REIMAGINING CONTENT MODERATION STRATEGIES IN THE AGE OF GENERATIVE AI

CUTS International 1983-2023

#2511

Reimagining Content Moderation Strategies in the Age of Generative AI

Published by

CUTS International
D-217, Bhaskar Marg, Bani Park, Jaipur 302016, India
Tel: +91.141.2282821, Fax: +91.141.2282485
Email: cuts1@cuts.org, Web site: www.cuts-international.org

Author: Asheef Iqubbal, Senior Research Associate, CUTS International

Citation: Iqubbal, A. (2025). Reimagining Content Moderation Strategies in the Age of Generative AI. CUTS International.

#2511

Contents

Foreword

In the digital age, information is not only abundant; it is alive, evolving, and increasingly synthetic. As Generative Artificial Intelligence (Gen-AI) reshapes our ability to produce and manipulate text, image, video, and audio; humanity finds itself standing at the edge of a new epistemic frontier. The question is no longer simply how we govern speech, but how we govern its simulation. This seminal report could not be more timely or urgent. At a moment when the velocity of AI innovation far outpaces the response time of laws and institutions, this work brings both clarity and depth to an increasingly complex conversation.

The phenomenon of synthetic media, while rooted in innovation, now sits at the intersection of trust, privacy, speech, and identity. Its applications span the hopeful and the horrific. On one hand, synthetic media enables educational transformation, multilingual accessibility, immersive storytelling, and creative empowerment. On the other hand, it is weaponised to mislead electorates, impersonate public figures, facilitate gendered violence, and destabilise societal cohesion. With great care and analytical rigour, this report explores the multifaceted promise and peril of this technological revolution.

What sets this report apart is not merely its comprehensiveness, but its conscientiousness. It recognises that the harms of synthetic media are not abstract—they are intimate, often traumatic, and disproportionately borne by women, children, and marginalised communities. The author gives voice to those whose identities are most vulnerable to being commodified, distorted, or violated in the synthetic realm.

Importantly, Iqubbal challenges us to reimagine moderation not as a censorial impulse, but as a collective responsibility shared by developers, platforms, users, and the state. This reframing is both ethically significant and practically necessary. The status quo, which reacts only after harm has occurred, fails to account for the psychological, reputational, and social costs that accompany synthetic abuse. The call for proactive, rights-sensitive safeguards rooted in constitutional values of dignity, free expression, and due process is an essential corrective.

The report's global comparative lens also strengthens its value. By drawing lessons from jurisdictions like the European Union and the United States, and examining approaches such as labelling, watermarking, transparency obligations, and AI safety institutes, the author offers a global perspective with local adaptability. And yet, the work does not fall into the trap of techno-solutionism. It remains attuned to the moral dilemmas posed by content moderation in democracies, such as "How do we protect speech while preventing harm? How do we ensure platform accountability without enabling surveillance? How do we distinguish satire from deception, art from abuse, and innovation from manipulation?"

These are not easy questions, and this report does not pretend to offer simple answers. Instead, it offers a principled framework rooted in context, complexity, and compassion. It invites us to imagine a future in which AI serves the public good, but does not overpower the human condition; where innovation is not pursued at the expense of justice; and where digital spaces remain inclusive, safe, and authentic.

To policymakers, this report is a roadmap. To researchers, it is a foundation. To civil society, it is a call to action. And to citizens, it is a reminder that in the synthetic age, our vigilance, values, and veracity matter more than ever.

As India and the world stand on the cusp of a synthetic media explosion, "Reimagining Content Moderation Strategies in the Age of Generative AI" must be read not just as a policy report but as a social imperative.

Sujeet Kumar
Member of Parliament (Rajya Sabha)
Member, Science & Technology, Environment & Forests
Member, Consultative Committee for External Affairs
Member, Indian Council of World Affairs
Member, Executive Committee, Bar Association of India

Acknowledgement

I acknowledge the numerous individuals who contributed to the success of this report through their inputs, discussions, reviews, encouragement, and guidance. I extend my sincere gratitude to stakeholders from research institutions, civil society organisations, start-ups, fact-checking organisations, think tanks, and industry representatives in India for their valuable insights.

