# ***Comments on the Draft amendments to Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 – in relation to synthetically generated information***

**Submitted by:** CUTS International
Consumer Unity & Trust Society
www.cuts-international.org
October 2025

**Prepared by:** Sohom Banerjee, Senior Research Associate, CUTS International (Consumer Unity & Trust Society)

CUTS International
Since 1983

# Table of Contents

## Executive Summary

These comments present CUTS International's analysis and recommendations on the Government of India's proposed amendments to the *Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021*, particularly concerning synthetically generated or AI-generated information. The submission aims to support a balanced framework that enhances accountability and transparency while safeguarding innovation, user rights, and freedom of expression.

Key highlights include:

- **Clarity of Definitions:** The terms "synthetically generated information," "label," and "identifier" should be precisely defined and applied consistently to avoid over-broad interpretation and uneven compliance.
- **Proportionality and Risk-Based Labelling:** Labelling obligations should be graded according to the risk and intent of AI modification, distinguishing harmless creative or enhancement uses from deceptive or harmful manipulations such as impersonation or misinformation.
- **Feasibility and Innovation Safeguards:** The mandatory "≥10% visual area or first 10% of audio" labelling rule could unintentionally burden startups, creators, and small enterprises, degrade content quality, and limit innovation. A flexible, proportionate approach—supported by technical guidance—is recommended.
- **Procedural Safeguards and Due Process:** Platforms must not act as de facto adjudicators. Any moderation or takedown under these rules should follow transparent, auditable, and appealable procedures with user notice and response mechanisms.
- **Alignment with Global Standards:** India should harmonise its approach with international initiatives such as the Content Authenticity Initiative (CAI), C2PA standards, and the Partnership on AI to ensure interoperability and global consistency.
- **Regulatory Impact Assessment (RIA):** RIA and post-implementation review (within 12 months) should assess alternatives, stakeholder costs and benefits, and the rule's actual impact on innovation, usability, and compliance.
- **Public Awareness and Digital Literacy:** Effective labelling must be complemented by user education so that individuals understand the meaning and implications of AI-generated content.
- **Inclusive Oversight:** Civil-society and consumer organisations should be involved in monitoring compliance and evaluating the real-world effectiveness of labelling and watermarking mechanisms.

Overall, CUTS International supports MeitY's proactive initiative but emphasises the need for clarity, proportionality, and procedural fairness to ensure that the regulatory intent of transparency does not inadvertently stifle innovation or legitimate speech.

## 1. Introduction:

CUTS International commends the Ministry of Electronics and Information Technology (MeitY) for addressing the challenges posed by synthetically generated information (AI-generated content or "deepfakes"). The draft amendments to the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 are a timely step to strengthen user safety and platform accountability in the face of rapid advancements in generative AI[1]. We support the Government's vision of an Open, Safe, Trusted and Accountable Internet, and offer constructive inputs to ensure the rules are effective, balanced, and innovation friendly.

## 2. Rule 2(1)(wa) – Definition of "Synthetically Generated Information":

### 2.2 Text of the Draft Rule:

*"Synthetically generated information" means any 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* [2].

### 2.3 Comment:

CUTS International welcomes the introduction of a formal definition for "synthetically generated information" under Rule 2(1)(wa), which provides much-needed legal clarity in addressing the growing challenges posed by AI-generated content. The proposed definition is broad and technology-neutral, appropriately covering various formats such as text, audio, images, and video. However, we believe the framing could be further strengthened to improve precision, protect digitally vulnerable users, and avoid unintended overreach.

First, the definition should account for scenarios where only part of the content is artificially manipulated—by including the phrase "or any part thereof"—as many harmful instances involve hybrid combinations of real and synthetic elements. Second, the phrase "reasonably appears to be authentic or true" may place an unfair burden on users to discern manipulation, especially those less familiar with digital technologies. To shift responsibility toward those better positioned to act responsibly—namely creators and disseminators—we recommend reframing this as "intended to appear authentic or true." This adjustment would help protect users who may be misled by deceptive content, while also shielding artistic or humorous expression (e.g., satire, parody, or clearly fictional content) from disproportionate regulation. Enforcement must also be guided by a harm-based approach. Not all synthetic content is inherently problematic—examples such as fictional CGI, edited professional portraits, or clearly labeled AI-generated satire are not intended to deceive.

In contrast, content like deepfake impersonations, manipulated news items, and audio simulations for fraud purposes pose clear risks. The rule should therefore focus on harmful synthetic content

[1] https://www.meity.gov.in/static/uploads/2025/10/8e40cdd134cd92dd783a37556428c370.pdf#:~:text=The%20Government%20of%20India%20is,amendments%20to%20the%20Information%20Technology

[2] https://www.meity.gov.in/static/uploads/2025/10/9de47fb06522b9e40a61e4731bc7de51.pdf

where there is either demonstrable deceptive intent or an absence of safeguards. Additionally, the interpretation of this provision must be linked to Rule 2(1A) and Rule 3(1)(b), both of which provide essential due diligence guardrails by identifying specific categories of unlawful content. In this regard, it is helpful to explicitly list the kinds of content considered unlawful—such as impersonation, defamation, misinformation, non-consensual intimate imagery, or fraud. Importantly, intent and context should be codified in the law as guiding principles for both platforms and enforcement agencies. Finally, to enable effective implementation, appropriate capacity-building should be undertaken through training for content moderators and law enforcement agencies.

**2.4 Suggestion:**

CUTS International recommends that Rule 2(1)(wa) be retained with the following refinements:

1. Modify the language to include "or any part thereof" to reflect the partial nature of many harmful synthetic manipulations.
2. Replace "reasonably appears to be authentic or true" with "intended to appear authentic or true" to align liability with creators and better protect vulnerable users.
3. Ensure enforcement of this provision is explicitly linked to Rule 2(1A) and Rule 3(1)(b), which define unlawful content and intermediary obligations.
4. Issue accompanying guidelines listing both harmful (e.g., deepfake impersonations, fake news, fraudulent audio) and harmless (e.g., satire, artistic AI creations, edited professional images) examples to aid interpretation.
5. Codify the role of intent and context within the legal framework to avoid disproportionate regulation.
6. Invest in training and capacity-building to help platforms and law enforcement effectively identify and handle synthetically generated content.

These clarifications will help establish a more precise, proportionate, and user-centric governance framework for synthetically generated content in India's evolving digital ecosystem.

## 3. Rule 2(1A) – Applying "Information" to Include Synthetically Generated Information:

### 3.1 Text of the Draft Rule:

*Any reference to "information" in the context of an unlawful act in these Rules (including under Rule 3(1)(b), Rule 3(1)(d), Rule 4(2), and Rule 4(4)) shall include synthetically generated information*[3].

### 3.2 Comment:

The proposed Rule 2(1A) serves as a crucial clarificatory provision that formally brings synthetically generated content within the scope of the term "information" under the IT Rules, 2021. This inclusion helps eliminate ambiguity around whether AI-generated or manipulated content falls under the regulatory framework, especially in the context of unlawful acts as defined under Rules 3(1)(b), 3(1)(d), 4(2), and 4(4). However, it is important to underscore that this inclusion does not imply a blanket restriction on all synthetic content. Rather, regulatory consequences under these provisions should only be triggered when the content in question is unlawful or harmful in nature. This interpretation maintains a necessary balance between free expression and user safety.

