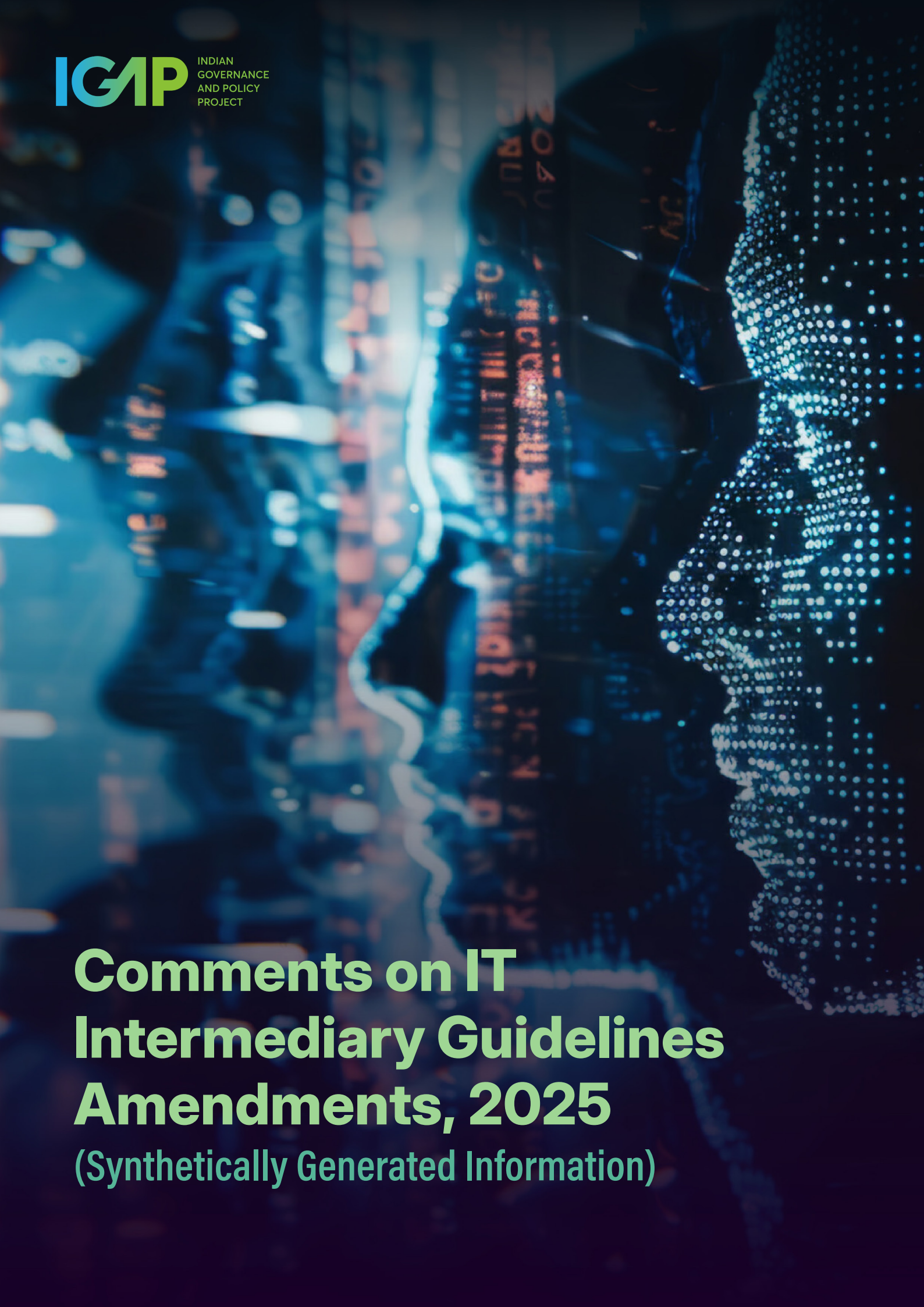

INDIAN GOVERNANCE AND POLICY PROJECT

Comments on IT Intermediary Guidelines Amendments, 2025

(Synthetically Generated Information)

COMMENTS ON IT INTERMEDIARY GUIDELINES AMENDMENTS, 2025 (SYNTHETICALLY GENERATED INFORMATION)

November 2025

Published by
Indian Governance and Policy Project (IGAP)

Authored by
Soumya AK

Designer
Manoj Murali

About IGAP
The Indian Governance and Policy Project (IGAP) is a policy, business advisory, and research studio working at the intersection of governance, technology, markets, and national development.

Grounded in a clear understanding of how state capacity, market forces, and emerging technologies shape India's strategic trajectory, IGAP addresses key questions that define the country's future — from the governance of AI and digital infrastructure to financial innovation, sustainability, and national security.

