower users, promote accountability while preserving privacy, artistic creation, satire and freedom of expression.

2. WITNESS' global experience on content provenance and transparency

WITNESS approaches this consultation from over a decade of work on media authenticity, provenance and transparency standards in the context of human rights and information integrity. WITNESS launched the first global mechanism to analyse deepfakes through the Deepfake Rapid Response Force.[1] WITNESS has spent decades working globally to ensure all work we do is informed by the needs of frontline communities, especially in the global majority.[2] Through our leadership in pushing for the Coalition for Content Provenance and Authenticity (C2PA)[3] and related multistakeholder initiatives to set interoperable standards based on fundamental rights, WITNESS has helped shape global approaches to signalling how digital content is created, edited, or generated by AI while protecting privacy, anonymity, and freedom of expression. Our engagement with transparency frameworks, ranging from the EU AI Act's Article 50 to emerging disclosure rules in the United Kingdom, Brazil and jurisdictions such as California emphasize that provenance and transparency should always work in service of users and creators rather than expose them to surveillance or arbitrary and overbroad moderation (or censorship). WITNESS' advocacy always emphasizes three principles:

1. **Process transparency over content control:** users should understand *how* content was generated or altered and move away from the false and misleading binaries of "AI" or "not AI," which leads to the wrong framework of "real" or "not".
2. **Privacy preserving provenance:** any and all transparency mechanisms must protect privacy.
3. **Interoperability and proportionality:** technical and legal frameworks must work across platforms and borders, with obligations proportionate to the actor's role in the AI pipeline (developer, deployer, or intermediary).

India is well positioned to emerge as a leading voice in how to approach synthetic content made with or assisted by generative AI if it builds thoughtful and carefully defined principles around interoperability, proportionality, and human-rights based transparency.

[1] Read about WITNESS' Deepfake Rapid Response Force here.
[2] Read about WITNESS' region specific consultations here.
[3] On our work in C2PA, see our report "Embedding Human Rights in Technical Standards."

3. Global precedents and good practices

There is a growing convergence around three core principles when it comes to regulating synthetic and generative AI content: **proportionality, pipeline responsibility and interoperability**. This has been a trend that WITNESS has contributed to and welcomed as a step towards both rights-respecting regulation and technical feasibility. Platform-centric moderation and fixed labelling mandates directly counter this trend that needs process-based disclosure and provenance infrastructure. Now is the time for India to align with the global consensus and become a leader in developing and enforcing a transparency framework for AI-generated and AI-assisted content that's grounded in fundamental rights and feasibility. Here we look at three regulations WITNESS has contributed to.

European Union: Article 50 of the EU AI Act on Transparency

Article 50 of the EU Artificial Intelligence Act (2024) identifies responsibility with deployers[4] to create clear disclosures for artificially generated or manipulated content only when it is likely to mislead a person into believing it is authentic. This is only applied to content that poses a realistic risk of deception or harm with flexibility in what this disclosure can look like -including uses of water-marking, labelling and other contextual signals. Unlike what MeitY has proposed, there are no quantitative requirements for these labels and there are the necessary exemptions for artistic, satirical and journalistic expression for when the use of AI is evident or otherwise disclosed. What is most important about Article 50 is that it won't be enforceable until August 2026 when the accompanying Code of Practise on Transparency is finalised and adopted.[5] The process and thought going into the enforcement of the central AI Transparency regulation in the EU is one we urge the Indian government to follow. The current Code of Practice is being worked on by a multi-stakeholder working group, of which WITNESS is a part of, that includes civil society, media and the private sector. This multi-stakeholder approach helps bridge practicality and feasibility with risk mitigation and fundamental rights.

The development of AI transparency regulation in the EU underlines the need for stakeholder consultation, interoperability and technical input. We urge India to move away from its current intermediary liability framework towards a measured, evidence based regulation worthy of a democracy with a thriving and sophisticated technology ecosystem.