In this context, the rule rightly ensures that existing due diligence and takedown obligations apply equally to synthetically generated content, without creating new categories of illegality. For example, non-consensual explicit deepfake images would logically fall under the existing category of "obscene or sexually explicit material" that violates an individual's dignity or privacy. However, we recommend that the government consider issuing guidance or illustrative lists—either in rules or an advisory—that explicitly mention such examples, as this will assist platforms and enforcement agencies in applying the rule with clarity and consistency.

On content such as deepfake audios that impersonate public figures or spread disinformation, the focus should remain on demonstrable harm and the requisite mental element—particularly when platforms are expected to act under the "knowingly false or misleading" standard. Given the challenges in establishing intent, especially at scale, enforcement mechanisms must be proportionate and context-sensitive. Similarly, enforcement in cases involving synthetic content that incites violence or promotes enmity must be guided by clearly defined legal thresholds to prevent overreach. We urge that this rule not be used as a basis for discretionary or excessive takedowns, particularly of political or satirical content.

With respect to Rule 4(2), the inclusion of synthetically generated content under the scope of traceability is consistent with the IT Rules' framework, but must be exercised with strong safeguards. The provision does not create new powers but clarifies that if an unlawful act involves synthetic content, originator information may be sought—subject to judicial oversight, necessity, and proportionality as per the Puttaswamy judgment. We reiterate our earlier recommendations to

[3] https://www.meity.gov.in/static/uploads/2025/10/8e40cdd134cd92dd783a37556428c370.pdf#:~:text=true.%20%28ii%29%20Clarificatory%20Inclusion%20,the%20context%20of%20unlawful%20acts%E2%80%94

the Ministry regarding guardrails for traceability: it should be invoked only in exceptional cases involving serious harm, and where alternate means have been exhausted. Further, clarity should be provided—either through FAQs or official commentary—that this inclusion does not expand the substantive scope of traceability, which remains confined to already-notified categories.

The clarification under Rule 4(4)—which mandates proactive detection of certain egregious categories such as CSAM or rape imagery—should similarly apply to AI-generated variants of such content. While this is a necessary step in future-proofing the regulatory response, we highlight that this alone may not substantially enhance enforcement outcomes. The availability and virality of AI-generated CSAM is an emerging concern, with global institutions such as NCMEC in the U.S., and regulators in the EU, Singapore, and Australia already responding with urgency. In India's context, we recommend that this be treated as a distinct area of policy and enforcement attention. MeitY may consider issuing sector-specific protocols for faster takedown, shared blacklists of known synthetic CSAM signatures, and collaboration with child safety NGOs and AI detection startups. The issue is too significant to be addressed solely through definitional expansion—it requires resource investment and institutional focus.

Overall, Rule 2(1A) is a necessary step to ensure that synthetically generated content does not escape accountability under India's digital governance framework. However, its operational success will depend on effective implementation, legal clarity, and consistent judicial interpretation that reinforces proportionality and safeguards civil liberties.

**3.3 Suggestion:**

CUTS International supports the intent behind Rule 2(1A), which seeks to clarify that synthetically generated content is subsumed within the broader definition of "information" under the IT Rules. However, we respectfully submit that the current framing of this provision may inadvertently apply a uniform regulatory lens to a highly diverse set of AI-generated content—without due regard to differences in intent**,** degree of modification, and platform reach. Such a "one-size-fits-all" approach may lead to disproportionate obligations on both creators and intermediaries and may chill innovation or lawful expression.

To ensure effective and proportionate implementation, CUTS suggests introducing a "harm-plus-intent" threshold in regulatory enforcement under Rules 3(1)(b), 3(1)(d), 4(2), and 4(4). Specifically, regulatory duties should be triggered only when synthetically generated content is demonstrably harmful or unlawful, and where there is reasonable indication that the modification was intended to deceive, mislead, or cause harm. This would insulate satire, parody, artistic expression, or minor aesthetic modifications from unintended regulatory scrutiny, provided that the intent to deceive is not apparent.

Additionally, we propose that the Ministry issue clarificatory guidance distinguishing between aesthetic or minor AI modifications (such as lighting adjustments or superficial image enhancement) and substantive synthetic alterations (e.g., face-swapping, synthetic voices, or misleading AI-generated imagery that alters meaning). Only the latter should attract regulatory scrutiny. Such nuance will allow both platforms and enforcement agencies to prioritize harmful cases without being overwhelmed by low-risk content.

Recognizing the diversity of digital platforms in India, we also recommend a tiered compliance model. Significant social media intermediaries (SSMIs) with large user bases and higher risk of virality can be expected to invest in proactive tools and swift takedown mechanisms. In contrast, smaller intermediaries should be required only to act upon complaints, without being burdened with technical mandates beyond their capacity. A stratified approach to intermediary obligations will promote balanced compliance without stifling smaller players.

To operationalize these principles, CUTS suggests including illustrative examples—either in the rules or an accompanying guidance note—of both harmful and innocuous synthetic content. For instance, deepfake pornography, impersonation of public figures for fraud, or AI-simulated riots could be cited as harmful use cases. Conversely, AI-generated political satire, stylized profile images, or fictional animation should be recognized as lawful or socially valuable. Examples will improve legal certainty and assist platforms and law enforcement alike.

In light of privacy concerns, particularly with respect to traceability under Rule 4(2), we urge that any effort to trace the originator of synthetic content be subject to rule-of-law safeguards. These include judicial oversight, necessity, proportionality, and demonstration that no less intrusive alternatives are available. The current amendment does not create new powers, but its interpretation must remain anchored to the original guardrails previously discussed when traceability provisions were introduced. We suggest that MeitY reiterate this understanding, including through FAQs or explanatory notes, to prevent overbroad interpretations.

Regarding Rule 4(4), we recognize the increasing threat posed by AI-generated child sexual abuse material (CSAM). However, we emphasize that the mere inclusion of synthetic content under "information" may not meaningfully advance enforcement unless accompanied by targeted protocols and technical investment. CUTS recommends development of a dedicated framework for synthetic CSAM, including platform guidance, detection benchmarks, training for law enforcement, and possibly a national reporting channel. Without this, regulatory inclusion may remain symbolic, despite the severity of the threat.

Finally, we stress that Rule 2(1A) must not be interpreted in the future to imply that all synthetic content is inherently unlawful or that failure to label it constitutes an offense. Any move toward mandatory labeling or default criminalization would require separate deliberation, stakeholder consultation, and appropriate safeguards to avoid curtailing lawful uses of AI tools.

## 4. Proviso to Rule 3(1)(b) – Protection for Removal of Synthetic Content (Safe Harbour):

### 4.1 Text of the Draft Rule:

*In Rule 3(1)(b), insert a proviso stating that if an intermediary removes or disables access to information (including synthetically generated information) in good faith, such action shall not invalidate the intermediary's protections under Section 79(2) of the IT Act* [4].

### 4.2 Comment:

CUTS International acknowledges the intent behind the proposed proviso to Rule 3(1)(b), which aims to clarify that intermediaries will not lose their safe harbour under Section 79(2) of the IT Act for removing or disabling access to synthetically generated content in good faith. This clarification addresses a longstanding concern in platform regulation: whether voluntary or proactive moderation by an intermediary may be construed as editorial action, potentially converting the platform into a "publisher" and risking loss of immunity. However, we believe that the current formulation may benefit from more defined contours to ensure it does not inadvertently encourage arbitrary, inconsistent, or opaque moderation practices.

