IT for Change's Submission on the Draft Amendments to the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, Relating to Synthetically Generated Information

We appreciate the Ministry's intention to address the harms caused due to deepfakes and other synthetically generated information through the proposed amendment to the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 ("IT Rules"). We would like to highlight some of our concerns with the proposed amendment, so that its objectives may be met effectively.

1. Labeling and user declarations are not adequate to address the virality of synthetically generated content: The harm from synthetically generated content often stems from its virality on digital platforms like social media. Therefore, while user declarations and labeling of synthetically generated content may be useful measures, these do not directly address the propensity of platform algorithms to amplify harmful content. Therefore, we recommend that the due diligence requirements for significant social media intermediaries with respect to synthetically generated content must also mandate technological measures to arrest the spread of viral and harmful synthetically generated information, conducting of risk assessment, documentation, transparency, and audit of platform recommendation algorithms, and investments in training human moderators to determine the veracity and the unlawful or harmful nature of the content.

2. IT Rules cannot be a substitute for a comprehensive governance framework for AI: An important concern about the proposed amendment is that it could become an inadequate substitute for a much-needed, comprehensive regulation on AI services, especially content-generating AI services such as text and image generators. We submit that the regulation of AI services that generate content should not be brought under the IT Rules, as it may be interpreted to mean that entities that provide or deploy such services enjoy safe harbor protection for the content that they generate. This would be the very antithesis of intermediary liability law that pertains only to the liability of intermediaries for third-party content carried by them and does not apply to content that is generated by them. The recently released India AI Governance Guidelines already recognize this problem and the need to define clearly, through legislation, the roles of various actors in the AI value chain (developer, deployer, users, etc.). The Government of India should, therefore, enact a separate legislation that provides a comprehensive governance framework for AI that goes beyond mere labeling and metadata requirements and calls for risk assessment, documentation, audit, and transparency obligations.

Our detailed submissions are below.

1. Rule 3(3)

Recommendation:

Expressly exclude content-generating AI services from the scope of safe harbor protection, and delineate obligations of providers or deployers of such services through a separate law.

Rationale:

The proposed amendment is at best a partial response to the harms from synthetically generated content. The root issue here is that AI models deployed on digital platforms allow for the large-scale generation of content that is harmful, discriminatory and/or violent—ranging from sexually explicit deepfake images, stereotypical depictions of marginalized communities, to violent imagery. The algorithmic architecture of platforms—fed by data points, model weights, and parameters—actively facilitates the generation of such content; the lack of guardrails to prevent or mitigate the harm arising from such content compounds the issue. While the addition of labels, metadata, and identifiers, as proposed under the amendment, can help users in ascertaining the veracity of synthetically generated content on digital platforms like social media, it cannot address the broader risks posed by the large-scale generation of harmful content through AI services and its widespread circulation.

Moreover, the regulation of synthetically generated content through the IT Rules creates an unintended legal risk. Rule 3(3), which outlines due diligence requirements in relation to synthetically generated information, is intended to apply to cases where an intermediary offers a computer resource which may, amongst others, enable the creation or generation of information as synthetically generated information. This could be interpreted to include entities that provide content-generating AI services such as text and image generators. Such an interpretation would effectively bring providers or deployers of generative AI systems within the ambit of "intermediaries" for the purpose of this rule. This inclusion would be problematic, as it may allow such providers or deployers to claim safe harbor protections by merely complying with minimal obligations such as labeling and metadata tagging. In effect, this would exonerate them from liability for harmful content generated by their systems via user prompts. This is undesirable as the output of such systems is inextricably shaped by the technological choices of their developers or deployers — such as the selection of training data, model weights, parameters, and inadequate bias testing and risk assessment. It is important to disabuse the notion that AI is neutral and its results are solely based on user prompts.