California's Approach to AI Transparency: A Layered Approach with AB853 Central

In the United States and globally, California has been leading with its frameworks for AI accountability and transparency. It has a suite of laws sometimes referred to as "California's AI governance package". This approach has ensured that California does not fall victim to the narrow platform enforcement approach, but one where they can embed transparency and responsibility across the AI lifecycle and avoid blanket bans. Bills such AB2013 on AI Transparency and Political Communication (2024) and several different complementary laws that create a civil and criminal framework for non-consensual, sexually

[4] Those who generate or manipulate image, audio, text, or video content using AI systems

[5] See WITNESS' submission to the EU consultation on the Code of Practice here. And WITNESS' position upon submission here.

explicit deepfakes demonstrate targeted, risk-specific transparency and accountability frameworks that endeavour to apply disclosure and removal duties in a narrow and clear way for high-risk and harmful contexts (i.e elections and NCII) as opposed to all AI-assisted media.

The most important legislation that WITNESS contributed to however is AB853, the California AI Transparency Act (2025), which we believe provides the technical backbone for California's transparency regulations. The Act requires AI developers, large platforms and device manufacturers (i.e. camera manufacturers) to support and preserve verifiable system-level provenance data. This will provide the "recipe" for if or how AI content was involved in the creation of content, moving away from the "real" or "AI" false binary. AB853 does not mandate visible labels or overlays but requires an architecture of trust that is interoperable and privacy-preserving while aligning with open standards. These standards most importantly are guarded against personal provenance data which the Act emphasizes by safeguarding privacy while it pursues accountability and authenticity. California's rights-based, infrastructure-oriented model reflects one way of embedding accountability within a mature technology ecosystem. India, similarly another key player in the global digital landscape, could consider adapting comparable principles within its own context, rather than relying on a platform liability approach.

Brazil: A Rights Based Framework in Development

Brazil's proposed Artificial Intelligence Legal Framework (PL 2338/2023), approved by the Senate and under review in the Chamber of Deputies provides a good model for an emerging rights-based approach to AI governance in the Global South. It establishes risk-based classification and transparency obligations for developers and deployers of AI systems while using Brazil's constitutional and data-protection framework to embed protections for privacy and freedom of expression. WITNESS is actively engaging in the development of this framework and sees potential for the bill to include specific provisions on synthetic or generative media. WITNESS has been hopeful by the government's ongoing consultations on provenance and watermarking standards and their interest in developing their regulations with pipeline accountability in mind.

As a fellow BRICS member country, Brazil's model offers relevant lessons for India. Brazil, if it continues down this trajectory, can demonstrate how national AI regulation can integrate transparency and provenance within the dimensions of human rights and proportionality.

4. Rule-by-rule recommendations

While WITNESS urges MeitY to develop a dedicated AI Transparency regulation, we have made some suggestions below in reference to each element in the proposed Amendment. These recommendations are grounded in our experience contributing to the EU AI Act's Code of Practise on Transparency (and the drafting working groups), providing a viewpoint on California's AB853, and supporting the adoption of open provenance standards such as the Coalition for Content Provenance and Authenticity (C2PA). Our recommendations and revisions are done with the intention to be constructive, evidence-based and to make India's framework proportionate, interoperable, and rights consistent should MeitY proceed with this amendment process.

Rule 2 (i): Definition of "Synthetically Generated Content"

Current Text:

> "(wa) 'synthetically generated information' means information which is artificially or algorithmically created, generated, modified or altered using a computer resource, in a manner that such information reasonably appears to be authentic or true."

Proposed Revision:

""(wa) 'synthetically generated content' refers to content that has been wholly or partly produced or substantially modified through automated or algorithmic processes, including the use of generative AI systems, where such processes materially influence the content's form or meaning and are intended, or reasonably likely, to mislead users as to its origin or authenticity.``

Explanation: *For the avoidance of doubt, this shall not include content that employs such processes for non-substantive purposes such as for accessibility, assisted editing, restoration, translation, captioning or similar technical assistance. Under these Rules, such content with benign uses of AI shall not be considered "synthetically generated", unless there is material alteration of meaning or reasonably likely to deceive viewers. Artistic, satirical, journalistic, educational, or research uses that clearly communicate their context are likewise exempt from this definition."*

Rationale:

If AI simply helps communicate human intent (i.e. translation, subtitling), it's assistive. If AI creates new meaning or false representation, it's synthetic and requires disclosure. This revision makes it clear that not all AI involvement requires regulation. The definition therefore narrows the scope to focus on intent and likelihood to deceive. Most importantly it avoids treating AI-generated content as inherently "artificial" or suspect, and introduces necessary exemptions for accessibility and legitimate creative or journalistic use. It also shifts emphasis from binary "AI-made" labels to process transparency, ensuring users understand how content was generated rather than categorising it as real or fake. To clarify, this definition does not remove the obligation to disclose or label synthetic content, but ensures the duty to disclose is risk-based and tied to potential harm, preventing routine AI assistance. This approach mirrors risk-based standards in the EU AI Act (Article 50) and California AB 853.

