The Dialogue™
INFORM ENGAGE IDEATE

WHITE PAPER

PREVENTION, DETECTION, REPORTING, AND COMPLIANCE

A COMPREHENSIVE APPROACH TOWARDS TACKLING DEEPFAKES IN INDIA

White Paper

PREVENTION, DETECTION, REPORTING, AND COMPLIANCE

A COMPREHENSIVE APPROACH TOWARDS TACKLING DEEPFAKES IN INDIA

Authors - Shruti Shreya & Pranav Bhaskar Tiwari

Copyeditor - Akriti Jayant

Designer - Shivam Kulshrestha

The Dialogue™ is a public policy think tank with a vision to drive a progressive narrative in India's policy discourse. Founded in 2017, we believe in facilitating well-researched policy debates at various levels to help develop a more informed citizenry, on areas around technology and development issues. The Dialogue™ has been ranked as the world's Top 10 think tanks to watch out for, by the Think Tank and Civil Societies Programme (TTCSP), University of Pennsylvania in their 2020 and 2021 rankings.

For more information Visit
thedialogue.co

Suggested Citation
Shreya. S, and Tiwari, PB. (2024) Prevention, Detection, Reporting, and Compliance: A Comprehensive Approach towards Tackling Deepfakes in India. The Dialogue™

Catalogue No
TD/PR/WP/0224/01

Publication Date
February 19, 2024

Cover Design
Cover prominently features John McCarthy, founding father of AI, in recognition of his groundbreaking contributions.

Disclaimer
The facts and information in this report may be reproduced only after giving due attribution to the authors and The Dialogue™.

Contents

Figures

Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
C2PA	Coalition for Content Provenance and Authenticity
CSAM	Child Sexual Abuse Material
GAC	Grievance Appellate Committee
IPC, 1860	Indian Penal Code, 1860
IT Act, 2000	Information Technology Act, 2000
LAION	Large-scale Artificial Intelligence Open Network
LEAs	Law Enforcement Agencies
LM	Language Modelling
NATO	North Atlantic Treaty Organisation
NPC	Non-Playable Character
VQGANs	Vector Quantised Generative Adversarial Networks

4K 29.97p, Stereo
90 °
Settings
28%
View
NORMAL
Text
Basic
Font
Size
Alignment
Vertical Alignment
Line Spacing
Tracking
Baseline
183,0
-69,0
2,01 %
-4,0
Index
00:00:00:00
00:00:05:00
00:00:10:00
00:00:08:22
10:00 / 03:15:06

I. Introduction

A. Unpacking the Technology

In today's digital age, synthetic media has become a testament to the astounding capabilities of modern technology. These digitally crafted creations, made possible through advancements in artificial intelligence (AI) and machine learning, have opened up a world of creative possibilities. Among these advancements, however, emerges a distinct and concerning subset known as 'deepfakes.' Deepfakes can be defined as hyper-realistic audio or video manipulations made using sophisticated AI techniques with a malicious intent to defraud, deceive or manipulate someone. These manipulations are so convincingly executed that they often become indistinguishable from genuine content, making them potentially deceptive and harmful. Therefore, while synthetic media as a whole contributes positively to various sectors, deepfakes necessitate a specific and rigorous regulatory approach.

With that being said, it is also crucial to ensure that regulations targeting deepfakes do not inadvertently hinder the positive growth and innovation associated with other forms of synthetic media. This careful delineation underscores the need for policies that are finely tuned to address the unique challenges posed by deepfakes, while simultaneously fostering a supportive environment for the continued development of beneficial synthetic media.

The objective of this paper is threefold. Firstly, it provides an understanding of the policy landscape surrounding deepfakes in India, offering insights into the regulatory mechanisms and initiatives undertaken so far. Secondly, it analyses the existing policy measures and technical efforts being undertaken in addressing this complex issue. Lastly, it charts a path forward focusing on enhancing the present initiatives and processes. The following box encapsulates how AI-generated synthetic media can be utilised positively in education, cinema, and other sectors.

Figure 1: Use Cases of Synthetic Media

- **Enhancing Content Creation for Global Reach**

  Consider a small-scale filmmaker aiming to release their movie globally. With synthetic media, they can effortlessly dub the film in multiple languages without requiring separate actors or compromising the original emotion and tone of the voices. This not only reduces costs but also broadens the content's reach, transcending language and creative barriers.[1]

- **Combating Online Child Exploitation**

  Law enforcement agencies are using synthetic media to create virtual decoy profiles online.[2] These profiles can appear as real children but are, in fact, synthetic creations. They can be used in sting operations to attract and catch predators online, preventing harm to real children and aiding in the fight against online child exploitation. This said, the ethics of using the technology is this way remains open.

- **Promoting Small and Medium Enterprises**

  In the Cadbury campaign, small business owners in different regions of India could feature Shahrukh Khan in

[1] Davenport, T. H., & Mittal, N. (2023, August 15). *How Generative AI Is Changing Creative Work*. Harvard Business Review. https://hbr.org/2022/11/how-generative-ai-is-changing-creative-work

[2] BBC. (2013, July 11). *"Virtual Lolita" aims to trap chatroom paedophiles*. BBC News. https://www.bbc.com/news/technology-23268893

personalised ads for their stores.[3] The synthetic media was generated with due permission from Mr Khan. A kirana shop in Mumbai, for instance, could have its own version of the ad where Shahrukh Khan talks about the store's special Diwali offers. This hyper-personalised marketing approach using synthetic media demonstrates how local businesses can leverage the star power of celebrities in a cost-effective manner, previously unthinkable due to budget constraints.

- **Simplifying Content Creation for End Users**

  AI in video generation is revolutionising content creation for consumers, making it easier and more accessible. For instance, Google's recently launched Lumiere is a text-to-video model that enables users to generate videos from textual descriptions.[4] This technology is particularly user-friendly, allowing even those without advanced video editing skills to create high-quality, coherent motion videos. Lumiere's ability to process videos in multiple space-time scales means it can handle a variety of video editing tasks, from converting images to videos to video inpainting and stylised generation. This marks a significant step forward in empowering consumers to create complex video content with ease.

- **Transforming Gaming through GenAI's Impact on NPC and Asset Creation**

  In the gaming industry, Generative AI is revolutionising NPC (non-playable character) creation and asset development. Custom language models, attuned to specific game lore, enable NPCs to interact in a human-like, context-aware manner, evolving with the player's journey. Tools like NVIDIA Riva and Audio2Face[5] aid in generating realistic voices and facial animations, enhancing real-time character expressiveness. Additionally, diffusion models streamline asset creation, a traditionally labour-intensive process, allowing for more efficient and creative development phases. This innovative use of Generative AI is just the beginning, with vast potential yet to be fully explored in gaming.

- **Revolutionising Education through Interactive and Personalised Experiences**

  In education, teachers can use synthetic media to create virtual historical figures for interactive learning.[6] Students could engage in a 'live' debate with a synthetic avatar of Abraham Lincoln, deepening their understanding of history through immersive experiences. Similarly, a lesson on environmental science could be taught in Marathi, Tamil, or any other local language, ensuring that students who are not proficient in English or Hindi can still access quality education.

- **Facilitating Hyperlocal Awareness of Welfare Schemes**

  Synthetic media can play a pivotal role in disseminating policies and schemes to a diverse, multilingual populace.[7] A health campaign on Pulse Polio could use synthetic media to create the message of 'Do Boond Zindagi Ke' and deliver it in the local dialect, significantly enhancing the campaign's impact and reach.

While acknowledging the positive applications of AI-generated synthetic media, the following box delves into the complex safety concerns and ethical dilemmas posed by creation of deepfakes, emphasising the need for awareness and regulation.