Bringing together lawyers, policy thinkers, and strategists with deep business and geopolitical insight, IGAP delivers solutions that balance India's developmental and security priorities with its democratic values and constitutional principles.

This study is published under the Creative Commons Attribution–CC BY-SA License. This license allows others to copy, distribute, remix, adapt, and build upon the material in any medium or format, provided appropriate credit is given to the creator and any derivative works are shared under the same license.

Introduction

We at the Indian Governance and Policy Project (IGAP) thank the Ministry of Electronics and Information Technology (MeitY) for inviting stakeholder inputs on the Draft Amendments to the Intermediary Rules. We appreciate the Government's continued commitment to a transparent and consultative process while updating India's digital regulatory framework to address emerging technological and societal challenges.

The Government's attention to the emerging risks arising from the misuse of deepfakes and other forms of synthetically generated content is commendable. The intention underlying the proposed amendments is both timely and laudable, reflecting a clear commitment to safeguard India's *Digital Nagriks* and ensure that online ecosystems remain secure, trustworthy, and resilient. We fully support efforts to strengthen user protection, enhance the safety and integrity of the internet, and modernize India's cyber-regulatory framework in line with technological advances, global best practices, and the evolving challenges posed by illegal or harmful content.

The proposed amendments **(2025 Amendments)** to the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 **(IT Rules)** mark India's first formal step toward governing AI-generated and algorithmically processed content. As highlighted in the explanatory note, the amendments aim to address harms such as AI-generated misinformation, deepfake media capable of reputational injury, electoral manipulation, and fraud. Unlike jurisdictions that have enacted dedicated AI legislation, India's approach relies on updating intermediary rules to regulate synthetic content.

Against this backdrop, the key questions now relate not to the intent of the amendments, which is clear and commendable but to how these obligations will function in practice. The draft rules introduce important concepts such as "synthetically generated information," mandatory labelling and provenance requirements, and platform-level verification duties. To be effective and proportionate, these obligations must be operationalised with clarity on what constitutes synthetic media, how "reasonably appears to be authentic or true" can be applied consistently, and how to distinguish high-risk generative outputs from benign or routine editing tools.

Practical implementation will also require workable approaches to labelling - whether through visible markers, metadata, or watermarking that persist across formats, edits, transcodes, and reposts, without imposing excessive burdens on creators, smaller developers, or ordinary users.

The proposed definition of “synthetically generated information” is so broad that it may inadvertently encompass a vast range of ordinary digital content including routine enhancements, filters, editing tools, autocorrect, and translation simply because the output “appears authentic or true.”

Without a materiality threshold or a distinction between harmless processing and genuinely deceptive manipulation, the Rules risk applying obligations intended for high-risk deepfakes to everyday digital activity. Such an expansive scope goes significantly beyond MeitY’s stated objective of addressing harmful synthetic misinformation and creates considerable uncertainty for both users and intermediaries in determining what content the definition actually captures.

This broad definitional trigger, combined with the structure of the compliance obligations, may also create strong incentives for intermediaries to over-remove or restrict content to avoid any risk of non-compliance. Since platforms face potential liability for failing to act but no corresponding consequence for over-compliance, the rational response is to default to pre-emptive takedowns, down-ranking, or the aggressive application of labels particularly given the inherent inaccuracies and false positives in current detection technologies. Such risk-averse behaviour can produce a chilling effect on lawful expression, disproportionately affecting creators, journalists, and dissenting or marginalised voices. A narrower, harm-aligned definition and proportionate, risk-based obligations would enable the framework to target genuinely deceptive synthetic media while avoiding unnecessary over-compliance that suppresses legitimate speech and innovation.

Similarly, the proposed requirements for platforms to verify user declarations and deploy “reasonable technical measures” must reflect the technical limits of detection systems today. Feasible verification will involve a mix of metadata checks, selective AI-based classifiers, and risk-based sampling rather than intrusive or generalised monitoring. This raises critical design questions - what false-positive and falsenegative thresholds are acceptable, how intermediaries can scale verification across India’s diverse, multilingual media landscape, and how implementation costs can be balanced to avoid over-blocking, chilling effects on creativity, or disproportionate burdens on smaller players.

1

DEFINITION OF "SYNTHETICALLY GENERATED INFORMATION" – RULE 2(1)(WA)

The 2025 Amendments introduce Rule 2(1)(wa), which 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." The definition is technology-neutral and focuses

(a) on the *process* (artificial or algorithmic creation) and

(b) the *effect* (appearing authentic). It also includes modification and alteration alongside creation and generation, significantly widening the ambit.