Rule 2(ii): Application to Unlawful Acts

Current Text:

> "(1A) For the purposes of these rules, any reference to 'information' in the context of information being used to commit an unlawful act, including under clause (b) and (d) of sub-rule (1) of rule 3 and sub-rules (2) and (4) of rule 4, shall be construed to include synthetically generated information, unless the context otherwise requires."

Recommendation:

WITNESS recommends that the proposed Rule 2(ii) be deleted. Its blanket inclusion of "synthetically generated information" under all "unlawful-act" provisions is unnecessary and risky. India already has sufficient legal tools to address AI-generated harms through targeted provisions on defamation, fraud, sexual exploitation, and election integrity. MeitY should work with relevant ministries and regulators to clarify and update existing provisions to cover AI-generated forms of those offences instead of expanding these unlawful act clauses to monitor all AI related content.

A dedicated AI framework would provide a far more coherent, targeted and proportionate mechanism for managing risk. India needs risk-based classification, transparency obligations, and clear accountability along the AI pipeline rather than relying on broad takedown duties that risk overcompliance and censorship. The EU AI Act follows this approach by imposing targeted transparency and disclosure duties for manipulative or deceptive content, while California's laws focus narrowly on specific harms such as election misinformation and non-consensual deepfakes, avoiding broad intermediary liability. WITNESS urges MeitY to develop a comprehensive AI Act.

Rule 3: Accountability of Platform Enforcements

Current Text:

"In the said rules, in rule 3, in sub-rule (1), in clause (b), before the explanation, the following proviso shall be inserted, namely:—

> "Provided that the removal or disabling of access to any information, including synthetically generated information, data or communication link within the categories of information specified under this clause as part of reasonable efforts or on the basis of grievances received under sub-rule (2) by such intermediary, shall not amount to a violation of the conditions of clauses (a) or (b) of sub-section (2) of section 79 of the Act.""

Proposed Revision:

"Where an intermediary, acting in good faith and on the basis of verified grievances, removes or disables access to information—including synthetically generated content—it shall not, solely by reason of such removal, be considered to have violated clauses (a) or (b) of sub-section (2) of Section 79 of the Act: Provided further that, where feasible, such removal or disabling shall include:

a. *Timely notice to the user who shared the content, explaining the reason for action;*
b. *A clear and accessible appeals mechanism to contest mistaken or unfair removals; and;*
c. *Inclusion of the action in the intermediary's periodic transparency reports, indicating the volume and grounds of such takedowns."*

Rationale:

In draft's current form, Rule 3's proviso risks reinforcing the over-reach created by Rule 2(ii) by encouraging intermediaries to take down any AI-assisted content to avoid liability. WITNESS recommends retaining this clause only if Rule 2(ii) is deleted and the definition of "synthetically generated content" is narrowed to focus on intent to deceive.

In that context, the rule should only serve to protect intermediaries acting in good faith while guaranteeing users' rights to notice, appeal, and transparency. These safeguards would prevent the chilling of lawful expression, particularly from journalists, human-rights defenders, and civic actors who rely on digital platforms for expression. We will further touch on the dangers of relying on platforms to conduct post-facto detection under comments for Rule 5 when we argue for the unreliability of this method (referenced in this amendment as "automated verification") required to implement this proviso.

This balanced approach aligns with the EU Digital Services Act, which embeds notice-and-appeal obligations. It ensures India's intermediary framework promotes responsible moderation and procedural fairness, rather than incentivising over-compliance or pre-emptive censorship.