To begin with, we seek clarification on whether the act of takedown alone—without evidence of content curation or promotion—can in fact transform an intermediary into a publisher. Indian jurisprudence has generally held that passive intermediaries do not become publishers unless there is active involvement in content creation or curation. Therefore, this risk may be overstated and should be clearly delineated in the rule's intent note or guidance, to avoid confusion.

More importantly, the phrase "in good faith" while well-intentioned, lacks operational clarity and opens the door to over-removal or misuse. Without a clear standard or process, this could result in platforms being overcautious—removing content not because it is harmful or unlawful, but to avoid any perceived risk. Moreover, in the absence of procedural safeguards, there is a risk that informal government suggestions or social pressure could be used to justify takedowns under the guise of "good faith." We believe that removal of content, even in good faith, must still comply with basic rule-of-law principles: there should be an internal rationale, documented reasoning, and wherever possible, an opportunity for the content uploader to be notified and heard before or soon after the removal.

This is particularly important given that different platforms may assess harm or legality differently. One platform's takedown may not set a precedent that others are obligated to follow. Hence, the rule must not be interpreted to imply that failure to act similarly by other platforms constitutes lack of "good faith." Such peer pressure-based enforcement risks creating unintended uniformity and may suppress diverse interpretations or legitimate content.

[4] https://www.meity.gov.in/static/uploads/2025/10/8e40cdd134cd92dd783a37556428c370.pdf#:~:text=true.%20%28ii%29%20Clarificatory%20Inclusion%20,the%20context%20of%20unlawful%20acts%E2%80%94

Furthermore, while platforms should be encouraged to act against abuse, we recommend that the protection under this proviso be limited to cases involving clearly abusive, exploitative, or harmful synthetic content, such as deepfake pornography, impersonation for fraud, or synthetic CSAM. Expanding the scope too broadly—especially to ambiguous cases like satire, political parody, or opinion-based content—without rigorous standards risks subjective interpretation and selective enforcement. For example, if a platform removes a derogatory deepfake about one public figure but retains similar content about another, it may be accused of acting inconsistently or discriminatorily. To guard against this, platforms must establish and publish content governance protocols that are transparent, consistent, and subject to audit.

Finally, there is a need for clear accountability if platforms' internal detection mechanisms (such as AI-based moderation tools) result in wrongful takedowns. Given that such tools can be fallible, we recommend that platforms record strong reasons for removal, maintain an auditable log of their moderation decisions, and provide a mechanism for users to appeal mistaken removals. A standardized notice-and-response framework, even post-facto where immediate removal is necessary, will ensure that user rights are protected without weakening the incentive for platforms to act responsibly.

### 4.3 Suggestion:

While CUTS International supports the intent of the proposed proviso—to offer intermediaries clarity and legal confidence in removing harmful synthetic content—we believe that the operational framework for implementing this protection must be carefully structured to avoid unintended consequences such as arbitrary censorship, inconsistent takedown practices, and erosion of user rights. The following recommendations aim to ensure the safe harbour is exercised with accountability, proportionality, and transparency.

#### 4.3.1 Clarify the Scope of "Good Faith" and Tie it to Defined Harm

The term "good faith" must be clarified to ensure that it is not used as a blanket justification for over-removal or discretionary action that lacks transparency. The safe harbour protection should be clearly limited to instances where the content is demonstrably abusive or harmful—such as synthetic child sexual abuse material, non-consensual deepfake pornography, impersonation for fraud, or clearly inciteful or violent misinformation. We recommend that MeitY issue guidance or FAQs to outline what qualifies as "harm" in this context. This would help platforms distinguish between protected speech (including satire and political commentary) and content that genuinely warrants intervention. Without this framing, platforms may default to an overly cautious stance and suppress borderline content.

#### 4.3.2 Strengthening Due Process and Accountability in Content Moderation

There is a significant risk that the safe harbour provision, if not accompanied by procedural guardrails, could lead platforms to act as de facto adjudicators of legality and harm, which is neither their mandate nor an appropriate substitute for rule-of-law processes. Platforms must not use this provision as a shortcut to avoid formal takedown mechanisms or as a shield for arbitrary

or opaque moderation decisions. We recommend that platforms be required to develop and publish internal community standards and harm assessment guidelines, particularly on how they determine what constitutes "synthetic misinformation," "harmful impersonation," or "manipulated content." These standards should be made publicly accessible and subjected to regular audits.

In addition, any takedown made under this provision must be accompanied by a notice-and-response framework. Content creators or uploaders must be notified of the takedown, provided with a brief explanation, and given a reasonable opportunity to appeal the decision. In urgent cases (e.g., involving graphic synthetic abuse or threats), the takedown may be immediate, but post-facto notice should still be mandatory. Without such processes, the provision risks being interpreted and applied inconsistently across platforms and could lead to legitimate content being suppressed without recourse.

#### 4.3.3 Establish Clear Documentation and Transparency Requirements

To enhance accountability and build user trust, we recommend that platforms maintain detailed records of each instance where content is removed under this safe harbour provision. These records should include the rationale, the type of content involved, and whether the action was triggered by user complaints, AI detection tools, or internal policy enforcement. Importantly, intermediaries—particularly Significant Social Media Intermediaries (SSMIs)—should incorporate disaggregated data on synthetic content takedowns into their mandatory transparency reports under the IT Rules. Such reporting would allow regulators and civil society to assess patterns, flag anomalies, and ensure that safe harbour protections are being applied to safeguard user rights—not curtail them.

#### 4.3.4 Limit to Clearly Abusive Contexts and Avoid Setting Precedents for Informal State Pressure

While voluntary removal in good faith may be necessary to address urgent harms, this must not be seen as a blanket endorsement of state nudges or informal diktats compelling platforms to take down content without legal scrutiny. Any implication that platforms are expected to act preemptively at the behest of non-transparent processes must be avoided. The law should clearly state that due diligence and internal assessments—not external pressures—must drive takedown decisions under this provision. This is particularly important given the sensitive political and social environment in India, where the line between misinformation and dissent can sometimes be blurred.

#### 4.3.5 Mitigate Risks of AI-Driven Errors

Where platforms rely on automated tools to detect harmful synthetic content, they should be held to a duty of care to verify such detections before removal. False positives—such as AI mistakenly flagging satire or art as harmful—can result in chilling effects on expression. Therefore, before invoking this safe harbour, platforms should ensure human review of flagged content and offer a clear redress mechanism. If content is wrongly removed, platforms should correct the decision promptly and notify the affected user.

## 5. Rule 3(3) – Due Diligence by Services Enabling Creation of Synthetic Media:

### 5.1 Text of the Draft Rule:

*Intermediaries that provide services for creating or modifying content using computer resources (leading to synthetically generated information) must ensure: (a) such content is labelled or embedded with a permanent unique identifier indicating it is synthetic, (b) the label/identifier is visibly displayed or audible in a prominent manner (covering ≥10% visual area or present in first 10% of audio), enabling immediate user identification of the content as synthetic, and (c) the intermediary shall not permit removal or alteration of this label or identifier* [5]

### 5.2 Comment:

The proposed Rule 3(3) introduces a critical requirement for services that enable the creation or modification of synthetic content through computer resources. It mandates that such content must carry a visible or audible label—embedded prominently—indicating that the information is synthetic in nature. The provision further prohibits removal or alteration of such labels.