[3] Exchange4media. (2021, October 26). *Shah Rukh Khan-Cadbury campaign: How to create free ads to support your local stores*. exchange4media. https://www.exchange4media.com/digital-news/shah-rukh-khan-cadbury-campaign-how-to-create-free-ads-to-support-your-local-stores-116514.html

[4] Google Research Lumiere. *A Space-Time Diffusion Model for Video Generation.* Lumiere. https://lumiere-video.github.io/

[5] Omniverse Audio2Face App. NVIDIA. https://www.nvidia.com/en-in/omniverse/apps/audio2face/

[6] U.S. Department of Education, Office of Educational Technology, *Artificial Intelligence and Future of Teaching and Learning: Insights and Recommendations*, Washington, DC, 2023. https://www2.ed.gov/documents/ai-report/ai-report.pdf

[7] Misuraca, G., & Noordt, C. van. (2020, July 1). *AI Watch Artificial Intelligence in public services - europa.eu*. Science for Policy Report. https://publications.jrc.ec.europa.eu/repository/bitstream/JRC120399/jrc120399_misuraca-aiwatch_publicservices_30062020_def.pdf

Figure 2: Deepfake Threat Landscape

- **Exacerbating Geopolitical Tensions**

  The era of 'Intelligentised Warfare'[8] is epitomised by deepfakes, such as of Ukrainian President Zelensky[9] and Russian President Putin[10], falsely portraying surrender appeals. This advancement marks a significant shift in warfare, targeting the cognitive faculties of adversaries and reshaping the battleground of perception and belief.

- **Challenges to Individual Reputations and Privacy**

  Deepfakes of public figures like Rashmika Mandana[11] and Sachin Tendulkar[12] underscore the threat to privacy and personality rights in addition to violations under the Indian Penal Code (IPC) & IT Act. These concerns are not limited to public figures; platforms are tasked with safeguarding the rights of both public figures and private individuals, albeit with varying thresholds of protection.[13] This misuse of personal identity can be particularly damaging when used for blackmail or character assassination.

- **Synthetic Child Sexual Abuse Material**

  The disturbing emergence of AI-generated Child Sexual Abuse Material (CSAM) in databases like LAION-5B[14] signals a harrowing misuse of deepfake technology, raising serious ethical and legal concerns.

- **Election Integrity**

  The spectre of deepfakes looms large over electoral integrity, as seen in the alleged Chinese meddling in Taiwanese elections.[15] While deepfakes pose threats to election integrity, the simpler and less expensive ‘cheapfakes’ present risks too.[16] Cheapfakes, created using basic methods like altering the speed of a video or selective cutting, can be more easily produced.

- **Fanning the Flames of Discord**

  Deepfakes have escalated the potential for disinformation, aggravating tensions across religious and ethnic lines.[17] Their heightened realism and believability intensify the challenges in discerning truth from fabrication, fueling conflict and misunderstanding.

[8] *Artificial intelligence in warfare*. (2022, November 17). Manohar Parrikar Institute for Defence Studies and Analyses. https://www.idsa.in/idsanews/Artificial_Intelligence_in_Welfare

[9] Allyn, B. (2022, March 17). *Deepfake video of Zelenskyy could be “tip of the iceberg” in info war, experts warn*. NPR. https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia

[10] The Guardian. (2023, December 14). *AI-generated Putin asks Putin about his rumoured body doubles – video*. The Guardian. https://www.theguardian.com/world/video/2023/dec/14/ai-generated-vladimir-putin-rumoured-body-doubles-video

[11] HT Tech Team. (2023, November 7). *Rashmika Mandanna deepfake row: What happened and how to identify such videos*. https://tech.hindustantimes.com/tech/news/rashmika-mandanna-deepfake-row-what-happened-and-how-to-identify-such-videos71699337039075.html

[12] Shreya, S. (2024, January 16). *Sachin Deepfake: Rethink India's approach to combating this menace*. Moneycontrol. https://www.moneycontrol.com/news/opinion/sachin-deepfake-rethink-indias-approach-to-combating-this-menace-12060431.html

[13] Yanisky-Ravid, S., & Lahav, B. Z. (2016). Public interest vs. Private lives-affording public figures privacy in the digital era: The three principle filtering model. *U. Pa. J. Const. L.*, *19*, 975. https://scholarship.law.upenn.edu/cgi/viewcontent.cgi?article=1633&context=jcl

[14] Thiel, D. (2023, December 20). *Investigation finds AI image generation models trained on child abuse*. FSI. https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse

[15] WION Web Team. (2024, January 11). *Taiwan election: Pro-China disinformation targets voters with deepfakes and manipulative videos*. WION. https://www.wionews.com/world/taiwan-election-pro-china-disinformation-targets-voters-with-deepfakes-and-manipulative-videos678779

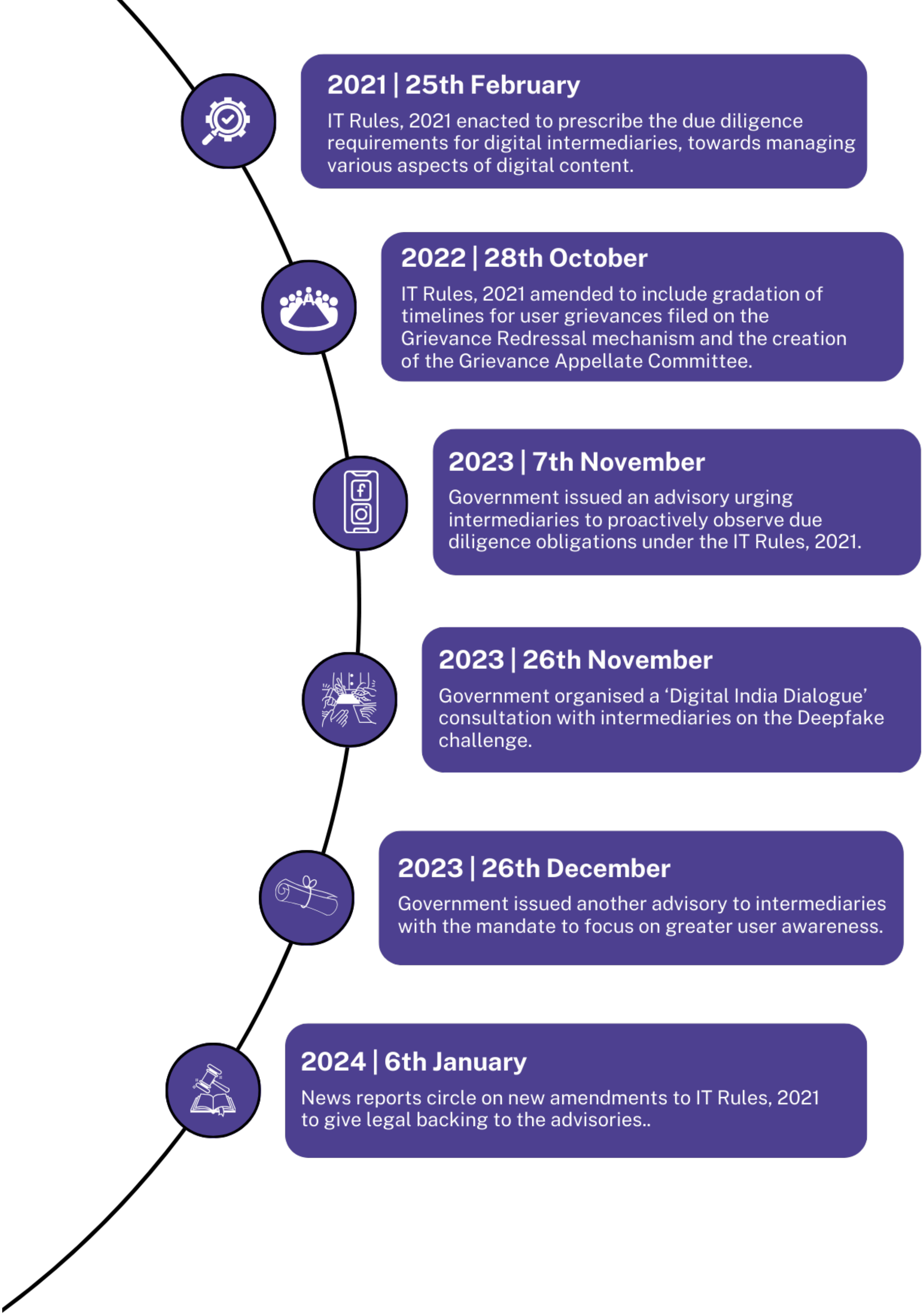
[16] Schick, N. (2020, December 24). *Don’t underestimate the cheapfake*. MIT Technology Review. https://www.technologyreview.com/2020/12/22/1015442/cheapfakes-more-political-damage-2020-election-than-deepfakes/. *See also,* Elliott, V. (2023, December 18). *Worried about deepfakes? don’t forget “cheapfakes.”* Wired. https://www.wired.com/story/meta-youtube-ai-political-ads/