I deeply appreciate the support and feedback from colleagues at CUTS International. Special thanks go to Sohom Banerjee (Research Associate), Samridh Shastry (Policy Analyst), and Krishaank Jugiani (Senior Research Associate) for their invaluable inputs throughout various stages of the project, from data analysis and stakeholder consultations to editing and structuring.

I am particularly grateful to Amol Kulkarni, Director (Research), CUTS International, for his guidance and faith. His ability to see connections between disparate ideas consistently opens paths forward when the research seems fragmented. His presence is felt throughout the report, not as impositions, but as kind yet firm nudges toward greater clarity and insight.

I also acknowledge the efforts of Akshay Sharma and Yamini Kumawat (Programme Team), Sweepthish Jayan, and Keval Sharma (Information Technology Team), Madhuri Vasnani and Mukesh Tyagi (Publications Team) at CUTS International for their exceptional support in bringing this report to fruition.

I express gratitude to all individuals, whether named above or not, without whom the publication of this report would not have been possible.

Any remaining errors are solely my responsibility.

Asheef Iqubbal
Senior Research Associate, CUTS International

Executive Summary

Generative Artificial Intelligence (Gen-AI) has transformed the creation of synthetic media. Gen-AI tools offer remarkable speed and accessibility, supporting applications across sectors such as marketing, education, translation, and media generation. For example, a university in India can use synthetic media tools to convert online lectures into multiple regional languages. With voice cloning and text-to-speech technology, it can create realistic voiceovers of professors, allowing students to learn in their native languages without losing tone or clarity. This would improve access to quality education for students in remote and underserved areas.

However, they also pose risks, including misinformation, identity-related fraud, and non-consensual synthetic media. For instance, nearly 35,000 publicly available synthetic media model variants have been identified, many fine-tuned to create synthetic media of identifiable individuals.[1] Notably, 96% of synthetic media outputs target marginalised social groups including women, from celebrities to everyday social media users.[2] These synthetic media are so sophisticated that even expert observers and basic fact-checking tools like reverse image searches or detailed verification are insufficient to debunk them and often require forensic analysis.[3]

The technology itself is not harmful; however, its potential for malicious use raises questions about how to balance its benefits with the growing risks. Policymakers, industry actors and civil society are implementing and/ or proposing measures to address synthetic media risks, including platform obligations, restrictions on non-consensual synthetic media, and labelling of synthetic media.[4] While steps in the right direction, these efforts face practical and implementational challenges which create enforcement gaps that malicious actors can exploit.

For example, the absence of a coordinated approach to labelling synthetic media can result in over labelling or under labelling, potentially misleading users into believing

1 Dramatic rise in publicly downloadable deepfake image generators. Available at: https://www.oii.ox.ac.uk/news-events/dramatic-rise-in-publicly-downloadable-deepfake-image-generators/
2 Dramatic rise in publicly downloadable deepfake image generators. Available at: https://www.oii.ox.ac.uk/news-events/dramatic-rise-in-publicly-downloadable-deepfake-image-generators/
3 Synthetic Media Policy: Provenance and Authentication — Expert Insights and Questions. Available at: https://www.techpolicy.press/synthetic-media-policy-provenance-and-authentication-expert-insights-and-questions/
4 Artificial illusion: Global governance challenges of deepfake technology. Available at: https://iapp.org/news/a/artificial-illusion-global-governance-challenges-of-deepfake-technology

that unlabelled media is human-generated or that all labelled media is synthetic. Effective labelling requires collaboration across platforms, AI developers, internet service providers, law enforcement agencies, and civil society. However, there is no clear framework that assigns responsibility across the AI value chain. This includes determining whether developers, deployers, or users should be accountable.

Techniques like watermarking systems face challenges that undermine their effectiveness as a solution to synthetic media concerns. These systems would only function correctly when detection tools are specifically calibrated to recognise the watermarks, an alignment that may not always exist across the diverse AI ecosystem. The current landscape of AI tools is fragmented, with proprietary systems using incompatible approaches that may not be able to communicate with one another, limiting the practical utility of watermarking as a broad solution. This fragmentation creates the possibility of incompatible watermarking systems, making it increasingly difficult for users to determine which authentication signals deserve trust.