The definition of SGI brings under its ambit even innocuous digital processing of information and does not exclude enterprise use cases. Additionally, classic exemptions such as parody, satire and artistic works are not provided within the text of the proposed amendment. The primary issue with the definition is that it covers technologies and use cases that neither use AI systems nor carry the risks intended to be addressed by the proposed amendments as per the explanatory note. Whereas other jurisdictions take a more nuanced and narrow approach. The EU AI Act specifically defines "deepfakes" as AI-altered images, audio, or video that resembles existing persons, objects, places etc. and would falsely appear to a person to be truthful[1]; China enumerates categories of content such as text generated via chatbots, synthetic voices, AI-generated faces or videos, and virtual scenes, that may cause misrecognition or confusion among the public.[2] South Korea explicitly defines outputs from generative AI (text, sound, images, and video).[3] India's provision, *by contrast*, could theoretically encompass any content processed algorithmically if it "appears authentic," regardless of whether AI is involved.

a. Overbroad Definition

First, the breadth of the definition raises substantial concerns about unintended capture of routine, benign, and even beneficial uses of digital technology. Nearly all digital content involves some degree of algorithmic processing. A further difficulty arises with mixed or hybrid media combining authentic and synthetic elements. Increasingly, real photographs contain digitally generated details, while AI-generated visuals embed authentic human features or voices. The current definition fails to accommodate this.

CATEGORY	EXAMPLE	WHY WOULD IT BE COVERED
Smartphone Photos	Photos automatically enhanced for colour balance, sharpness, or noise reduction	The image is "modified or altered through algorithmic means" and appears authentic
Edited Videos	Clips with background replacement, lighting correction, or automatic stabilization	These edits involve algorithmic changes that make the content appear realistic
Audio Recordings	Recordings using noise cancellation, equalization, or auto-tuning	Algorithms modify the audio signal to make it clearer or more natural
Social-Media Filters	Filters that smooth skin tone, adjust lighting, or apply AR effects	The final image/video is algorithmically altered but looks authentic
Automatic Image Enhancements	Software like Photoshop "auto enhance" or HDR adjustments	The tool algorithmically modifies pixels to optimize the image
Speech-to-Text Transcripts	AI-generated transcripts of real speech or automated summarisation and translation tools	Algorithmic generation of text that appears truthful or authentic, though used for practical and beneficial purposes
Spellcheck or Grammar Corrections	Edited text with AI or algorithmic correction tools	Words have been "created or altered" algorithmically to appear correct/authentic
Autocorrect and Predictive Text	Messages adjusted by phone keyboard suggestions	Text is algorithmically modified before being sent, appears real
AI-Assisted Text Creation	Emails, internal memos, meeting notes, or marketing copy generated or refined using AI tools	Algorithmic text that appears authentic but poses no risk of deception—illustrating how textbased outputs could fall under SGI even though they lack the disinformation potential of images or video
Virtual Backgrounds in Video Calls	Zoom or Teams background blurring/replacement	Algorithmic alteration creates a realistic image that appears true
Photo Compression or Resizing Tools	Online images that are compressed or resized algorithmically	Alters the data through algorithmic means, even if imperceptible

Under a literal reading of the proposed definition, all such illustrative content will qualify as synthetically generated information if it "reasonably appears to be authentic or true."

Recital 133 of the EU AI Act[4] explicitly guards against such overreach by excluding AI systems that offer assistive or enhancement functions that do not substantially alter the semantics or factual content of the input data. Without such a threshold, the definition risks sweeping in routine technological processes that bear no resemblance to synthetic deception. This would not only create compliance uncertainty but also trivialise the genuine mischief the amendment seeks to address.

Second, the threshold of "reasonably appears to be authentic or true" lacks objective benchmarks for assessment. What appears authentic to one observer may not to another, and what appears authentic in one context may clearly be fictional in another. The EU's formulation uses "falsely appear to a person to be authentic or truthful",[5] which, while also subjective, incorporates the element of falsity, implying deceptive potential rather than mere technical characteristics. China's focus on content that "may cause confusion or misrecognition" similarly centres on actual or potential deceptive effect".

This subjectivity creates uncertainty for intermediaries attempting to determine whether content falls within the definition, potentially leading to inconsistent application across platforms and excessive caution that manifests as overbroad censorship of legitimate content.[6]

b. Divergence from Legislative Intent

The Ministry's explanatory note[7] accompanying the draft rules indicates an intent to address AI-generated misinformation and deepfake media capable of reputational harm, electoral manipulation, or fraud. However, the operative definition extends far beyond this to any algorithmic modification, regardless of intent or effect. Such divergence between legislative purpose and statutory text may produce both underand over-inclusive outcomes as it may fail to capture non-AI but deceptive manipulations, while simultaneously burdening legitimate innovation in digital imaging, audio enhancement, or accessibility technologies.