While the intent behind this rule—enhancing transparency and traceability in synthetic content—is laudable, certain operational and definitional ambiguities need careful reconsideration to ensure the regulation remains proportionate, enforceable, and does not stifle innovation or user rights.

First, the terminology used in the draft—particularly "label," "identifier," and "watermark"—needs to be clarified and standardized. We recommend consistently using the term "label" as the official language in the rule and explanatory note to avoid confusion in implementation. Variance in interpretation of these terms could result in uneven compliance and inconsistent user experience across platforms.

Second, while a mandatory, permanent, non-removable label may strengthen traceability, it also imposes significant costs—technical, financial, and reputational—on developers, especially smaller startups and non-commercial creators. Embedding irreversible labels might degrade user experience, disincentivize legitimate AI-enabled creativity (such as artistic enhancement), and affect engagement levels. These unintended consequences merit further assessment.

Third, not all AI-generated or modified content carries the same risk or public interest implication. For instance, minor cosmetic alterations or AI-assisted enhancements (e.g., light adjustments, smoothing, background blur) are typically not harmful and do not mislead users. Subjecting such content to the same labeling standard as deepfake impersonations or synthetic pornography may dilute the seriousness of labeling, confuse users, and generate over-reporting.

Moreover, the current framing risks overbreadth. The threshold for what qualifies as "synthetically generated or modified content" is not clearly defined. Who decides what constitutes a "minor"

[5] https://www.meity.gov.in/static/uploads/2025/10/8e40cdd134cd92dd783a37556428c370.pdf#:~:text=3%283%29%5D%3A%20,computer%20resources%20enabling%20creation%20or

versus a "material" alteration? Without clarity, platforms may adopt an overcautious approach to avoid regulatory scrutiny, potentially harming innovation and speech rights.

There is also a risk of treating all labeled content as equally harmful, which may diminish the ability of users and platforms to distinguish genuinely misleading or malicious synthetic content from harmless or beneficial uses. This "flattening" of content risks overblocking or creating distrust in content labeling systems.

Further, we must highlight the importance of proportionality in visual and audio labeling requirements. The "10% of visual area" requirement, while possibly aimed at deterring removal, may significantly interfere with the aesthetic and usability of legitimate content—especially for creators working with detailed imagery. Similarly, applying a 10% time rule to audio (e.g., a 1-minute disclosure in a 10-minute file) may discourage engagement or cause confusion.

Finally, the comment section currently explains implementation mechanics in detail (e.g., watermark placement, encryption standards), which may be better suited for technical guidance rather than a regulatory submission. CUTS International recommends focusing instead on higher-level principles—transparency, proportionality, enforceability, and user rights protection.

### 5.3 Suggestions:

#### 5.3.1 Revisit the Labelling Requirement Through a Risk-Based Approach:

The government should adopt a graded framework for labeling synthetic content based on its potential for harm or misuse. This would help distinguish between benign uses (e.g., AI-assisted photo editing) and malicious manipulation (e.g., impersonation deepfakes or misinformation). This approach is consistent with emerging best practices globally and ensures that the regulation remains proportionate.

#### 5.3.2 Clearly Define Scope and Thresholds:

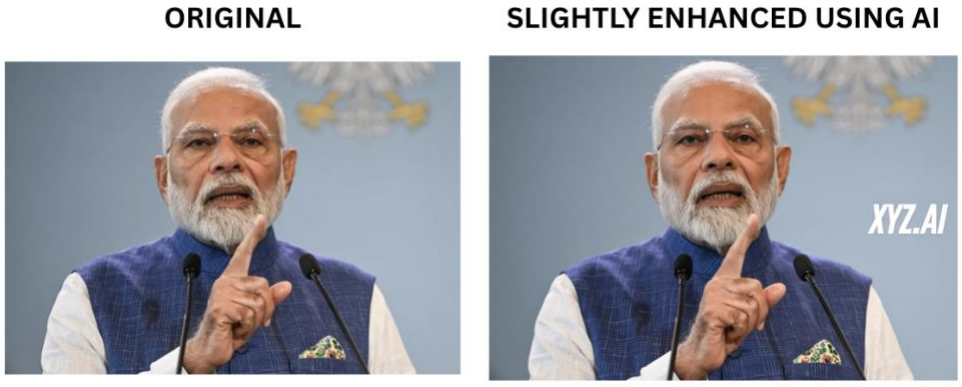
The rule should clarify what types of AI-generated or modified content are subject to mandatory labeling. We suggest incorporating a clause that excludes "minor cosmetic or aesthetic alterations that do not significantly alter the meaning, context, or authenticity of the content." For example, minor lighting or contrast adjustments using AI filters should not be within the scope of labeling.

#### 5.3.3 Assessing Real-World Implications of Mandatory Labelling

To support a proportionate, risk-based approach, we propose to undertake, for MeitY's consideration, a short evidence-gathering demonstration that tests the real-world effects of the mandatory "≥10% visible / first 10% audio" labelling requirement across common, benign AI uses (e.g., basic lighting/clarity adjustments, noise reduction) versus high-risk manipulations (e.g., impersonation deepfakes). The exercise will (i) produce before/after artefacts in image, video, and audio (about ten exemplars), (ii) document usability, aesthetic degradation, user comprehension, and engagement effects, and (iii) include targeted interviews with creators, SMEs, platforms, and consumer advocates.

**Illustrative Example: When Minor AI Enhancements Get Labelled**

The image below demonstrates a common, everyday use of artificial intelligence — AI-based enhancement for visual clarity and lighting correction. The photograph on the left is the original, while the one on the right has been slightly refined using an AI-assisted editing tool to adjust colour balance and sharpness.




***Source:*** *Image sourced from Gettyimages and subsequently edited by the author*
***Note:*** *The image has been generated and included only for illustration; it has no other intent or meaning.*

Under the proposed IT Rules [Rule 3(3)], even this minimal AI intervention would require the creator to label the content and cover at least 10% of the surface area with a prominent label. This could potentially compromise the messaging, diminish utility, raise concerns around authenticity of the entire content, without resulting in any effective benefits. This example underscores the practical challenge of over-regulation**.** Everyday users and professionals using AI for simple, non-deceptive tasks like improving lighting, contrast, or focus would be subject to the same compliance requirements as those creating fully synthetic or deceptive content. Problematic abuses of synthetic content like producing non-consensual intimate or obscene imagery will again be subjected to similar compliance requirements.

Consequently, the proposal disregards the intended use and extent of AI modification, and treats all uses (including abuses) and all kinds of AI modifications (from touch-ups to complete generation) similarly. Such blanket rule and one-size-fits-all approach risks degrading the visual experience**,** increasing compliance burdens**,** discouraging legitimate creative and professional uses of AI tools, while being of limited utility.

Findings will be consolidated into a brief to MeitY recommending a tiered labelling threshold that distinguishes minor cosmetic edits from materially altered or deceptive synthetic media.

### 5.3.4 Standardization Through Expert Consultation:

A technical standards body or expert group (comprising civil society, technologists, consumer advocates, and industry representatives) should be constituted to establish common guidelines on labeling format, placement, language, and duration. For instance, overly prescriptive requirements such as "10% area" or "first 10% of audio" may not be suitable across all use cases. Flexibility combined with recommended templates would be more effective.

### 5.3.5 Consider Practical Constraints on Small and Medium Enterprises:

The government should assess the cost, development time, and usability impacts of mandatory labeling—especially for startups and small developers. Open-source SDKs or implementation toolkits could be developed in collaboration with industry to facilitate adoption. There is also a need to anticipate potential negative user reactions (e.g., lower engagement with visibly labeled content) and balance user expectations with transparency needs.