Further, methods like hashing and metadata tracking may require collecting or analysing personal data, creating potential privacy risks and a breach of anonymity. Online anonymity is essential for multiple reasons and should not be equated with malicious activity. For example, anonymity provides a safe space for survivors of domestic violence, sexual abuse, or LGBTQ+ individuals in hostile conditions to seek support without exposing their identities. By tracking everyone, the regulatory mechanisms risk treating all users as potential offenders, shifting the burden of proof onto individuals. Assigning identifiable and unique identities does not always ensure accountability, and anonymity does not imply non-compliance with the law.

Detection technologies and synthetic media creation tools are locked in a constant arms race, with detectors often falling behind. Creators only need to produce synthetic media convincing to humans, while detectors must identify subtle. Moreover, detection systems trained on existing synthetic patterns become vulnerable to future techniques, leading to false positives and undermining reliability. As public awareness of synthetic media grows, a wider crisis of trust emerges–people may doubt all digital media. This scepticism may fuel the 'liar's dividend' where genuine media are dismissed as AI fabrications to evade accountability. Together, these issues can destabilise the information system and threaten the notion of shared factual reality.

While the challenges are on spectrums, existing regulations remain largely reactive, focused on addressing harm only after it occurs. They fail to establish proactive safeguards to prevent the creation and spread of malicious synthetic media. This is compounded by the fact that malicious use often outpaces local law enforcement, which lacks the skills to investigate such harms, especially when perpetrators use tactics like VPNs to hide their identity.

Moreover, labelling synthetic media as fake after it has been created does little to mitigate the harm experienced by victims. For example, when a synthetic pornographic image is made, even if it is flagged, the damage has already been done. For victims, the violation persists regardless of classification. However, broad content moderation can lead to the removal of legitimate uses of synthetic media, like educational or satirical media. Therefore, proactive safeguards must be carefully designed to mitigate harm without inadvertently restricting legitimate use cases.

Key Recommendations

Reimagining Liability: Liability frameworks for synthetic media must be context-sensitive. A risk-based approach should calibrate regulatory oversight to potential harm, guided by collaboratively developed harm classifications and principles. Foundational mandatory safety codes and standards should be integrated throughout the AI system lifecycle, with protective measures scaled proportionally to the level of risk. Independent oversight bodies, constituting diverse stakeholders, should monitor compliance and publish public reports to enhance transparency and build trust. In addition, establish specialised AI risk mitigation mechanisms, such as liability insurance pools and collective reserves, funded by deployers and developers of AI, to distribute risk across the AI ecosystem and ensure compensation, regardless of individual entity resources.

Collaborative AI Governance: Relying solely on voluntary commitments creates problems, as it fails to establish binding obligations and leaves AI systems' operations opaque. Therefore, the proposed AI Safety Institute (AISI) should play a key role in facilitating an iterative, collaborative process for synthetic media governance by involving civil society, academics, and experts to set benchmarks, establish transparency standards and code of practice. Through collaboratively designed codes and standards, stakeholders can build a framework that promotes innovation, protects individual rights, and mitigates risks. AISI can leverage these partnerships without imposing rigid rules that negatively impacts innovation or legitimate expressions.

Empowering the Public: Top-down interventions are important, but grassroots action is essential to bridge the gap between intent and practice. Stakeholders must prioritise developing reliable, accessible AI detection and verification tools to protect users from harmful synthetic media. Instead of government-mandated tools, India should promote market innovation and strengthen independent fact-checking networks. Inspired by Taiwan's "nerd immunity" approach, which combines professional fact-checkers with public participation, India could facilitate a collaborative ecosystem where social media platforms integrate fact-checking directly into their apps to inform users without being restrictive.