[17] Correspondent, D. (2023, November 20). *Deepfake video of BSP's Praveen Kumar stirs religious row*. Deccan Chronicle. https://www.deccanchronicle.com/nation/politics/201123/deepfake-video-of-bsps-praveen-kumar-stirs-religious-row.html

B. Recent Developments

In India, the Information Technology Act, 2000 (IT Act) serves as the primary legislative framework regulating digital content, with detailed guidelines under the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 (IT Rules, 2021) addressing online safety challenges, including misinformation and privacy breaches. This table delves into the developments in India concerning deepfakes:

Figure 3: Deepfake Related Policy Developments in India



2021 | 25th February

IT Rules, 2021 enacted to prescribe the due diligence requirements for digital intermediaries, towards managing various aspects of digital content.

2022 | 28th October

IT Rules, 2021 amended to include gradation of timelines for user grievances filed on the Grievance Redressal mechanism and the creation of the Grievance Appellate Committee.

2023 | 7th November

Government issued an advisory urging intermediaries to proactively observe due diligence obligations under the IT Rules, 2021.

2023 | 26th November

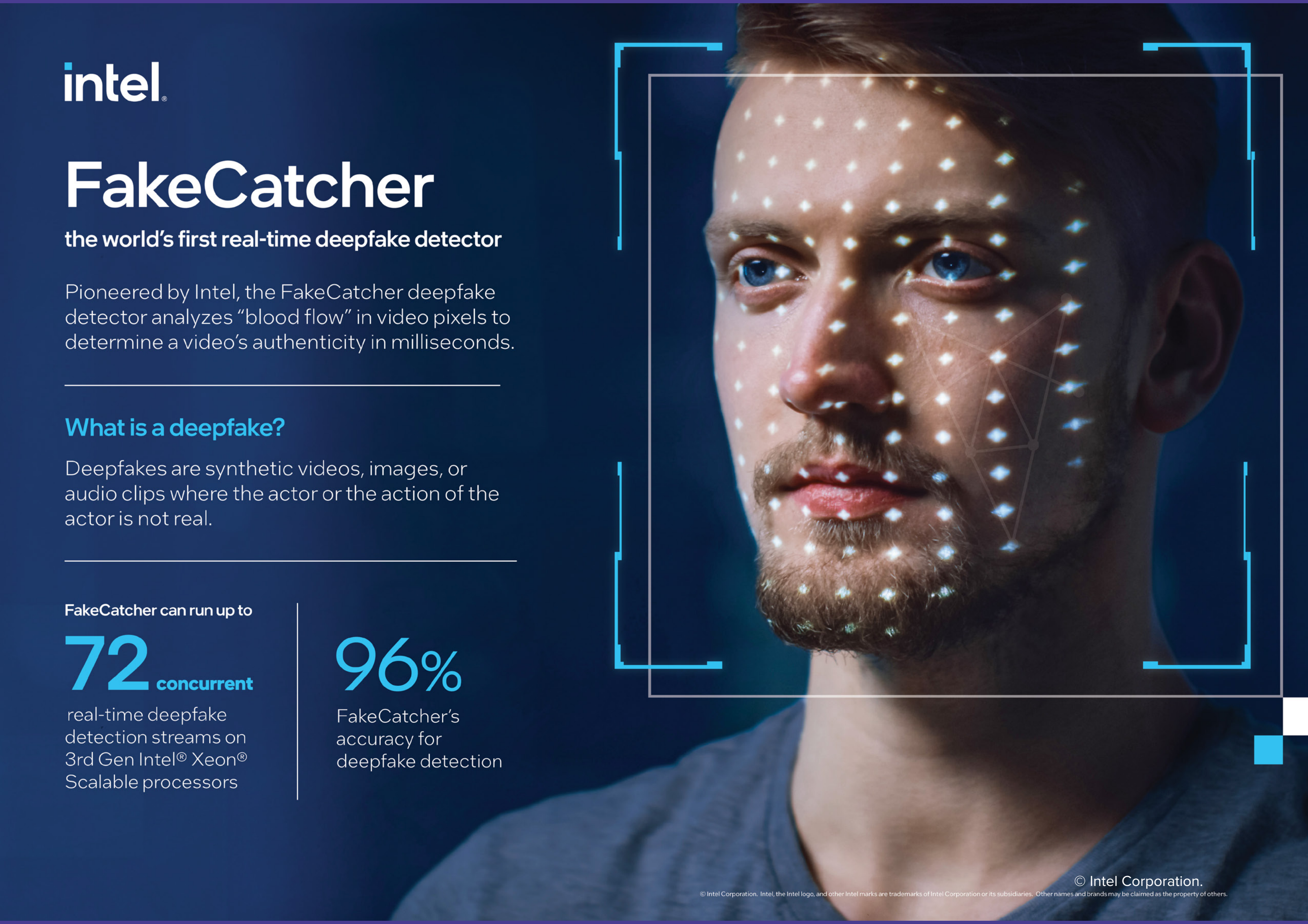
Government organised a ‘Digital India Dialogue’ consultation with intermediaries on the Deepfake challenge.

2023 | 26th December

Government issued another advisory to intermediaries with the mandate to focus on greater user awareness.

2024 | 6th January

News reports circle on new amendments to IT Rules, 2021 to give legal backing to the advisories..



intel

FakeCatcher

the world's first real-time deepfake detector

Pioneered by Intel, the FakeCatcher deepfake detector analyzes "blood flow" in video pixels to determine a video's authenticity in milliseconds.

What is a deepfake?

Deepfakes are synthetic videos, images, or audio clips where the actor or the action of the actor is not real.

FakeCatcher can run up to

72 concurrent real-time deepfake detection streams on 3rd Gen Intel® Xeon® Scalable processors

96% FakeCatcher's accuracy for deepfake detection

II. Current Industry Efforts

A. Technical Efforts

The battle against deepfakes started soon after the deepfake technology itself. These developments trace back to the early days of deepfake emergence, around 2017, when this menace started gaining public attention.[18] There onwards, industry efforts, crucial in the broader fight against digital manipulation, evolved significantly over the years. For instance, by 2018-2019, there was a notable increase in the development of advanced tools and strategies for deepfake detection.[19] Since then, the momentum has continued, with the companies constantly innovating and refining approaches to counter AI-related risks.