Rule 4: 3(3) Labelling and Metadata Requirements

Current Text:

"In the said rules, in rule 3, after sub-rule (2), the following sub-rule shall be inserted, namely:—

> "(3) Due diligence in relation to synthetically generated information:
> (a) Where an intermediary offers a computer resource which may enable, permit, or facilitate the creation, generation, modification or alteration of information as synthetically generated information, it shall ensure that every such information is prominently labelled or embedded with a permanent unique metadata or identifier, by whatever name called, in a manner that such label, metadata or identifier is visibly displayed or made audible in a prominent manner on or within that synthetically generated information, covering at least ten percent of the surface area of the visual display or, in the case of audio content, during the initial ten percent of its duration, and can be used to immediately identify that such information is synthetically generated information which has been created, generated, modified or altered using the computer resource of the intermediary;
> (b) the intermediary under clause (a) shall not enable the modification, suppression or removal of such label, permanent unique metadata or identifier, by whatever name called.""

Proposed Revision:

"(3) Due diligence in relation to synthetically generated content:

a) Where an intermediary, developer, or deployer offers or uses a computer resource that enables or facilitates the creation or substantial modification of content through automated or algorithmic processes, each actor shall ensure that such content includes, or is accompanied by, tamper-evident provenance metadata compliant with open, interoperable standards and/or a permanent identifiers such as an invisible watermark or fingerprint.

b) Such provenance metadata or identifier shall be designed to communicate the process and tools used to create or alter content, without disclosing personal or identifying information *about the creator or user without their explicit and informed consent. The metadata or identifier should be privacy-preserving, machine-readable, and resistant to unauthorised removal or falsification.*

c) *Visible or audible indicators shall be visibly displayed or made audible in a prominent manner. They shall be context-sensitive and proportionate, for example, a visual or textual marker displayed at the start or in an accompanying notice, rather than a fixed overlay occupying a percentage of the display or audio duration.*

Pipeline responsibility:

a) *Developers, deployers, and intermediaries shall share responsibility for maintaining provenance signals throughout the lifecycle of the content. Intermediaries shall not be held liable for the absence of provenance data where such signals were not attached at the point of creation or deployment.*
b) *This rule shall not be interpreted to require general monitoring or automated scanning of all content by intermediaries."*

Rationale:

This section introduces the most far-reaching labelling obligation in the draft amendment. To ensure proportionality and technical feasibility, WITNESS emphasises pipeline responsibility as the organising principles: disclosure and provenance duties must be shared among developers, deployers and intermediaries consistent with global practise. Our engagement here is to propose revisions that would create minimum necessary improvements, but we remain steadfast in urging MeitY to develop a multistakeholder approach towards an Indian AI Act with comprehensive transparency provisions. The current quantitative requirement of 10% overlay is impractical, inaccessible, and incompatible with existing global standards. It risks embedding trackable identifiers that compromise privacy and usability, without actually improving authenticity or user understanding.

WITNESS supports transparency through provenance and disclosure, but these goals are best achieved through tamper-evident, rights-respecting metadata built on open and interoperable standards such as the Coalition for Content Provenance and Authenticity (C2PA), and/or other forms of identifiers, such as resilient invisible watermarks or fingerprints. These approaches enable verifiable transparency into if or how AI tools were used without compromising privacy or accessibility.

Fixed visual or audio overlays should play a role in promoting transparency when part of a holistic approach that combines it with tamper-evident metadata or other identifiers captured and shared amongst complaint tools and services, distributed through pipeline responsibility across the AI lifecycle spanning developers, deployers, and intermediaries. This would position India as a leader of technically sound, rights-based transparency consistent with the EU AI Act (Article 50) and California AB 853, and away from a platform-centric system that risks overcompliance and censorship without achieving genuine accountability.