### 5.3.6 Embed Rule of Law Principles and User Rights:

Platforms should not only be required to label synthetic content but must also follow transparent processes, provide reasons for content classification, and allow users to contest wrongful labeling. Labels should not be used as a blanket justification for suppression or de-prioritization of content without a reasoned process. This ensures compliance with due process and safeguards freedom of expression.

### 5.3.7 Engage with International and Multistakeholder Ecosystems:

India should align with ongoing international efforts such as the Content Authenticity Initiative (CAI), C2PA standards, and Partnership on AI. Doing so would help Indian regulations remain interoperable globally, support Indian developers building for global markets, and reduce fragmentation in content attribution.

### 5.3.8 Enforcement Should Be Proportionate and Tiered:

Any enforcement measures—such as withdrawal of safe harbour or takedown orders for non-compliance—must be gradual, beginning with warnings and implementation support. Severe penalties should be reserved for repeated, deliberate defaulters. App stores and hosting services may be involved in implementation.

### 5.3.9 Regulatory Impact Assessment (RIA) and Periodic Review:

We recommend conducting Regulatory Impact Assessment (RIA)[6] before full rollout, followed by a post-implementation review within 12 months. The RIA should systematically compare regulatory and non-regulatory alternatives, assess stakeholder costs and benefits (users,

[6] https://cuts-ccier.org/ria-research/

startups/SMEs, SSMIs, law enforcement), and identify likely impacts before arriving at a reasoned conclusion. The assessment should capture:

- Impact on innovation and startup ecosystem
- Efficacy of labels in informing user behaviour
- User engagement metrics and public perception
- Evidence of shifts to non-compliant or grey-market tools

### 5.3.10 Include Civil Society and Consumer Groups in Compliance Monitoring:

Civil society organisations and consumer rights groups should play a role in verifying whether AI services are labeling content correctly and whether users are aware of these labels. This will improve accountability and allow early feedback on problems faced by users.

### 5.3.11 Educate the Public on Labeling and Its Meaning:

Government-led digital literacy initiatives should include clear, accessible materials explaining what an "AI-generated label" means, how to interpret it, and what rights users have. Without public awareness, even well-labeled content may fail to achieve its intended purpose.

## 6. Rule 4(1A) – Additional Duties of Significant Social Media Intermediaries for Synthetic Content:

### 6.1 Text of the Draft Rule:

*Significant Social Media Intermediaries (SSMIs) providing services for users to upload/publish content shall, prior to publishing such content: (a) require the user to declare whether the content is synthetically generated; (b) deploy reasonable and proportionate technical measures to verify the accuracy of the user's declaration (taking into account the nature, format, source of the content); and (c) when a declaration or verification indicates the content is synthetic, ensure the content is clearly and prominently labelled as such on the platform*[7]. If an SSMI knowingly allows unlabeled synthetic content or fails to comply with these steps, it will be deemed as non-compliance with due diligence requirements. *(This rule applies only to content publicly posted via the platform, not to private or otherwise unpublished content.)*

### 6.2 Comment:

This is a pivotal provision that shifts some onus onto the major platforms where synthetic content eventually gets disseminated. Even with all upstream efforts (definition, labels by generators, etc.), people might still share unlabeled or malicious deepfakes on social media, messaging boards, video sites, etc. Rule 4(1A) ensures that *platforms act as gatekeepers*, adding checks at the point of content upload to catch synthetic media.

We support this concept because SSMIs (like Facebook, YouTube, Instagram, X/Twitter, ShareChat, etc.) are often where deepfake content goes viral and causes harm. These companies also have some of the most advanced AI capabilities and resources to tackle the problem, so it's appropriate to ask them to do more.

The two-layer approach (user self-declaration and platform verification) is wise:

- **User Declaration:** Requiring users to self-declare whether uploaded content is synthetically generated introduces a layer of accountability. While the intent is to create a culture of transparency, it is important to acknowledge that many users—especially those less familiar with the technical underpinnings of generative AI—may not clearly understand what constitutes "synthetic content." For instance, individuals using apps with automatic filters or minor AI-based enhancements (such as brightness adjustment, beautification, or background edits) may be unsure whether such modifications qualify as synthetic. Without clear guidance, these users may over-disclose (labelling trivial edits unnecessarily) or under-disclose (assuming more significant AI edits don't count), leading to inconsistent compliance and potential user fatigue.

  To mitigate this, platforms must provide clear, accessible, and multilingual explanations (e.g., through tooltips or brief examples) distinguishing between content that must be

[7] https://www.meity.gov.in/static/uploads/2025/10/8e40cdd134cd92dd783a37556428c370.pdf#:~:text=%E2%80%A2%20Enhanced%20Obligations%20for%20SSMIs,Open%2C%20Safe%2C%20Trusted%20and%20Accountable

labelled and content that does not require declaration. A graded approach could also be considered, where only substantial or manipulative AI modifications trigger a declaration requirement. This would help prevent confusion and ensure the system remains user-friendly, particularly for creators or individuals with limited digital literacy.

- **Automated Verification:** Given not all will declare truthfully or even realize something they share is synthetic (imagine someone forwarding a deepfake video not knowing it's fake), the platform's tools act as a safety net. "Reasonable and proportionate" suggests it's not an absolute duty to catch everything, but to have a sensible system given the state of tech and risk. So, an image posted might be scanned against known deepfake databases or analyzed by an AI model to assign a probability it's AI-generated. If high probability, and user said "No", then platform can flag it.

- **Labelling on Platform:** If content is identified as synthetic, the platform must label it clearly. This likely means adding a visible disclaimer overlay or caption like "Manipulated Media" or "AI-generated" on the post. Some platforms already do something similar (Twitter/X tags some fake media as "manipulated", Instagram has worked with fact-checkers to label misinformation images, etc.). This ensures that even if the content itself wasn't watermarked by its creator, the platform will inform its users. It's an important transparency measure and it prevents the scenario of e.g. a deepfake video appearing in someone's feed with no context – instead, users will see a notice that it's not real, possibly with an explanation or link to fact-check.

### 6.3 Global Alignment:

As noted earlier, the EU AI Act will make such verification and labeling obligatory for platforms by 2026[8]. The EU's Digital Services Act (DSA) also has related provisions where very large platforms must assess and mitigate risks of disinformation, including those possibly amplified by deepfakes[9]. We also have seen voluntary movement: Facebook/Meta announced in 2024 that they would start labeling "AI-generated" content in certain contexts; TikTok instituted a policy requiring users to label AI content and even introduced an automatic label for some content in 2023. So, the rule essentially codifies what is emerging as a best practice: platform-level content authenticity checks.

### 6.4 Feasibility:

There will be challenges. Technically, how effective are current deepfake detectors? As discussed, not perfect, but they are getting better. Also, different media types require different detectors (image vs video vs audio vs text). SSMIs will likely focus on the media types they host (e.g., a text-based platform might look for AI-written text that impersonates someone, which is a different

[8] https://blackbird.ai/blog/deepfake-detection-required-eu-ai-act-blackbird-ai-compass/#:~:text=readable%20formats,this%20reality%2C%20requiring%20deployers%20to

[9] https://iapp.org/news/a/artificial-illusion-global-governance-challenges-of-deepfake-technology#:~:text=Election%20integrity%20and%20misinformation%20laws,of%20deceptive%20deepfake%20election%20content

challenge than images). They may also have to rely on user reports in some cases where automated detection fails. The rule just says they must deploy measures; it doesn't guarantee they catch all. And that's fine if they try.