In the escalating battle against deepfakes, technology companies and academic institutions globally are uniting with various stakeholders to deploy technologies innovatively. These efforts are compartmentalised into distinct yet collaborative initiatives:

Coalition for Content Provenance and Authenticity (C2PA), a joint endeavour by technology giants like Adobe and Microsoft, spearheads the development of standards for media content certification.[20] Their initiatives, like the Content Authenticity Initiative[21] and Project Origin[22] emphasise cryptographic hashing and digital watermarking for content integrity.

Figure 4: Efforts by the Industry to establish Content Provenance

These initiatives certify the source of digital content.

Coalition for Content Provenance and Authenticity (C2PA)

- **Members:** Adobe, Arm, Intel, Microsoft, and Truepic etc.
- **Mission:** Development of technical standards for certifying the source and history of media content; combating misleading information online
- **Initiatives:** Oversees Content Authenticity Initiative and Project Origin

Content Authenticity Initiative

- **Members:** Industry leaders and technology experts in digital content like Adobe, Airbus, American Association of Insurance Services, BBC, Black Women In Artificial Intelligence, and The Wall Street Journal
- **Technology:**
    - ➢ Utilises cryptographic asset hashing for verifiable, tamper-evident signatures
    - ➢ Ensures that image and metadata alterations are traceable

[18] Westerlund, M. (1970, January 1). *The emergence of Deepfake Technology: A Review*. Technology Innovation Management Review. https://timreview.ca/article/1282

[19] Knight, W. (2020, April 2). *A new deepfake detection tool should keep world leaders safe-for now*. MIT Technology Review. https://www.technologyreview.com/2019/06/21/134815/a-new-deepfake-detection-tool-should-keep-world-leaders-safefor-now/

[20] Coalition for Content Provenance and Authenticity. (2023) . *Overview*. C2PA. https://c2pa.org

[21] Content Authenticity Initiative. (2023). *Authentic storytelling through digital content provenance*. Content Authenticity. https://contentauthenticity.org

[22] Project Origin. (2023). *What Origin Does*. Origin Project. https://www.originproject.info

- **Unique Characteristics:**
  - Integrates with tools like Photoshop for secure metadata preservation during editing
  - Enables consumers to view historical information about content through the Verify site

Project Origin

- **Members:** A coalition led by Microsoft and BBC, including other technology and media companies
- **Technology:**
  - Utilises digital watermarking credentials for content integrity
  - Implements measures to trace media content back to publishers
- **Unique Characteristics:**
  - Aimed at maintaining integrity of news and information content
  - Integrates with tools like Photoshop for secure metadata preservation during editing

Efforts to combat deepfakes are advancing with the application of watermarking technology, as seen in Meta's Stable Signature,[23] Google DeepMind's SynthID,[24] and Amazon's Titan Image Generator.[25] These initiatives focus on embedding invisible yet resistant watermarks into AI-generated content. This watermarking method integrates data related to the content's origin, acting as an imperceptible signature that remains intact even if the content is copied or distributed without authorisation. This approach is essential in maintaining the integrity and ownership of digital content, providing a robust solution against the manipulation and misuse of AI-generated media.

Figure 5: Efforts to tackle Deepfake with Watermarking

These solutions embed invisible and irremovable watermarks in digital content leading to detection of AI-generated content.

Stable Signature

- **Creator:** Launched by Meta in October 2023
- **Technology:** Embeds invisible watermarks into images created by open-source generative AI models, traceable even after editing
- **Unique Characteristics:**
  - ➢ Watermark is rooted in the model's digital data, resistant to removal
  - ➢ Capable of reducing false positives in AI-generated image detection
  - ➢ Compatible with popular AI models like VQGANs and latent diffusion models

SynthID

- **Creator:** Beta version launched by Google DeepMind, in partnership with Google Cloud
- **Technology:** Embeds imperceptible digital watermarks into AI-generated images or audio; includes detection capability for identifying content created by Google's AI models
- **Unique Characteristics:**
  - ➢ Extends to watermarking AI-generated music and audio
  - ➢ Detectable watermark remains intact even after image modifications
  - ➢ Offers three confidence levels for result interpretation in identification

Titan Image Generator

- **Creator:** Preview Announced by Amazon in November, 2023
- **Technology:** Empowers rapid production of high-quality images with built-in invisible watermarks integrated into image outputs to identify AI-generated images
- **Unique Characteristics:**

[23] Douze , M., & Fernandez, P. (2023, October 6). *Stable signature: A new method for watermarking images created by open source generative AI.* AI at Meta. https://ai.meta.com/blog/stable-signature-watermarking-generative-ai/

[24] SynthID. *Identifying AI-generated content with SynthID*. (2023, November 16). Google DeepMind. https://deepmind.google/technologies/synthid/

[25] Wiggers, K. (2023, November 29). *Amazon finally releases its own AI-Powered Image Generator at AWS re:invent 2023*. TechCrunch. https://techcrunch.com/2023/11/29/amazon-finally-releases-its-own-ai-powered-image-generator/

- ➢ Focus on responsible AI at every stage of model development
- ➢ Watermarks designed to be resistant to alterations
- ➢ Part of Amazon Titan models, offering a variety of image, multimodal, and text model options

The classifier-based deepfake detection efforts, incorporating tools such as Sentinel,[26] FakeCatcher,[27] Sensity,[28] Quantum Integrity,[29] Google's Audio LM Classifier,[30] and Microsoft Video Authenticator,[31] utilise advanced AI technologies for multi-layered analysis and real-time detection. These systems employ classifier-based detection methods, where machine learning algorithms are specifically trained to distinguish between authentic and manipulated media. They analyse key features in images or videos, like facial expressions and lighting inconsistencies, to identify signs of digital alteration. The classifiers then assess the probability of the content being a deepfake, providing a likelihood score. Continuously updated with new data, these tools adapt to evolving deepfake techniques, making them efficient and effective in the identification and mitigation of manipulated media in various applications.

Figure 6: Classifier-based Deepfake detection

These solutions utilise AI innovatively to detect deepfakes.

Sentinel

- **Creator:** Founded by ex-NATO and U.K. Royal Navy members with a tech background
- **Technology:** Utilises a multi-layered AI approach, combining hashing, metadata, audio analysis, and facial scrutiny for deepfake detection
- **Unique Characteristics:** Houses the largest verified deepfake collection, offering resistance against reverse-engineering for robust authenticity checks

FakeCatcher

- **Creator:** Developed by Intel and Umur Ciftci from SUNY Binghamton
- **Technology:** Analyses blood flow signals in video pixels, converting them into spatiotemporal maps to detect deepfakes
- **Unique Characteristics:** Achieves 96% accuracy with real-time detection, running 72 streams simultaneously and optimised with tools like OpenVino and OpenCV

[26] *Defending against deepfakes and information warfare*. Sentinel. (2020). https://thesentinel.ai

[27] Intel. (2022, November 14). *Intel introduces real-time Deepfake Detector*. https://www.intel.com/content/www/us/en/newsroom/news/intel-introduces-real-time-deepfake-detector.html

[28] Sensity. (2024, January 17). *Strengthen your detection capabilities against AI-Manipulated content*. Sensity. https://sensity.ai/

[29] *A.I. Powered Deep Fake & Image Forgery Detection*. Q- Integrity. https://q-integrity.com/

[30] Borosis Z., Zeghidour N. (2022, October 26). *AudioLM: a Language Modeling Approach to Audio Generation*. Google Research. https://blog.research.google/2022/10/audiolm-language-modeling-approachto.html#:~:text=For%20this%20purpose%2C%20we%20trained,with%20a%20simple%20audio%20classifier

[31] Horvitz, E., & Burt, T. (2021, May 5). *New steps to combat disinformation. Microsoft On the Issues.* Microsoft. https://blogs.microsoft.com/on-he-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/