By contrast, major platforms have adopted more targeted disclosure frameworks that distinguish between routine editing and materially deceptive content.[8] These policies typically require disclosure only for content that:

(i) makes a real person appear to say or do something they didn't;

(ii) alters footage of a real event or place; or

(iii) generates a realistic-looking scene that didn't occur.

Crucially, such frameworks exclude adjustments limited to colour correction, brightness, or other technical enhancements, as well as content where synthetic elements are obviously unrealistic or clearly labelled as creative effects.

RECOMMENDATION

To prevent overreach and ensure proportional application of obligations, the definition of *SGI* under the Draft IT Rules, 2025, should be modified keeping the following principles at the centre:

- ***Aligning the definition with the intent:*** Since the policy objective is to regulate *AI-generated content* (such as deepfakes in misinformation, including electoral misinformation and voter deception, non-consensual intimate imagery, infringement of personality rights, financial frauds and other similar serious harms), the definition needs to be narrow and aligned to this objective.
- **Materiality threshold:** Obligations to apply only to content that is *materially altered* in a manner that affects its authenticity or factual accuracy, and that *reasonably appears to depict real persons, events, or statements that did notoccur.*
- **Adoption of an exclusion framework:** Similar to Recital 133 of the EU AI Act, the rules may exclude:

  (a) *Minor or assistive modifications* (e.g., colour correction, background blur, accessibility edits);

  (b) *Routine technical processing or enhancement* (e.g., compression, upscaling, format conversion); and

  (c) *Edits that do not substantially alter the semantic meaning or factual truthfulness* of the material.

LABELLING AND METADATA REQUIREMENTS FOR INTERMEDIARIES – RULE 3(3)

Clause 4 of the Amendments inserts new sub-rule (3) into Rule 3 of the IT Rules, imposing due diligence obligations on intermediaries that offer computer resources enabling the creation, generation, modification, or alteration of synthetically generated information.

Under proposed Rule 3(3)(a), such intermediaries must ensure that every piece of synthetically generated information *is prominently labelled or embedded with a permanent unique metadata or identifier*, in a manner that is *visibly displayed or made audible* in a prominent way. The rule further specifies that this label must cover at least **ten percent** of the surface area of the visual display or, in the case of audio content, be audible during the initial ten percent of its duration.

Additionally, Rule 3(3)(b) prohibits such intermediaries from enabling the modification, suppression, or removal of the label, metadata, or identifier, thereby mandating its permanence and detectability.

a. Multiplicity of Actors in the AI Value Chain

The proposed obligation that any intermediary offering a computer resource "which may enable, permit or facilitate" the creation or dissemination of SGI must ensure such information is labelled or embedded with identifiers overlooks the complex structure of the AI value chain.

A single system may involve multiple actors i.e. infrastructure providers, developers, vendors, deployers, and users, each with varying levels of control and visibility over outputs. Imposing identical obligations across this chain is technically infeasible and risks both overreach and regulatory gaps. Without clear delineation of roles, compliance could become fragmented - multiple entities might duplicate labelling efforts, or each might assume the other is responsible, resulting in non-compliance.

Comparatively, international frameworks adopt more granular role definitions. The EU AI Act distinguishes between *providers* i.e. developers who develop AI systems with a view to placing it on the market or putting it into service under their own name or trademark and *deployers* i.e. users of AI systems under its authority, in professional

capacity[9]. China regulates all internet information service providers and has specific provisions for service providers of "deep synthesis" AI content. South Korea's law uses "AI business operator," a broad category covering both developers and deployers of AI services[10]. In South Korea's framework, transparency and disclosure obligations are primarily imposed on deployers that provide AI products and services to end-users, rather than on upstream actors[11]

.

These distinctions acknowledge the distributed responsibility within AI ecosystems, ensuring obligations align with actual operational control and capacity, rather than applying a single standard to technically and functionally distinct actors.

b. Technical Feasibility: Labelling Types, Provenance, and Interoperability Challenges

(i) Conflation of Labelling Methods

Intermediaries are required to ensure that all such information carries a visible label or an embedded, permanent identifier, and to prevent its removal. However, the draft conflates these two distinct approaches and places disproportionate emphasis on visible or explicit identifiers.

- Visible labels (e.g., watermarks or on-screen notices) inform viewers directly but can be cropped, edited out, or obscured. They also offer no technical verification and can disrupt visual or audio quality, especially in professional or creative contexts.