Another important consideration is the potential friction introduced into user workflows. Requiring users to make a declaration—such as ticking a checkbox to indicate whether content is AI-generated—may appear minimal, but at scale it could lead to significant user fatigue, particularly for everyday creators or casual users. This added step in the upload flow might discourage engagement, especially among those unfamiliar with AI tools or unsure about whether their content qualifies as synthetic.

To mitigate this, the design of the declaration interface must be simple, context-aware, and supported by clear examples or prompts. Platforms could consider adaptive prompts that only appear when AI-generated characteristics are detected or suspected, rather than applying a universal friction layer. For instance, a lightweight back-end risk scoring mechanism could trigger the declaration step selectively—this would reduce unnecessary burdens on the majority of users uploading non-synthetic content.

It is also important that this requirement not create disincentives for content creation or sharing, especially by independent creators, educators, and artists. Thus, implementation must balance the need for accountability with preserving a smooth and inclusive user experience. Any UI change should be tested with user feedback loops before full deployment. If the experience feels punitive or unnecessarily bureaucratic, it risks eroding user trust and platform vitality.

Many users might not fully understand the definition, but one can hope the platform provides a tooltip ("e.g. image created with AI, deepfake, etc."). The verification step (scanning content) could introduce some processing delay, but likely negligible for images (a second or two maybe) and more for videos. SSMIs handle millions of uploads daily, so the systems must be efficient and scalable. They might not scan everything fully, but maybe use risk-based triggers (e.g., new account uploads video to scan thoroughly; known verified news outlet uploads to maybe trust more). The rule says "having regard to nature, format, source" which hints at such context-based approaches – meaning an SSMI can calibrate their verification intensity based on the scenario. This flexibility is important for feasibility.

### 6.5 Enforcement and Avoiding Overreach:

The rule's strong linkage between non-compliance and the loss of due diligence protections creates significant pressure on SSMIs to detect and appropriately label synthetic content. While this acts as a deterrent, it also risks incentivizing platforms to adopt overly cautious or aggressive moderation approaches. For instance, platforms may label content as "synthetic" merely on suspicion or based on imperfect detection tools, resulting in mislabelling of genuine user-generated content. This can diminish user trust and lead to inadvertent censorship. To mitigate such risks, platforms must be required to act responsibly, transparently, and in accordance with rule of law principles—such as providing prior notice, opportunities for user response, and reasoned decisions. Additionally, civil society organisations and consumer groups must be involved in the oversight and evolution of these mechanisms to ensure user rights and fairness are safeguarded.

Collaborative enforcement through multi-stakeholder dialogue, including public consultation and pilot testing, is essential before any punitive consequences are imposed. Such a participatory approach will ensure that enforcement is effective without being excessive or arbitrary.

**6.6 Privacy Consideration:**

The rule explicitly does not apply to private content – that's good. So, if an SSMI has both public and private channels (like Facebook has both timeline posts vs private Messenger chats), only the public ones are governed here. That distinction maintains privacy of personal communications (where automated scanning for AI content would be invasive and maybe not legally tenable without specific law). For public posts, scanning them is fair game since platforms already do it for moderation. No new privacy issue there beyond what's normal content moderation.

**6.7 User Rights:**

It is essential that any action by platforms to label or remove content on the basis that it is synthetically generated be undertaken in a manner that upholds fundamental principles of due process. Users must be provided with prior notice before their content is tagged as AI-generated or removed for failing to carry a label. Crucially, they should be afforded an opportunity to explain the nature of their content and respond to the platform's assessment before any final action is taken. Platforms must ensure that decisions are reasoned, transparent, and based on clearly articulated standards, rather than automated or opaque determinations. As a consumer-facing regulatory measure, this rule must be implemented in line with rule of law principles. Furthermore, in case of disputes, there should be accessible redressal pathways—such as an appeals mechanism involving expert review panels that understand the nuances of synthetic media and its detection. These procedural safeguards will protect users from arbitrary actions, preserve trust in the system, and ensure that legitimate expression is not unintentionally curtailed.

**6.8 Benefits:**

If effectively implemented, Rule 4(1A) can serve as an important user-facing measure to raise awareness about synthetic content on platforms. For instance, in scenarios such as a fake video falsely attributed to a government official, a clear synthetic label can act as a warning, helping to prevent the spread of panic or misinformation. Over time, such markers—if reliable and responsibly applied—can create user expectations and drive cautious sharing behavior, much like the "forwarded" label on WhatsApp or blue tick verifications.

However, we caution against the emergence of a binary perception among users—that all unlabelled content is authentic, and all labelled content is necessarily synthetic. Due to the limitations of detection technology and the sheer volume of uploads, many AI-generated pieces may go undetected and remain unlabelled. Conversely, real content could occasionally be misclassified as synthetic due to technical false positives. Therefore, users must be educated that labels are indicators, not absolute judgments. This nuance must be reflected in platform messaging and public awareness campaigns.

We also recommend that MeitY, civil society organisations, and platforms work together to design transparent, balanced communication strategies that explain what the labels mean, their limitations, and how users should interpret them. This approach will help prevent misplaced trust in unlabelled content and undue skepticism towards genuine materials.

### 6.9 Suggestion:

We recommend that MeitY:

- **Provide clarity on enforcement**: Maybe issue standard operating procedures or a guidance note for SSMIs on how to demonstrate compliance. This could include maintaining logs of how many uploads declared as AI, how many flagged by the system, etc. Possibly an audit mechanism can be introduced where SSMIs submit annual reports on this (maybe as part of their transparency reports).

- **Set up a technical expert group** (perhaps under the AI Safety Institute or a new taskforce): Where SSMIs can share challenges and best practices in implementing these tools. This could be an informal knowledge-sharing platform. Since all SSMIs (the big ones) face the same problem, collaboration could be beneficial, even on things like creating open databases of known deepfake hashes or developing open-source detection models. A government-facilitated forum including consumer groups, CSOs and Think Tanks might accelerate progress.

- **Gradual ramp-up:** Possibly start with certain content types (like deepfake videos, which are high-risk) for aggressive detection, and expand as technology matures (like detecting AI-generated text is still very hard, so maybe focus less on that at first).

- **Public communication:** When these mechanisms are in place, the government and platforms should inform the public. A positive PR could be: "To protect users, platforms in India will now label AI-generated content – here's what to look for." This will prepare users and put would-be malicious users on notice that their fake posts will likely get flagged.

- **Alignment with Legal Framework:** It might be prudent to check that this requirement meshes with any obligations under the upcoming Digital India Act or other guidelines, to ensure consistency. Since IT Rules might eventually fold into a new framework, continuity of such pro-user provisions should be preserved.

***Illustrating Potential Impact – Both Positive and Negative Scenarios***

To understand how Rule 4(1A) might function in practice, consider the following example during an election season. A video clip surfaces online showing a leading political candidate making inflammatory remarks. With the proposed rule in place, the uploader is prompted to declare whether the video is AI-generated. If they lie or are unaware, the platform's AI detection systems may still flag the content based on known manipulation patterns. It is then labelled as "Synthetic/Altered Content," enabling users, media outlets, and election monitors to treat it with caution, potentially preventing widespread misinformation and electoral disruption.