Microsoft Video Authenticator

- **Creator:** Developed by Microsoft Research, in collaboration with Microsoft's Responsible AI team
- **Technology:** Employs AI algorithms to analyse photos and videos for subtle alterations and blending boundaries
- **Unique Characteristics:** Provides a real-time confidence score for media manipulation; part of Microsoft's Defending Democracy Program, emphasising ethical journalism and disinformation combat

Sensity: Visual Threat Intelligence Platform

- **Creator:** A Dutch startup focusing on digital security and deepfake detection
- **Technology**: Offers a platform and API for detecting and countering deepfakes by analysing various sources
- **Unique Characteristics:** Collects extensive visual threat intelligence, leveraging deep learning for comprehensive risk assessment in audio-visual content

Quantum Integrity

- **Creator:** A Swiss startup specialising in AI-based digital security solutions
- **Technology**: Uses deep learning algorithms for detailed analysis and detection of deepfake manipulations in images and videos
- **Unique Characteristics:** Provides adaptable algorithms for diverse use cases, enhancing fraud prevention and decision-making in different sectors

Collaborative and mixed-method initiatives spearheaded by academics, scientists, community leaders, and industry experts are making significant strides in the realm of deepfake detection. Innovative approaches like WeVerify,[32] Phoneme-Viseme Mismatch Detection,[33] DeepWare AI,[34] Human Vocal Tract Modeling System,[35] and PhaseForensics[36] are at the forefront of this effort. These techniques employ cross-modal content verification and intricate analysis of human speech and facial movements, playing a pivotal role in safeguarding the digital environment from deceptive manipulations. These concerted efforts, a blend of academic research, scientific innovation, community involvement, and industry application, are key in combating digital manipulation. They highlight the critical importance of technological advancements in preserving the integrity of digital content, demonstrating a unified response from various sectors in addressing this challenge.

[32] WeVerify. (2021, May 21). *Wider and enhanced verification for you*. WeVerify. https://weverify.eu/

[33] Agarwal, S., Farid, H., Fried, O., & Agrawala, M. (2020). *Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches*. Open Access. https://openaccess.thecvf.com/content_CVPRW_2020/papers/w39/Agarwal_Detecting_Deep-Fake_Videos_From_Phoneme-Viseme_Mismatches_CVPRW_2020_paper.pdf

[34] Deepware.ai. (2021). *Scan & Detect Deepfake Videos*. Deepware.ai. https://deepware.ai

[35] Blue, L., Warren, K., Abdullah, H., Gibson, C., Vargas, L., O'Dell, J., Butler, K., & Traynor, P. (1970, January 1). *Who are you (I really wanna know)? detecting audio {deepfakes} through vocal tract reconstruction*. USENIX. https://www.usenix.org/conference/usenixsecurity22/presentation/blue

[36] Prashnani, E., Goebel, M., & Manjunath, B. S. (2022, November 17). *Generalizable deepfake detection with phase-based motion analysis*. arXiv.org. https://arxiv.org/abs/2211.09363

Figure 7: Collaborative and Mixed Approaches to tackling Deepfake

These solutions utilise a combination of technical capabilities, existing databases, social network analysis, lip motion, and other mechanisms to detect deepfakes.

WeVerify: Wider and Enhanced Verification For You

- **Creator:** The WeVerify consortium, led by Ontotext AD, includes the University of Sheffield, Deutsche Welle, and Agence France-Presse, among others
- **Technology:** Employs cross-modal content verification, social network analysis, and blockchain-based database for fact-checking
- **Unique Characteristics:** Features an open-source browser plugin, a collaborative cross-media verification workbench, and a citizen-oriented verification chatbot

Phoneme-Viseme Mismatch Detection

- **Creator:** Developed by experts at Stanford University and the University of California
- **Technology:** Analyses mouth movements (visemes) versus spoken sounds (phonemes) using AI algorithms
- **Unique Characteristics:** Excels in identifying subtle mismatches indicative of deepfake videos with minor or localised manipulations

PhaseForensics

- **Creator:** Developed by researchers at the University of California, Santa Barbara, including Ph.D. student Michael Goebel
- **Technology**: Focuses on lip motion analysis across various frequencies for neural network analysis
- **Unique Characteristics:** Exhibits high generalizability and accuracy across different datasets, employing a balanced detection methodology

DeepWare AI

- **Creator:** A community-driven, open-source initiative focused on Deepfake detection
- **Technology:** Utilises a comprehensive database of over 124,000 videos, including live content, for accurate synthetic media identification. Trained on diverse sources like YouTube, 4Chan, and Celeb-DF videos
- **Unique Characteristics:** Adapts to the evolving online environment, maintaining relevance with new online trends. Its vast and varied video collection ensures robustness in detecting a wide range of deepfakes

Human Vocal Tract Modeling System

- **Creator:** Developed by scientists at the University of Florida
- **Technology:** Models the human vocal tract to distinguish between real and synthetic audio. Trained on both authentic and fake audio recordings to create realistic values for different parts of the vocal tract
- **Unique Characteristics:** Highly accurate, with an ability to determine the biological plausibility of audio samples. Not reliant on prior exposure to specific deepfake audio, making it versatile in defending against various deepfake techniques

B. Policy Efforts

In addition to technical advancements, innovative policy measures and sensitisation efforts are equally critical to creating a holistic response to digital challenges. To that end, this section delves into the policy and awareness-building efforts being undertaken by the industry.

1. Collaborative Partnerships and Global Initiatives

A key strategy in combating deepfakes involves forming collaborative partnerships and participating in global initiatives, drawing on a diverse range of expertise from academia, industry stakeholders, government bodies, and experts around the world. A notable example is the Frontier Model Forum, which includes leading tech companies such as Microsoft, Google, OpenAI, and Antropic.[37] This forum leverages the technical and operational expertise of its member companies to advance AI safety research and support the development of AI applications aimed at addressing society's most pressing needs. This collective effort enriches the AI ecosystem, providing a platform for shared learning and innovation.

In a similar vein, the Partnership on AI supported by leading tech companies globally stands out as a non-profit collaboration among academic institutions, civil society organisations, industry leaders, and media entities.[38] This partnership focuses on creating solutions that ensure AI advances yield positive outcomes for people and society. By pooling resources and knowledge from diverse sectors, these partnerships and initiatives are pivotal in shaping an AI landscape that is safe, ethical, and beneficial for all.

Another notable example of innovative collaboration is Fox's partnership with Polygon Labs.[39] This partnership aims to tackle deepfake distrust through the development of blockchain technology. They have developed an open-source protocol called Verify, allowing media companies to register their content and establish its authenticity. This system cryptographically signs individual pieces of content on the blockchain, helping to establish the origin and history of original journalism. Such an approach not only helps in detecting deepfakes more effectively but also promotes a culture of trust and authenticity in digital content.

2. Content moderation through Automated Tools

To effectively combat the challenges posed by deepfakes, tech companies have been deploying automated content moderation systems that integrate advanced AI classifiers, machine learning algorithms, and human oversight. This approach has been a mainstay in the industry, with platforms long employing AI to detect various types of violative content. Now, their focus is increasingly on enhancing these systems to effectively detect deepfakes.

AI classifiers are crucial in this setup. They are programmed to swiftly scan and analyse large volumes of digital content, pinpointing potential deepfakes or other forms of manipulated media. Some companies like Google, for instance, also employ tools like the Audio LM classifier, part of their Audio Language Model, for more nuanced detection of AI-generated content.[40]

Machine learning algorithms complement these classifiers. They continuously evolve, improving their accuracy through the analysis of extensive datasets containing both genuine and deepfake content. This learning process enables them to identify subtle discrepancies and signs of manipulation in digital media. While AI provides scalability for filtering vast amounts of data, human moderators add a layer of contextual understanding, applying their knowledge to AI-flagged content. This blend of AI efficiency and human judgement ensures a more accurate and sensitive approach to content moderation.