Rule 5: 4(1A) Obligations for Significant Social Media Intermediaries (SSMIs)

Current Text:

"In the said rules, in rule 4, after sub-rule (1), the following sub-rule shall be inserted, namely:—

> "(1A) A significant social media intermediary which enables displaying, uploading, or publishing any information on its computer resource shall, prior to such display, uploading, or publication,—
> (a) require users to declare whether such information is synthetically generated information;
> (b) deploy reasonable and appropriate technical measures, including automated tools or other suitable mechanisms, to verify the accuracy of such declaration, having regard to the nature, format, and source of such information; and
> (c) where such declaration or technical verification confirms that the information is synthetically generated, ensure that the same is clearly and prominently displayed with an appropriate label or notice, indicating that the content is synthetically generated:
>
> Provided that where such intermediary becomes aware, or it is otherwise established, that the intermediary knowingly permitted, promoted, or failed to act upon such synthetically generated information in contravention of these rules, such intermediary shall be deemed to have failed to exercise due diligence under this sub-rule.
>
> **Explanation.—** For the removal of doubts, it is hereby clarified that the responsibility of the significant social media intermediary shall extend to taking reasonable and proportionate technical measures to verify the correctness of user declarations and to ensure that no synthetically generated information is published without such declaration or label."”

Recommendation:

WITNESS recommends this Rule be withdrawn or restructured. While we encourage accountability against harmful deepfakes as well as provenance and transparency standards, this sub-rule places disproportionate responsibility on platforms. This emphasis risks mass surveillance and over-removal of lawful content. It also conflates transparency with content control, whereas provenance standards such as C2PA show that trust can be built through interoperable, privacy-preserving infrastructure applied at the point of creation rather than through platform-level verification. India's framework should make the shift from platform-centric to pipeline model by enacting the following:

a. impose disclosure and provenance duties on developers and deployers of generative-AI systems
b. require intermediaries only to preserve and display provenance signals, not to generate or verify them
c. maintain strong safeguards for privacy, encrypted communications, and user autonomy

Rationale:

The requirement for "automated verification" is technically unreliable and unsuited to legal obligations. WITNESS' work with the Deepfake Rapid Response Force; building awareness about AI detection tools and advocating for benchmarks for these tools (the TRIED Benchmark)[6], demonstrate they are inconsistent, and prone to false positives. AI detection, or "automated tools" as per the language of the draft, remain inconsistent across files, modalities, and contexts and are prone to false positives and false negatives. Additionally, AI detection tools cannot differentiate between varying uses of manipulation. If a detection tools identifies use of AI, it will automatically designate a piece of content as synthetic without providing the explanation as to how or for what purpose it was used, reinforcing the harmful binary of "real" and "fake" and neglecting the need to distinguish between different types of AI uses outlined in the comment on Rule 2. Building regulatory enforcement on such uncertain technology would therefore be both ineffective and risky, while also imposing unrealistic burdens on platforms to assess the authenticity of every piece of user content.

A rights-consistent alternative would mirror the EU AI Act (Article 50) and California AB 853, which locate transparency duties with those who create or deploy AI systems, not with platforms. These frameworks achieve accountability through provenance infrastructure, not through content screening.

WITNESS urges MeitY to reconsider the architecture of Rule 4(1A) and instead embed transparency requirements within a dedicated AI framework that clarifies pipeline responsibilities, encourages open standards, and protects encryption and freedom of expression.

5. Conclusion

WITNESS welcomes that MeitY is acting swiftly to respond to the growing challenge of synthetic and AI-generated content and its intention to promote transparency and accountability. However, these current amendments to the existing IT Rules are not the appropriate framework to achieve these objectives. This draft amendment extends intermediary liability in ways that risk overbroad censorship, privacy violations, and technical infeasibility, while failing to build the kind of trust infrastructure required to address AI-driven harms.

Our revisions and recommendations here represent the minimum necessary improvements to make the current amendment more practical, privacy-preserving, and interoperable with international standards. However, even with these changes, the IT Rules framework remains too limited to serve as India's primary instrument for AI governance or transparency.

The European Union and California have both advanced beyond piecemeal content regulation, developing comprehensive AI transparency regimes that embed provenance, accountability, and rights safeguards across the AI pipeline. India now has an opportunity to take a similar leadership path. India needs and deserves a comprehensive AI Act with specific and thoughtful carveouts for transparency. India needs to lead in enabling coherent, risk-based, and globally interoperable governance of synthetic and generative media.

[6] To read about WITNESS' Deepfake Rapid Response Force, the first global mechanism to respond to the threat of deepfakes, see here. WITNESS has also developed this tip sheet on educating users about AI detection tools here. WITNESS has developed a TRIED Benchmark for AI Detection tools, see here.