- Embedded metadata or "content credentials" provide machine-readable provenance but serve a different function - tracing origin rather than signalling to users. Each of these methods have distinct technical characteristics and practical constraints.[12]

(ii) Limitations of Provenance and Metadata Systems

The draft amendments also assume the availability of reliable and tamper-proof systems for embedding and preserving provenance data in synthetic media. In practice, such technologies are still evolving, and no current standard has proven immune to removal or corruption.[13]

- Visible labels are simple to apply but lack technical verifiability. They can be cropped, edited out, or removed during redistribution, and their overuse may even erode user trust.[14] Embedded metadata (such as EXIF or XMP) can record provenance, but major platforms routinely strip this data upon upload, undermining its persistence.[15]

- Machine-readable watermarking tools introduce imperceptible signals within content to identify AI-generated material. However, these are also vulnerable to transformations (post-processing, cropping) and adversarial attacks.[16] Google's SynthID, for instance, can embed invisible watermarks in AI-generated text and images, but detection can be easily evaded through edits, reformatting or manipulations.[17]

- Similarly, content credential frameworks such as the C2PA standard aim to create tamper-evident provenance chains using cryptographically signed manifests combined with metadata to establish tamper-evident provenance chains.[18] Yet, these too can be removed or corrupted when content passes through non-compatible platforms.[19] While newer C2PA versions integrate watermarking to strengthen resilience, this approach depends on broad ecosystem adoption and significant infrastructure investment - a burden that may be prohibitive for smaller intermediaries or tool providers.

(iii) Interoperability and Cross-Platform Enforcement Gaps

Even where labelling or provenance systems are implemented, content can easily migrate beyond controlled environments. Users can download, screenshot, or re-upload material across services that use different or no provenance standards. Once content leaves the original platform, labels and metadata are frequently lost, making consistent enforcement impossible.

Without universal technical standards and cross-platform adoption, provenance and labelling obligations risk being fragmented and ineffective.

c. Proportionality and Impact on Legitimate Uses

The labelling requirements under the proposed rule are overly rigid, prescribing that identifiers must cover at least ten per cent of a visual display or play during the first ten per cent of an audio clip. Such uniform, quantitative mandates are not found in other jurisdictions and fail to account for differences in format, purpose, or use. The proposed rule also provides no exemption for legitimate professional, creative, scientific, or accessibility-related uses of synthesis technologies, risking disruption of lawful and socially valuable applications and may impact protected speech and expression.

This creates two key challenges:

***First,* the rule contains no exemptions for low-risk or legitimate uses.** As detailed in the illustrative table below, it would require labelling even for routine applications such as VFX in films, auto-tuning in music, or AI-assisted

transcription in journalism and academia. By contrast, the EU AI Act exempts disclosure when AI is used for criminal investigation or prosecution, and allows artistic or creative content to disclose AI use in ways that don't interfere with the work's presentation or enjoyment.[20] Similarly, South Korea's Framework Act[21] allows flexibility for artistic and creative expressions, permitting notifications that don't hinder exhibition or enjoyment.

***Second,* the rule fails to differentiate risk levels or manipulation severity.** It applies the same ten percent labelling threshold to both innocuous AI-assisted tools and high-risk deceptive content. This undifferentiated approach risks label fatigue - when users encounter excessive or irrelevant warnings, they become desensitized, undermining the effectiveness of labels meant to flag genuinely misleading or harmful content[22]. Several jurisdictions like Brazil[23], South Korea[24] and Singapore[25] have introduced targeted provisions to address risks arising from synthetic and manipulated content, particularly in electoral contexts. These examples highlight how other jurisdictions are adopting risk-differentiated and context-specific approaches to managing synthetic media.

The table below illustrates how such rigid labelling could affect different sectors and typical uses of AI tools

CATEGORY	TYPICAL USE OF SYNTHETIC / AI TOOLS	RISK OF DECEPTION	IMPACT OF LABELLING
Film & Visual Media	VFX, CGI, colour correction	Low	Distorts artistic quality and viewing experience
Music & Audio Production	Auto-tuning, background mixing, restoration	Low	Disrupts creative process; unnecessary disclosure
Journalism & Editorial Work	Transcription, summarisation, captioning	Low	Burdens standard production workflows
Academic & Scientific Uses	Simulations, visualisations, accessibility tools	Minimal	Hinders research and inclusion
Software & Productivity Tools	Autocorrect, translation, predictive text	Minimal	Overregulates ordinary digital use
Marketing & Corporate Communication	AI-assisted copy, report generation	Low	Affects efficiency; adds redundant compliance

Personal / Social Media Users	Filters, voice enhancement, image retouching	Minimal	Regulates harmless personal expression
Malicious / Deceptive Uses	Deepfakes, impersonation, fabricated media	High	Legitimate focus of regulation

d. Compelled Speech: Constitutional Considerations

The proposed labelling requirement - mandating that synthetic content display a visible or audible warning covering at least ten percent of the media may raise concerns of compelled speech.