However, while these automated tools are highly effective, they are not without challenges. The accuracy of AI classifiers can vary based on the quality and diversity of the training data. There’s also the risk of false positives or negatives.[41] Moreover, as deepfake technology advances, it becomes more challenging for these systems to detect manipulated content accurately. Accordingly, it is critical that these AI models are continuously refined and the content moderation strategies are regularly updated.

[37] Frontier Model Forum. (2024). *Frontier Model Forum: Advancing Safe AI Development.* Frontier Model Forum. https://www.frontiermodelforum.org

[38] Partnership on AI. (2024). *About Us: Advancing positive outcomes for people and society*. Partnership on AI. https://partnershiponai.org/about/

[39] Wiggers, K. (2024, January 9). *Fox partners with Polygon Labs to tackle deepfake distrust*. TechCrunch. https://techcrunch.com/2024/01/09/2648953/

[40] Google Deep Mind. (2023, 27 October). *AI Safety Summit: An update on our approach to safety and responsibility*. Google Deep Mind. https://deepmind.google/public-policy/ai-summit-policies/#identifiers-of-ai-generated-material

[41] Silberg, J., & Manyika, J. (2019, June 6). *Tackling bias in artificial intelligence (and in humans)*. McKinsey & Company. https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans

3. Fact-Checking Efforts

Fact-checking and contextual analysis in the context of deepfakes extends beyond straightforward verification of facts. It necessitates a deep understanding of both the content and its context. This involves analysing not just the explicit content but also elements like tone of voice, setting, and the historical and relational context between communicators. Tools like Brand24 exemplify this approach by analysing the context of discussions, assessing the most commonly associated words with a monitored keyword and their sentiments, whether positive, neutral, or negative.

However, this method isn't infallible and can be influenced by biases in algorithms or the perspectives of the fact-checkers themselves. Accordingly, adopting a more balanced and transparent approach to fact-checking is crucial.

In addition to nuanced fact-checking, there's also a growing emphasis on elevating authoritative sources of information. This approach involves prioritising content from recognised authorities, such as election commissions or established news outlets, especially in situations prone to misinformation. By highlighting authentic news and providing contextual information from these sources, platforms can help ensure that the most reliable information is disseminated to the public. However, it's crucial to manage this approach carefully to avoid inadvertently suppressing legitimate alternative voices and perspectives, thereby maintaining a diverse and healthy information ecosystem.

4. User Empowerment and Media Literacy

User empowerment and media literacy initiatives are crucial in the fight against deepfakes, and several notable campaigns and programs highlight this approach. For instance, the Global Network Initiative and YouTube's 'Hit Pause' campaign represent significant efforts in this direction. This campaign is designed to encourage users to think critically and verify information before sharing online content, particularly videos. It aims to raise awareness about the potential manipulation of videos and the importance of cross-checking facts, thus empowering users to become more discerning consumers and sharers of digital content.

Jigsaw, a unit within Google, also contributes significantly to this area. Jigsaw focuses on tackling global security challenges through technology, including misinformation and cyber threats. They develop tools and initiatives to help users understand and navigate the complex digital landscape, including the threats posed by deepfakes and other forms of digital manipulation.

Additionally, initiatives like Meta's partnership with Reuters, which offers a free online training course on deepfake identification, are instrumental in educating users, especially those in newsrooms. Similarly, X's Birdwatch initiative is another innovative approach, where users are empowered to collaboratively annotate tweets they believe contain misleading information. This feature encourages active participation from users in identifying and contextualising potentially deceptive content, fostering a community-based approach to combating misinformation.

5. Research and Academic Collaboration

Investment in academic research and collaboration with educational institutions is also a crucial element. Google's funding to the Indian Institute of Technology, Madras, for establishing a Responsible AI centre is noteworthy in this regard.[42] This initiative aims to foster collective efforts involving researchers, domain experts, developers, community members, policymakers, and more to ensure responsible AI deployment and localisation tailored to the Indian context. Similarly, Meta's support for the Deep Fake Detection Challenge is an important initiative to advance deepfake detection capabilities through academic partnerships.[43]

[42] Browning, M. (2023, November 29). *Our approach to protecting users from the risks of AI generated media*. Google. https://blog.google/intl/en-in/company-news/technology/our-approach-to-protecting-users-from-the-risks-of-ai-generated-media/

[43] Meta. (2020, June 25). *Deepfake Detection Challenge Dataset*. AI at Meta. https://ai.meta.com/datasets/dfdc/

The Dialogue™
INFORM ENGAGE IDEATE
#SaferInternetDay
डीपफेक का विश्लेषण: सुरक्षित डिजिटल भविष्य की दिशा में
श्रुति श्रेया और प्रणव भास्कर तिवारी के साथ
हिंदी एपिसोड इंग्लिश सबटाइटल्स के साथ
TECH
Spotify
51 unread emails
Reminder for Adhering to the IPC and IT Act.
No, thanks

III. Evaluating the Feasibility of New Prescriptions under IT Rules, 2021

As India grapples with the challenges posed by deepfakes, the government is considering potential amendments to the IT Rules, 2021 to strengthen its regulatory framework. This section delves into the feasibility of the proposed amendments that are likely to be introduced as per press reports,[44] aiming to analyse their need and feasibility given the extensive mechanisms that are already in place.

A. The requirement of a definition

Reports suggest that the new amendment shall include a definition of deepfakes under the IT Rules, 2021. This proposal to define deepfakes must distinguish between malicious use and constructive and creative applications of synthetic media to prevent stifling innovation and free speech.[45]

Furthermore, there is also a concern about the risk of regulatory incoherence that this inclusion might bring. The IT Rules, 2021, already contain overly broad provisions related to misinformation[46] and privacy breaches,[47] which can easily subsume deepfake-related harms. Introducing specific definitions for every emerging technological challenge can lead to an overly complex regulatory environment. This complexity risks confusing intermediaries due to the multitude of encumbrances for similar types of harms. Legal amendments should be considered a last resort, with a preference for utilising the existing framework to address concerns as long as possible to ensure a fair balance between regulation *vis a vis* innovation and free expression.

B. The efficacy of constant reminders of legal penalties

Another proposed amendment might require platforms to send constant reminders to users about the legal penalties outlined in the IPC and the IT Act for posting harmful content.[48] While the intention to educate users about the consequences of their actions is commendable, the feasibility and effectiveness of constant reminders warrant careful consideration. There's a risk of 'legal fatigue' where users, overwhelmed by the repetitive and intricate legal language, may choose to ignore these warnings altogether. This desensitisation could lead to a paradoxical situation where increased warnings result in decreased attention and awareness. Research in psychology has shown that overexposure to warnings can lead to desensitisation. A study titled Warning Fatigue: Insights from the Australian Bushfire Context[49] found that repeated exposure to warnings could lead to diminished attention. Further, legal scholars and behavioural scientists have studied the effectiveness of legal notices and warnings. For instance, the paper Simplifying Privacy Disclosures – An Experimental Test[50] suggests that complex legal language is often less effective and that simplified disclosures are more likely to be understood by users.

A more nuanced approach, perhaps involving simpler language and more engaging methods of communication, might yield better results in terms of user awareness and responsible use of the platform. Using clear and straightforward language, coupled with interactive and informative approaches, can help users better understand the implications of their actions and encourage responsible platform usage.