However, things may also go wrong. Imagine a citizen journalist uploads a genuine video of a protest or police misconduct, but the platform's algorithm mistakenly flags it as synthetic—perhaps due to poor video quality, compression artifacts, or the presence of overlays. The video is labelled as AI-generated or blocked pending review. As a result, legitimate speech is suppressed, public awareness of a critical issue is delayed, and the uploader struggles to appeal the decision due to opaque review mechanisms. This chilling effect could discourage whistleblowers or independent voices from posting real content, fearing mislabelling or suppression.

These examples underscore the need for balance: while proactive labelling can mitigate harm, robust safeguards—including notice, an opportunity to respond, and transparent review standards—are essential to avoid unintended censorship or chilling effects on speech. Rule 4(1A) must therefore be implemented with a rights-respecting, proportional, and error-aware framework.

So, Rule 4(1A) is a critical accountability measure for platforms. It leverages the resources of large tech companies in service of truth and safety in the information ecosystem. We support its inclusion, with the caveat that it should be enforced in a manner that's *practical and fair*, allowing iterative improvements. If done right, it will make Indian social media space notably more resilient against AI-based misinformation than many other countries – potentially a model for others.

## 7. Global Best Practices and Indian Context:

In crafting these recommendations, we draw not only on CUTS International's own research and stakeholder consultations[10], but also on lessons from other jurisdictions and expert studies:

- The European Union (EU) has moved forward with robust AI governance through the EU AI Act, which among other things will *require AI-generated content to be labelled and watermarked*, and require platforms to monitor and mitigate AI-driven misinformation[11]. The EU also differentiates contexts (allowing exceptions for legitimate creative or journalistic use) and imposes steep penalties for non-compliance[12]. India can take cues from the EU in setting high standards for transparency while also incorporating context-based flexibility.
- In the United States, while no comprehensive federal law on deepfakes exists yet due to free speech considerations, the approach has been a mix of voluntary commitments and narrow laws. The White House secured voluntary pledges from AI companies to implement watermarking and other safety measures in 2023, and an Executive Order (Oct 2023) now pushes for industry standards on content authentication[13]. Some U.S. states have enacted laws targeting specific malicious deepfakes – for example, laws in California and Texas criminalize deepfakes intended to influence elections or deepfake pornography targeting individuals. These reflect a consensus that certain *uses* of synthetic media are so harmful that strong action (including criminalization) is warranted. India's draft amendments stop short of criminal law changes (focusing on intermediary duties), but we note that Indian Penal Code and IT Act provisions on impersonation, fraud, obscenity, etc., can already apply to the creation or misuse of deepfakes. Going forward, India may also consider targeted criminal penalties for worst forms of deepfake abuse (such as non-consensual sexual imagery), akin to laws in the UK, Australia and Singapore that directly outlaw such acts.
- Singapore has been proactive through the Protection from Online Falsehoods and Manipulation Act (POFMA), 2019. POFMA empowers authorities to issue swift correction orders or takedowns for online falsehoods, including deepfakes that could threaten public order or election integrity[14]. Under POFMA, if a video circulating is found to be a deepfake (say, misrepresenting a politician's speech), the government can require platforms to label it as false or remove it. This strong-arm approach has been effective in curbing

---

[10] https://cuts-ccier.org/pdf/reimagining-content-moderation-strategies-in-the-age-of-generative-ai.pdf#:~:text=Collaborative%20AI%20Governance%3A%20Relying%20solely,Through%20collaboratively%20designed%20codes

[11] https://iapp.org/news/a/artificial-illusion-global-governance-challenges-of-deepfake-technology#:~:text=Platform%20accountability%20and%20regulation,identities%2C%20and%20prevent%20deceptive%20deepfakes

[12] https://blackbird.ai/blog/deepfake-detection-required-eu-ai-act-blackbird-ai-compass/#:~:text=Every%20visual%20asset%20your%20organization,impact%20both%20finances%20and%20reputation

[13] https://www.getclarity.ai/ai-deepfake-blog/white-house-executive-order-demands-watermarking-of-ai-content#:~:text=The%20executive%20order%20directs%20the,advancements%20in%20generative%20AI%20technologies

[14] https://iapp.org/news/a/artificial-illusion-global-governance-challenges-of-deepfake-technology#:~:text=introduced%20election,of%20deceptive%20deepfake%20election%20content

misinformation, though it also raises debates on oversight and free expression. Additionally, Singapore amended its Penal Code in 2020 to criminalize the creation or distribution of non-consensual intimate deepfakes, recognizing the trauma such content causes. The Indian Government's draft rules resonate with Singapore's emphasis on traceability and labeling but take a less centralized approach by placing responsibility on platforms and creators rather than direct state intervention in each incident. This is appropriate for India's democratic context, but we can learn from Singapore's experience the importance of clarity in law and swiftness in response when dealing with viral deepfake harms.

- Other jurisdictions like China have implemented strict rules under the Deep Synthesis Regulations (2023), mandating not only labeling of AI-generated media but also requiring platforms to verify user identities and even pre-review certain content. While China's regulatory style differs (given its controlled internet space), the underlying concept of *provenance and accountability* is similar. We mention this to highlight that around the world, transparency requirements for synthetic media are becoming the norm, and India's framework will be part of a global mosaic of AI governance. Cooperation on standards (e.g., sharing best practices on watermarking tech) could be mutually beneficial.

## 8. CUTS International's Research & Stakeholder Insights:

Our own study, *"Reimagining Content Moderation Strategies in the Age of Generative AI*[15]*" (2023)*, involved extensive stakeholder consultations. We found that a risk-based, context-sensitive approach is vital. Broad-brush measures can inadvertently stifle permissible expression or technological innovation. For example, overly stringent labeling might discourage positive use-cases like AI in education or satire, while overly lenient rules could fail to prevent serious harm. Thus, we advocate for dynamic regulation: start with baseline rules (like these amendments) and refine through multistakeholder inputs. We echo the recommendation for establishing an AI Safety Institute (AISI) or similar independent body in India. Such a body can bring together government, industry, civil society, and technical experts to formulate detailed codes of practice, monitor the impact of the rules, and advise on updates. It can also help devise harm classification – distinguishing low-risk synthetic media (e.g. obvious filters) from high-risk ones (realistic political deepfakes) – and calibrate responses accordingly. Additionally, our research emphasizes empowering the public. This includes not only media literacy but also providing tools for users to verify content themselves. In the long run, innovations like browser plugins or platform features that allow users to check if an image has a verified origin (using the embedded metadata) could be promoted. By building a culture of verification ("*Don't trust until verified*"), we strengthen societal resilience against deepfake-driven deception.

[15] https://cuts-ccier.org/pdf/reimagining-content-moderation-strategies-in-the-age-of-generative-ai.pdf#:~:text=Collaborative%20AI%20Governance%3A%20Relying%20solely,Through%20collaboratively%20designed%20codes

## 9. Other Considerations and Recommendations:

In addition to the rule-wise comments above, we offer some cross-cutting suggestions to bolster the regulatory approach to synthetically generated content in India.