Under Article 19(1)(a) of the Constitution, the right to free expression includes the right not to be forced to convey a state-mandated message within one's own creative, journalistic, or communicative work. A uniform and intrusive labelling mandate, especially one that alters the aesthetics or substance of expression, can be difficult to justify for content that is lawful, non-deceptive, or obviously artificial. Such measures must satisfy the constitutional tests of necessity and proportionality, and the proposed amendments do not currently differentiate between contexts where a warning is essential and those where it may not be required. This concern is heightened by the fact that the labelling obligation applies equally to high-risk deepfakes and routine uses of digital tools such as filters, translations, editing software, or accessibility modifications.

By requiring creators and intermediaries to embed prominent warnings across all synthetic outputs regardless of harm, purpose, or audience the amendments risk inserting State-mandated speech into private expression in ways that may go further than necessary to address misinformation or deception.

A more narrowly tailored approach, focusing on materially deceptive content, would better align with constitutional principles while still addressing the harms the Rules seek to prevent.

RECOMMENDATION

- **Achievable technical standards:** Encourage adoption of open and interoperable standards for implicit (provenance based markers) as well as explicit labels. Imposing an obligation to label and detect without such standards will not achieve the goal of the amendments. This will also leave room for varying degrees of compliance and vagueness.

- **Categorical exemptions:** Exclude content where synthetic nature is contextually obvious and poses minimal deception risk, for instance, professional cinema and entertainment (disclosed through credits), scientific visualization and simulation, educational and journalistic content.

- **Revise visible label mandate:** Replace prescriptive requirements (e.g., "10% of visual surface") with flexible standards allowing platforms to implement disclosure methods appropriate to content type and context.

- **Risk-based differentiation:** Labelling requirements should reflect the potential risk or likelihood of deception posed by synthetic content. High-risk material - such as electoral misinformation, financial fraud, or non-consensual intimate imagery may warrant clear, visible labelling, whereas low-risk or contextually evident content (for example, creative, educational, or accessibility-related uses) should allow for flexible or minimal disclosure. This approach maintains clarity for users without over-labelling benign or clearly identifiable content.

OBLIGATIONS ON SIGNIFICANT SOCIAL MEDIA INTERMEDIARIES - RULE 4(1A)

Clause 5 of the 2025 Amendments inserts **Rule 4(1A)**, which sets out additional due diligence obligations for Significant Social Media Intermediaries **(SSMIs)**. SSMIs are required to ensure that users declare whether content being uploaded is synthetically generated and to deploy reasonable and appropriate technical measures to verify such declarations. They must also label or clearly identify synthetic content published on their platforms and prevent its dissemination without appropriate disclosure. The requirement to deploy "reasonable and appropriate technical measures" acknowledges that perfect detection is not currently possible and allows for evolution of technical approaches as technology advances. The provision appropriately recognizes that verification approaches may need to be tailored based on "the nature, format, and source of such information", allowing for context-sensitive implementation.

a. Concerns with User Declaration Requirements

The effort to place responsibility on users to declare synthetically generated content is a constructive step toward shared accountability. However, the challenge lies in the overbroad and ambiguous definition of synthetic content. Without clear parameters, users will struggle to determine what qualifies—potentially leading to inconsistent or inaccurate declarations. This will likely result in inconsistent or false declarations, whether out of confusion or to avoid restrictions. *Second,* the declaration requirement creates friction in the user experience that may deter content creation and sharing. Every upload would require users to pause and consider whether their content is synthetic, make a determination based on unclear criteria, and complete an additional declaration step. For platforms built on rapid, spontaneous sharing of content, this friction could substantially degrade user experience and platform functionality.

b. Verification Challenges: Detection tools and limitations

Detection of synthetic media remains an active area of research, with different content categories (audio/video) requiring distinct detection methods and tools, and presenting its own technical limitations.[26] Broadly, detection methods fall into two categories - inference-based and provenance-based. Inference-based techniques

analyse the content itself, identifying visual or acoustic inconsistencies such as irregular lighting, pixel artifacts, or disrupted motion coherence in videos. Provenance-based systems, by contrast, rely on embedded identifiers or cryptographic "content credentials" to authenticate the origin and modification history of media.[27]

Each approach faces practical challenges. Inference-based detection struggles with generalization failures,[28] as tools trained on specific datasets or known manipulation techniques perform poorly against new generative models or real-world compressed, and low-quality content—especially common on social media.[29] Detection of audio deepfakes is further complicated by background noise or overlapping speech, which degrade accuracy.[30] Provenance-based methods, meanwhile, depend on widespread adoption of consistent metadata or watermarking standards, yet many platforms strip or alter such data upon upload, undermining reliability.