[44] Agarwal A. (2024, January 06). *Centre likely to amend IT rules to define deepfakes*. Hindustan Times. https://www.hindustantimes.com/india-news/centre-likely-to-amend-it-rules-to-define-deepfakes-101704482610756.html

[45] Askari, J. (2022, November 25). *Deepfakes and Synthetic Media: What are they and how are techUK members taking steps to tackle misinformation and fraud*. The UK's Technology Trade Association. https://www.techuk.org/resource/synthetic-media-what-are-they-and-how-are-techuk-members-taking-steps-to-tackle-misinformation-and-fraud.html

[46] Rule 3 (1) (vi) IT Rules, 2021.

[47] Rule 3 (1) (ii) IT Rules, 2021.

[48] Agarwal A. (2024, January 06). *Centre likely to amend IT rules to define deepfakes*. Hindustan Times. https://www.hindustantimes.com/india-news/centre-likely-to-amend-it-rules-to-define-deepfakes-101704482610756.html

[49]Mackie, B. (2013). *Warning fatigue: Insights from the Australian bushfire context: A thesis submitted in partial fulfilment of the requirements for the degree of doctor of philosophy in media and communication in the University of Canterbury* (thesis). https://ir.canterbury.ac.nz/server/api/core/bitstreams/85f17b9e-cf99-4d8c-87c5-a3d1472cd309/content

[50] Adam S. Chilton & Omri Ben-Shahar,(2016) *"Simplification of Privacy Disclosures: An Experimental Test"* (CoaseSandor Working Paper Series in Law and Economics No. 737). https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=2443&context=law_and_economics#:~:text=The%20results%20of%20the%20experiment,3

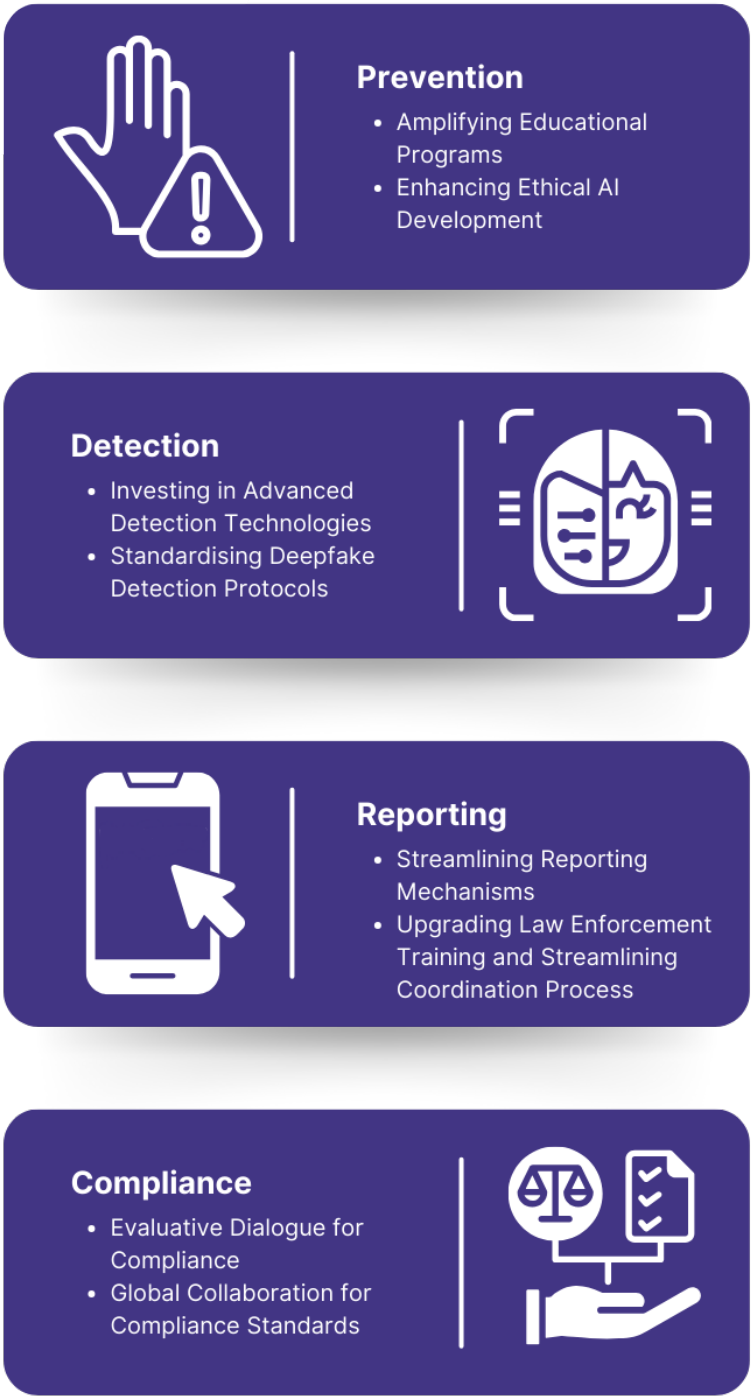
C. Enhancement of Grievance Redressal Mechanism

The third proposed amendment might make all user grievances appealable to the Grievance Appellate Committee (GAC).[51] Currently, only complaints made directly to the Grievance Officer are considered appellable, while grievances reported through in-app mechanisms are excluded from this process.

While this is aimed at enhancing accountability in the grievance redressal process, it will be important to ensure that the current dual-route system comprising in-app reporting and the option to approach a grievance officer continues. This is important to maintain flexibility and efficiency and allow users to choose the most appropriate path based on the nature of their grievance. Moreover, it is equally important to ensure that users are aware of the grievance redressal resources, including the appellate mechanisms. This requires proactive communication and education initiatives to inform users about their rights and the avenues available for addressing grievances. A well-informed user base is more likely to utilise the appeal mechanism effectively, contributing to the overall goal of enhancing accountability and improving the platform's governance.

[51] Agarwal A. (2024, January 06). *Centre likely to amend IT rules to define deepfakes*. Hindustan Times. https://www.hindustantimes.com/india-news/centre-likely-to-amend-it-rules-to-define-deepfakes-101704482610756.html

Figure 8: Recommendations for Leveraging and Improving Mechanisms



Prevention

- Amplifying Educational Programs
- Enhancing Ethical AI Development

Detection

- Investing in Advanced Detection Technologies
- Standardising Deepfake Detection Protocols

Reporting

- Streamlining Reporting Mechanisms
- Upgrading Law Enforcement Training and Streamlining Coordination Process

Compliance

- Evaluative Dialogue for Compliance
- Global Collaboration for Compliance Standards

IV. Recommendations

As we navigate the complexities of synthetic media and deepfakes, our policy measures must be in consonance with rapid technological advancements while aligning with the four foundational principles set forth by India's IT Minister: Prevention, Detection, Reporting, and Compliance.[52] Our existing regulatory and technological landscape is equipped with a multitude of processes and systems that form a solid base for tackling these challenges. The need of the hour is to focus on enhancing and refining these existing frameworks. This approach is preferable to the introduction of new due diligence requirements, which could potentially complicate or slow down the response to deep fakes.

This section presents recommendations that advocate for leveraging and improving the mechanisms already in place. Such a strategy strikes a delicate balance, fostering both accountability and innovation, and promotes a secure and responsible digital environment.

A. Prevention

1. Amplifying Educational Programs

Building upon existing digital literacy campaigns, there is a need to further amplify the reach and depth of educational programs that inform the public about deepfakes. Greater collaboration with educational institutions and leveraging media platforms can broaden the impact of these initiatives.