To holistically address the risks posed by the generation of synthetic content through AI services, and to avoid regulating providers or deployers of content-generating AI services under a provision meant to address the liability of intermediaries for third party content, we recommend the enactment of a separate law, which comprehensively defines the obligations and liabilities of all actors involved in the development and deployment of AI systems. These actors would include the developers, deployers, owners, and users of content-generating AI

services. The European Union's AI Act provides an instructive example of such a law. Our recommendation is echoed in the India AI Governance Guidelines, which call for legislative clarity on the respective roles, obligations, and liabilities of actors (developers, deployers, users etc.) in the AI value chain. Such a law should not only mandate the addition of labels, metadata, and identifiers, but should also require risk assessments, technical documentation of the AI models, including their training and testing, independent audits, and transparency obligations to ensure that the outputs of AI models are not harmful and discriminatory. If such a law is enacted, it would complement existing intermediary liability provisions. Intermediaries that both host user-generated content and also offer AI services (for example, chatbots such as Grok by X) will have legal obligations under both the intermediary liability rules as well as the proposed AI legislation.

2. Rule 4(1A)

Recommendation:

This Rule mandates significant social media intermediaries (SSMIs) to require user declarations and deploy automated tools to verify synthetically generated information.

We suggest the inclusion of the following three measures as part of the due diligence requirements for SSMIs in addition to adopting technological measures to detect synthetically generated content.

First, this provision should require platforms to deploy technical measures such as graduated "circuit breaker" protocols to curb the spread of harmful synthetic content that is in contravention of the IT Rules. Further, SSMIs should be required to impose stricter rate limits, demonetization, and link-out suppression with respect to repeat offender accounts sharing harmful synthetic content, with delivery reinstated only after verified remediation.

Second, SSMIs should be required to conduct periodic risk assessments of their platform recommendation algorithms, particularly their role in amplifying harmful content, including content that is synthetically generated. The results of such an assessment and the corrective action plan should be transparently disclosed. Furthermore, SSMIs should be required to publish standardized risk metrics, such as time-to-label and take down, recommendation rates for detected synthetic hate, and recidivism, disaggregated by language and sensitive nature of content, such as gender-based violence.

Third, it is also important to ensure that, in addition to technical measures, SSMIs invest in trained human moderators who can assess the veracity and harmful or unlawful nature of content.

Rationale:

The proposed measure of requiring a user declaration is not likely to be effective as users who seek to manipulate communication through synthetically generated content are not likely to

disclose the same. Research has shown that labeling has limited impact on the persuasiveness / perceived accuracy of content. At the same time, when systems for labeling misinformation are implemented, any misinformation that remains unlabeled tends to be perceived as more accurate. This underscores that while the detection of the synthetic nature of content is important, it is more important to determine the veracity and harmful or unlawful nature of content and to take appropriate action to thwart its circulation.

Further, technological measures to detect synthetic or AI-generated content continue to face gross accuracy challenges, especially when deployed on a large scale and with bad-faith actors evolving measures to continually evade deepfake detection. Therefore, making safe harbor protection of SSMIs subject to the adoption of imperfect technical measures to detect synthetically generated content may devolve into an excuse for such platforms to shirk responsibility for the circulation of harmful synthetic content. Studies have repeatedly shown the role of social media algorithms in amplifying harmful content. Therefore, the law should also mandate SSMIs to be transparent about the role played by their recommendation algorithms in amplifying harmful conduct and mandate the adoption of technical measures to arrest the virality of such content, whether synthetically generated or otherwise. SSMIs should be required to audit amplification, downranking efficacy, repeat-offender handling, and cross-language performance, with public summaries and corrective action plans. Furthermore, SSMIs should be required to publish standardized risk metrics, such as time-to-label and take down, recommendation rates for detected synthetic hate, and recidivism, disaggregated by language and state.

Finally, while technological measures may help in detecting the synthetic nature of content, what is important is the determination of the harmful nature of the content, which requires knowledge of language, local context, and cultural nuances. This calls for highly qualified, trained human content moderators; technological measures, however sophisticated, cannot be a substitute. Human review is also important to check against the misuse of law and platform reporting mechanisms, especially against dissenting voices from marginalized or historically vulnerable groups.