At scale, real-time detection and verification remain computationally expensive and difficult to operationalize. Applying deepfake detection to the billions of images, videos, and audio clips uploaded daily would demand vast computing power and introduce delays that could disrupt platform performance.[31] Despite rapid research progress, experts note that detection technologies are struggling to keep pace with advances in generative AI - turning this into a technical "arms race" in which detectors are continually outpaced by more sophisticated generation models.[32]

c. The "Knowingly Permitted, Promoted, or Failed to Act" Standard

The proviso to Rule 4(1A) deems an intermediary to have failed in due diligence if it "knowingly permitted, promoted, or failed to act upon" synthetically generated information. While the phrase "knowingly permitted" is consistent with the actual knowledge standard under Section 79 of the IT Act, the addition of "failed to act upon" is imprecise and could imply a general monitoring obligation, contrary to the safe-harbour framework. The rule also refers to the use of "reasonable technical measures", but without clarifying what these entail or when inaction amounts to a failure. A clearer standard linking liability to actual knowledge or wilful disregard, would preserve proportionality.

Moreover, treating algorithmic amplification as promotion, could expose intermediaries to liability for ordinary recommendation functions, compelling over-censorship or removal of all synthetic content - an outcome neither practical nor consistent with the principle that intermediaries should not be held strictly liable for third-party content.

d. False Positives and User Impact

Current detection systems not only fail to identify all synthetic content but also incorrectly flag authentic content as synthetic (false positives).[33] This creates serious problems: users may find their legitimate content incorrectly labeled or removed, journalists may have authentic investigative material flagged as manipulated, and creators may face reputational harm from incorrect synthetic labels. The draft provides no mechanism for users to appeal incorrect determinations or seek correction of erroneous labels, creating a one-sided system where algorithmic errors carry no accountability.

RECOMMENDATION

A pragmatic approach is recommended to ensure transparency and verifiability. Implementation should be risk-based, with stricter obligations for high-risk content such as electoral material, financial fraud, or non-consensual intimate imagery, and lighter requirements for low-risk or clearly synthetic content. Technical obligations relating to labelling and detection need interoperable standards to be developed before the law mandates the same. Mechanisms like user disclosures, notice and action, community notes, etc may serve the intended purpose without hampering innocuous technology.

Conclusion

There is an urgent need to make internet a safer, more trustworthy space, particularly as synthetic media becomes increasingly accessible and capable of causing real harm. The proposed amendments represent an important step in that direction. To ensure that these safeguards are both effective and proportionate, it is essential that the framework clearly distinguishes between high-risk, deceptive uses of synthetic media and the vast spectrum of benign, beneficial, or routine digital practices. A more targeted formulation that is supported by workable technical standards, contextual labelling approaches, and a risk-tiered compliance model will not only strengthen user protection but also avoid unintended burdens on creators, developers, and intermediaries.

As MeitY continues refining the Rules, it will be important to take into account the operational realities of detection technologies, the multiplicity of actors in the AI value chain, and the need to preserve innovation, legitimate expression, and ease of use for *Digital Nagriks.*

A balanced, collaborative approach - grounded in technical feasibility and global best practices will ensure that these amendments achieve their intended purpose - curbing harmful deepfakes and malicious synthetic content while fostering a healthy, open, and vibrant digital ecosystem.

We hope these comments assist the Ministry in shaping a regulatory framework that is robust, future-ready, and reflective of India's democratic and technological aspirations.

Endnotes

1 European Union, Article 3(6), EU AI Act, https://artificialintelligenceact.eu/article/3/.

2 China, Article 17, Administrative Provisions on Deep Synthesis in internet-based Information Services. https://cyrilla.org/api/files/1728288988021wmhrfp2o9t.pdf.

3 Republic of Korea, Article 2(5), Framework Act on the Development of Artificial Intelligence and Establishment of Trust, https://cset.georgetown.edu/wpcontent/ uploads/t0625_south_korea_ai_law_EN.pdf.

4 European Union. Recital 133 read with Article 50, EU AI Act, https://artificialintelligenceact.eu/article/50/.

5 European Union, Article 3(6), EU AI Act.

6 *Rishabh Dara, Intermediary Liability in India: Chilling Effects on Free Expression on the Internet,* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2038214 (2011); C Arun and S Singh, *Online Intermediaries in India,* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2566952 (2015); Aradhya Sethia, *The Troubled Waters of Copyright Safe Harbours in India,* (2017) 12 Journal of Intellectual Property Law & Practice 398, 404.

7 MeitY, *Explanatory Note on Synthetic Media,* October 2025, https://www.meity.gov.in/static/ /2025/10/8e40cdd134cd92d d783a37556428c370.pdf

8 Google Support, *YouTube: Disclose Process for AI-generated Content,* https://support.google.com/youtube/answer/14328491?sjid=7781109938293688022- NC#disclose_process; Meta, *Misinformation Policy*, https://transparency.meta.com/policies/community-standards/misinformation/ similarly requires users to disclose "photorealistic video or realistic-sounding audio that was digitally created or altered" through its AI-disclosure tool, and may apply an informative label to content that "creates a particularly high risk of materially deceiving the public on a matter of public importance".