### 9.1 Multi-Stakeholder Collaboration and Capacity Building:

The fight against malicious deepfakes cannot be won by government or industry alone. We recommend establishing a collaborative framework, perhaps through the proposed AI Safety Institute (AISI) or a dedicated advisory committee, that includes platform representatives, AI developers, civil society, academia, and law enforcement. This body can advise on evolving trends in AI misuse and craft codes of practice to supplement these rules. For example, a code of practice might detail how platforms should handle political deepfakes during elections, or how companies should design user interfaces for AI disclosure. It can also create training modules for police and judges on synthetic media issues, as well as awareness programs for journalists and educators (who can in turn inform the public). As generative AI evolves (deepfakes moving into real-time video calls, or synthetic text bots on forums), policies will need nimbleness – an ongoing conversation with stakeholders will provide that nimbleness through consensus rather than constant formal amendments.

### 9.2 Research and Innovation in Detection:

We applaud the work by Indian institutions like C-DAC in developing deepfake detection tools. Continued investment in R&D is crucial. The government could fund research challenges or hackathons to improve deepfake detection (especially for content in Indian languages and contexts, which global tools might not optimize for). Equally, research into robust watermarking, crypto-based provenance (like using blockchain to record content origin), and authentication techniques (e.g., hardware signing of camera-authored footage to distinguish from AI) should be encouraged. India's talented AI researchers and startups should be incentivized via grants or recognition (e.g., an "Open, Safe Internet" innovation award) to contribute solutions that help comply with these rules. This not only aids compliance but could spawn Indian IP and leadership in a cutting-edge tech domain.

### 9.3 Legal Alignment and Coherence:

The Government of India may consider whether other laws need updating to complement these rules. For instance, data protection law (the DPDP Act) could come into play if biometric data (like someone's face or voice) is processed by AI without consent – currently, the DPDP doesn't specifically address that scenario, but globally GDPR and others treat such misuse as a privacy violation.

As mentioned, IPC covers impersonation and defamation, but having a specific offense for egregious deepfake creation (especially of intimate images) could strengthen deterrence, similar to the UK's and Australia's approach for sexual deepfakes. Any such changes should be carefully

evaluated and are beyond the scope of this consultation, but we flag them for a holistic strategy. Meanwhile, the IT Rules amendments will serve the immediate need under existing law.

**9.4 Global Cooperation:**

Deepfake threats are global – fake videos cross borders, and the expertise to tackle them can be shared. India should actively participate in international dialogues on AI and online content regulation. For example, engagement with the EU's forthcoming AI Office once the AI Act is in force, or with UNESCO's initiatives on misinformation, could be beneficial. Perhaps bilateral cybersecurity or tech agreements can include cooperation on detecting and countering deepfakes (much like countries cooperate on cybercrime or terrorism content). If India leads by example with these rules, we can also lead in pushing for some norm-setting globally – such as advocating for an international standard on AI content disclosure. This will help Indian platforms if other jurisdictions also require similar behavior from the global tech giants.

**9.5 User Empowerment and Digital Literacy:**

We cannot overstate the importance of educating users in parallel with regulation. The government, along with NGOs and educational institutions, should integrate media literacy modules that include deepfake awareness. School curricula could cover basic lessons on identifying fake images or videos. Public service media and news outlets can run programming on "fake vs fact" showcasing deepfakes. The more aware the average user is, the less likely a deepfake will achieve its malicious goal. We suggest MeitY partner with the Ministry of Education and PIB's fact-check unit, etc., to disseminate knowledge. Additionally, enabling and promoting reporting channels is key – users should know where to report a suspected deepfake (to the platform, to law enforcement if it's criminal like morphing). A national online portal (perhaps under CERT-In or Cyber Crime reporting portal) could have a section for deepfake incidents, guiding victims on steps to take (legal aid, how to get content removed, psychological support if needed in cases of harassment). A holistic support system for victims will underscore the government's commitment to protecting citizens from this new-age menace.

**9.6 Balancing Act – Preserving Free Expression and Innovation:**

While tackling the dark side of AI, we must ensure not to cast a chilling effect on the beneficial uses of these technologies. India's regulatory stance, as indicated by the IT Minister's comments, favors a calibrated approach that is not overly heavy. We echo this – the goal is not to create onerous compliance that only Big Tech can meet. Thus, in rolling out these rules, authorities should be mindful of the capacity of startups and the creative community. If needed, consider a bit of flexibility or phased enforcement for smaller entities (while expecting higher standards from the biggest players who have more resources). Also, we should protect parody, satire, and artistic freedom. A comedian using a deepfake for parody, or a filmmaker using AI for special effects, should not fear legal reprisals if no harm is intended and perhaps a disclaimer is provided. The rules themselves don't criminalize any creation; they just impose labeling – which is fine. But how society and law enforcement react matters. We hope these rules won't be misused to target artists or political dissent under the guise of "it's synthetic". That circles back to having clear definitions

and context of "unlawful act" – i.e., only if the content violates an existing law or someone's rights is action triggered. Free speech principles as in the Constitution and court precedents remain paramount.

### 9.7 Monitoring and Review of Regulation:

CUTS International recommends that MeitY commit to a formal review mechanism to assess the effectiveness and proportionality of these rules after one to two years of implementation. Given the rapidly evolving nature of AI technologies—including emerging use cases such as real-time deepfakes, immersive virtual content, and AI-generated avatars—periodic recalibration of the regulatory framework is essential. A structured Regulatory Impact Assessment (RIA) process should be integrated into this review. This would allow policymakers to measure intended and unintended outcomes, assess cost-effectiveness, and evaluate whether the regulation continues to serve its objectives without imposing undue burden or stifling innovation.

Such a review process should be participatory and transparent, involving diverse stakeholders including civil society organisations (CSOs), consumer groups, academia, technologists, and platform representatives. In particular, consumer groups bring grounded insights into user harms and usability challenges, which are crucial for assessing the social impact of labelling, detection, and takedown systems. CSOs can help flag emerging misuse patterns or identify gaps in enforcement and redress, especially on behalf of vulnerable or digitally inexperienced users.

The review should consider metrics such as compliance rates among SSMIs, the number and nature of synthetic content labels issued, appeals and false positives, reported user harms, and the actual deterrence effect on harmful deepfake circulation. It should also include an evaluation of whether adequate notice and grievance mechanisms have been implemented by platforms. Moreover, MeitY could consider publishing periodic transparency reports or synthetic media monitoring bulletins, which highlight trends, enforcement benchmarks, and system-level learnings. This feedback loop would ensure the regulation remains dynamic, accountable, and responsive to the evolving AI ecosystem.

## 10. Conclusion:

CUTS International welcomes the Government of India's efforts to proactively address the growing risks posed by synthetically generated content through the proposed amendments to the IT Rules, 2021. We recognise the importance of enhancing transparency, accountability, and safety in India's digital ecosystem while safeguarding user rights and preserving space for innovation and expression.

Our comments and suggestions reflect a commitment to practical and rights-respecting implementation. We emphasise the need for proportionality, contextual interpretation, clear procedural safeguards, and multi-stakeholder engagement—including consumer groups and civil society—to ensure the rules are not only enforceable but also fair, inclusive, and future-ready. As generative AI continues to evolve, the regulatory framework must remain adaptive, grounded in rule of law principles, and responsive to the lived realities of Indian internet users.

We look forward to continued engagement with the Ministry and other stakeholders to ensure that India's approach to synthetic content governance sets a global benchmark for balancing innovation with harm mitigation.

*This submission has been prepared by CUTS International (Consumer Unity & Trust Society)*

*For any queries, please contact: Sohom Banerjee | Senior Research Associate |* sje@cuts.org