2. Enhancing Ethical AI Development

NITI Aayog[53] and the Ministry of Electronics and IT[54] are already focused on promoting responsible and trustworthy AI. The emphasis should now be on enhancing these efforts through deeper collaboration between government agencies and tech companies. Practices like Amazon Web Service's AI Service Cards[55] and greater investment by companies in new resources[56] for enhancing their response to AI based social challenges are commendable and should be adopted more widely. The aim is to improve responsible AI practices across the industry continuously.

B. Detection

1. Investing in Advanced Detection Technologies

Current efforts in developing AI for deepfake detection are commendable, yet there's a pressing need to boost investment in this domain significantly. Enhanced funding will sharpen the accuracy of these technologies. This initiative should encompass support for public-private partnerships and dedicated research into advanced detection technologies, underlining the fact that continuous innovation is crucial in the effective combat against deepfakes.

2. Benchmarking Deepfake Detection Protocols

On the foundation of existing efforts, standardising deepfake detection benchmarks is critical. This requires a collaborative push from tech companies, academic researchers, and government bodies to forge a unified set of detection benchmarks while balancing the need for independent strategies to distinguish between harmful deepfakes and legitimate synthetic media uses and evade bad actors.

C. Reporting

1. Streamlining Reporting Mechanisms

As discussed earlier, existing IT Rules and tech companies' internal systems already embed effective grievance redressal processes. The key focus now should be on intensifying user awareness of these mechanisms and improving their user-friendliness for optimal utilisation. The framework is in place;

[52] Vaishnaw A. (2023, November 23). *New regulations soon to tackle deepfake*. The Hindu BusinessLine. https://www.thehindubusinessline.com/info-tech/new-regulations-soon-to-tackle-deepfakes-vaishnaw/article67564946.ece/amp/

[53] NITI Aayog. (2022, November). *Working Document: Enforcement Mechanisms for responsible #AIforAll*. NITI Aayog. https://www.niti.gov.in/sites/default/files/2023-03/Responsible-AI-AIForAll-Approach-Document-for-India-Part-Principles-for-Responsible-AI.pdf

[54] Innovate India. 2024. *Call for expression of interest on responsible AI*. Innovate India. https://innovateindia.mygov.in/eoi-responsibleai/#:~:text=The%20Ministry%20of%20Electronics%20and,to%20its%20socio%2Deconomic%20realities.

[55] Philomin, V., & Hallinan, P. (2022, November 30). *Introducing AWS AI Service Cards: A new resource to enhance transparency and advance responsible AI*. Amazon. https://aws.amazon.com/blogs/machine-learning/introducing-aws-ai-service-cards-a-new-resource-to-enhance-transparency-and-advance-responsible-ai/

[56] Open AI. (2023). *Elections Program Manager*. (2023). Open AI. https://openai.com/careers/elections-program-manager

empowering users with the knowledge to access and utilise it efficiently is crucial.

2. Upgrading Law Enforcement Training and Streamlining Coordination Process

Law enforcement agencies already undergo training for handling cybersecurity cases, but this training must be consistently updated to incorporate the latest technological and legal advancements. Moreover, while the IT Rules provide a framework for coordination between intermediaries and LEAs[57] establishing a standardised operating procedure is essential to enhance this process's efficiency.[58] This standardised procedure should clearly outline the officials authorised to make requests and the format for these requests, ensuring streamlined and effective coordination.

D. Compliance

1. Evaluative Dialogue for Compliance

Instead of solely relying on increasing compliance duties, a more effective strategy would involve direct evaluation of the efforts put in towards compliance. This approach should include regular dialogues between regulatory bodies and companies, creating a platform for mutual understanding and agreement on the best ways forward. Such interactions can lead to a deeper understanding of the challenges faced by companies and enable the development of more practical and impactful compliance strategies.

2. Global Collaboration for Compliance Standards

Complementing this progressive approach, efforts in international forums for deepfake regulation should be intensified. India, by actively participating in shaping global consensus and harmonising compliance standards, can ensure that its domestic policies align with international best practices. This global collaboration is essential in developing a coherent and unified response to the challenges posed by deepfakes, considering the borderless nature of digital media and technology.

[57] Rule 3 IT Rules, 2021.

[58] Shreya S. & Tiwari P. (2022, July 4). *IT Rules, 2021: A Regulatory Impact Assessment Study (Vol. 1).* New Delhi. The Dialogue and Internet And Mobile Association of India. https://thedialogue.co/wp-content/uploads/2022/07/IT-RULES-2021-interactive.pdf

AUTHORS

SHRUTI SHREYA

Senior Programme Manager, Platform Regulation and Gender and Tech, The Dialogue

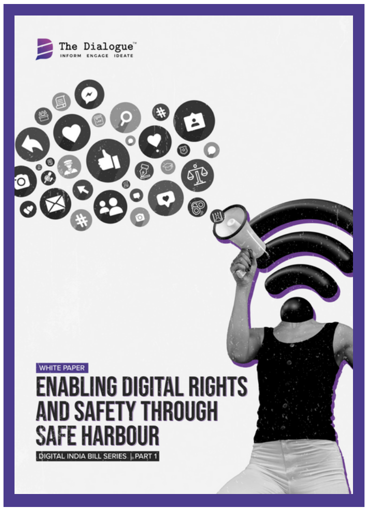
Shruti is a Senior Programme Manager for the Platform Regulation and Gender and Tech verticals at The Dialogue. A Gold Medalist lawyer by training, she is engaged in conducting interdisciplinary research on various aspects of social media governance, internet freedom, online safety, and feminist technologies.

PRANAV BHASKAR TIWARI

Non-Resident Fellow, The Dialogue

Pranav Bhaskar Tiwari, a Non-Resident Fellow at The Dialogue, contributes to the telecom, intermediary liability, and online gaming programs. A lawyer with specialised training in Intellectual Property and Technology Laws, Pranav's academic excellence is marked by a Gold Medal in International Law and Diplomacy from the Indian Society of International Law. His track record is characterised by extensive engagement with stakeholders in the tech policy arena and global fellowships with institutions including the Internet Society, ICANN, and the University of Chicago. Pranav's expertise extends to research-based advocacy, focusing on critical areas such as cybersecurity, online safety, data protection, and the implications of emerging technologies.

MORE FROM OUR RESEARCH


The Dialogue
INFORM ENGAGE IDEATE
WHITE PAPER
ENABLING DIGITAL RIGHTS AND SAFETY THROUGH SAFE HARBOUR
DIGITAL INDIA BILL SERIES | PART 1

WHITE PAPER

Enabling Digital Rights and Safety Digital India Bill Series Part-1

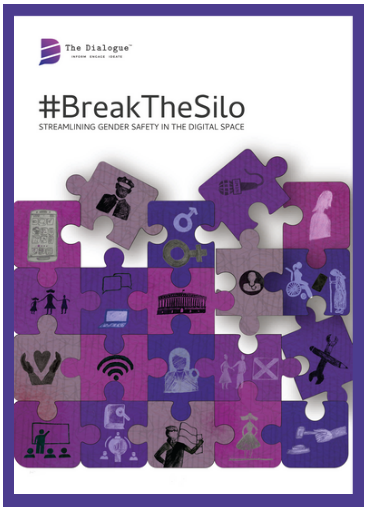
The Dialogue
INFORM ENGAGE IDEATE
Written Comments
RECOMMENDATIONS ON THE DRAFT BROADCASTING REGULATION SERVICE BILL, 2023

WRITTEN COMMENTS

Recommendations on the Draft Broadcasting Regulation Service Bill, 2023

RESEARCH REPORT

IT Rules, 2021: A Regulatory Impact Assessment Study Volume 2 | July 2023



POLICY FRAMEWORK

#BreakTheSilo: Streamlining Gender Safety in The Digital Space

LinkedIn | The Dialogue

Twitter | The Dialogue

Facebook | The Dialogue

Instagram | The Dialogue