9 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008.

10 Framework Act on the Development of Artificial Intelligence and Establishment of Trust, https://perma.cc/CL3T-VHZ6.

11 Ibid.

12 Mozilla Research, *In Transparency We Trust,* 2024, https://www.mozillafoundation.org/en/research/library/in-transparency-we-trust/

13 ANemecek, Alexander, Watermarking Without Standards is Not AI Governance, 2024, https://www.researchgate.net/publication/392315400_Watermarking_Without_Standards_Is_Not_AI_Governance.

14 ibid.

15 EXIFData.org (2025), *Do Social Media Sites Strip EXIF Data?*, https://exifdata.org/blog/do-socialmedia-sites-strip-exif-data-2025-test.

16 X Zhao (2024), Invisible Image Watermarks Are Provably Removable Using Generative AI. https://proceedings.neurips.cc/paper_files/paper/2024/file/10272bfd0371ef960ec557ed6c866058-Paper-Conference.pdf.

17 Sven Gowal and Pushmeet Kohli, *Identifying AI-generated images with SynthID,* 2023, https://deepmind.google/blog/identifying-ai-generated-images-with-synthid/ ; *SynthID: Tools for watermarking and detecting LLM-generated Text,* 2025, https://ai.google.dev/responsible/docs/safeguards/synthid.

18 C2PA (2024), C2PA Specification 2.1, https://spec.c2pa.org/specifications/specifications/2.1/specs/_attachments/C2PA_Specification.pdf.

19 C2PA (2023), Implementation Guidance 1.4. https://spec.c2pa.org/specifications/specifications/1.4/guidance/Guidance.html; DeepLearning.AI (2024), C2PA Introduces Watermark Tech to Combat Media

Misinformation. https://www.deeplearning.ai/the-batch/c2pa-introduces-watermark-tech-to-combat-mediamisinformation/.

20 Article 50, EU AI Act, https://artificialintelligenceact.eu/article/50/.

21 Republic of Korea, Article 31, Framework Act on the Development of Artificial Intelligence and Establishment of Trust, https://cset.georgetown.edu/wpcontent/uploads/t0625_south_korea_ai_law_EN.pdf.

22 IGAP, *Global Legal Responses to Deepfakes: A Regulatory Primer*, 2025

23 Ibid.

24 Ibid.

25 Ibid.

26 Amerini I., et al. (2025), *Deepfake Media Forensics: Status and Future Challenges. Journal of Imaging*,11, 73. https://doi.org/10.3390/jimaging11030073.

27 Ibid.

28 Science Direct (2025), *DeepFake Video Detection: Insights into Model Generalisation — A Systematic Review,* https://www.sciencedirect.com/science/article/pii/S2543925125000075.

29 Science Direct (2025), Deepfake Video Detection Methods, Approaches, and Challenges, https://www.sciencedirect.com/science/article/pii/S111001682500465X#bib21.

30 Reuters Institute (2024), *Spotting deepfakes in a year of elections: how AI detection tools work and where they fail,* https://reutersinstitute.politics.ox.ac.uk/news/spotting-deepfakes-year-electionshow- ai-detection-tools-work-and-where-they-fail.

31 *supra note 21.*

32 Williams, Kaylee (2025), *What Journalists Should Know About Deepfake Detection Technology in 2025: A Non-TechnicalGuide*, https://www.cjr.org/tow_center/what-journalists-should-know-aboutdeepfake- detection-technology-in-2025-a-non-technical-guide.php; Optica (2025), *Generating and Detecting Deepfakes: A 21st Century Arms Race,* https://www.opticaopn. org/home/articles/volume_36/february_2025/features/generating_and_detecting_deepfakes_a_2 1st-century_arms_race.

33 supra note 24; Sora 2 *Deepfake Risks: Realism & Authenticity Challenges,* 2025,https://geneo.app/blog/sora-2-deepfake-authenticity-2025/.

INDIAN GOVERNANCE AND POLICY PROJECT

The Indian Governance And Policy Project (IGAP) is an emerging think tank focused on driving growth, innovation, and development in India's digital landscape. Specializing in areas like AI, Data Protection, FinTech, and Sustainability, IGAP promotes evidence-based policymaking through interdisciplinary research. By working closely with industry bodies in the digital sector, IGAP provides valuable insights and supports informed decision-making. Core work streams include policy monitoring, knowledge dissemination, capacity development, dialogue and collaboration.

For more details visit: www.igap.in

Contact us: relations@igap.in