1. Introduction: Contextualising Synthetic Media Landscape

Generative AI (Gen-AI) has enabled the creation of synthetic media, including voices, images, and videos, offering valuable applications across fields.[5] For instance, in education, teachers can create virtual historical figures using Gen-AI to provide interactive and engaging learning experiences for students.[6] Producing a video using traditional methods involves resources such as script, locations, equipment, crew, and editing, all contributing to significant costs. In contrast, synthetic media requires none of these resources. With just a few prompts, Gen-AI can create a video, image, and audio.[7] This process is remarkably accurate and scalable.[8]

Due to advancements in Gen-AI, the use of synthetic media has increased multifield.[9] Owing to its positive potential, Gen-AI is expected to boost global productivity by adding US$2.6 to US$4.4tn value each year.[10] The full scope of the impacts of synthetic media is still evolving, along with the associated risks. For example, deepfakes, a subset of synthetic media, are increasingly being created both with and without the consent of the individuals depicted. These are often used to exploit social fault lines and enable fraud. Due to recent events of fraud, misinformation and non-consensual sexual media, the term "deepfake" is usually associated with negative connotations.[11]

Between 2022 and 2023, for example, the Internet saw a staggering 400 percent increase in synthetic sexual media[12] and an unprecedented 3000 percent rise in

5 We refer to synthetic media, including deepfakes, as images, videos, audio, or combinations created using generative artificial intelligence. We avoid the term 'synthetic content,' which includes AI-generated text, since our focus is not on text-based content due to its different challenges. We use 'authentic' for human-created media and 'synthetic' for artificial media. Available here: https://www.drcf.org.uk/siteassets/drcf/pdf-files/the-future-of-synthetic-media.pdf?v=385978

6 Deepfake Technology: The Risks, Benefits and Detection Methods. Available at: https://www.linkedin.com/pulse/deepfake-technology-risks-benefits-detection-methods-sahota-%E8%90%A8%E5%86%A0%E5%86%9B-/

7 Regulating Deepfakes: An Indian perspective. Available at: https://digitalcommons.usf.edu/cgi/viewcontent.cgi?article=2245&context=jss

8 5Examples of Generative AI-Made Videos for Corporate Use. Available at: https://hourone.ai/blog/5-examples-generative-ai-made-videos/

9 When AI Gets It Wrong: Addressing AI Hallucinations and Bias. Available at: https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias/

10 The economic potential of generative AI: The next productivity frontier. Available at: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier

11 Deepfakes are a real political threat. For now, though, they're mainly used to degrade women. Available at: https://www.vox.com/2019/10/7/20902215/deepfakes-usage-youtube-2019-deeptrace-research-report

12 State Of Deepfakes: Realities, Threats, and Impact. Available at: https://www.securityhero.io/state-of-deepfakes/#key-findings

synthetic media-enabled fraud.[13] A survey reveals that over 75 percent of India's online population have encountered such media, with at least 38 percent falling victim to related scams.[14] The World Economic Forum warns that AI-driven false and misleading information, capable of eroding democracy and deepening social polarisation, poses one of the immediate risks to the global economy.[15]

It may be tempting to argue that synthetic media is not a novel phenomenon. They can be viewed as equivalent to mis/disinformation. For instance, a synthetic media portraying a political leader as intoxicated and disoriented might not seem substantially different from a pre-AI doctored image, or someone falsely claiming that they saw the leader intoxicated and disoriented. Such statements are not typically regulated or prohibited unless there is an independent justification. Why, then, should synthetic media be treated differently? A probable answer is that synthetic media has a greater influence than verbal assertions or photoshop.

Synthetic media are self-authenticating: the human mind does not easily dismiss them, and even if it does, there remains a doubt.[16] Gen-AI, for example, can produce realistic media that deceives people into believing false news or impersonations, posing serious risks for misinformation, fraud, and identity-based crimes.[17] These risks are being compounded as India faces the highest risk of misinformation and disinformation, according to the 2024 Global Risk Report.[18] They pose a threat by worsening a scenario[19] in which users increasingly struggle to differentiate truth from falsehood, often ending up believing those who are most adept at deceiving them.[20]

Moreover, India's vulnerability to misinformation is linked to its high internet usage, limited media literacy, and fragility of mainstream media. Synthetic media is becoming faster, cheaper, and can be produced at an unimaginable scale which would only compound these challenges. A 2023 study reflected this trend, reporting a 550

---

13 Identity Fraud Report. Available at: https://onfido.com/landing/identity-fraud-report/

14 75% Indians have viewed some deepfake content in last 12 months, says McAfee survey. Available at: https://economictimes.indiatimes.com/tech/technology/75-indians-have-viewed-some-deepfake-content-in-last-12-months-says-mcafee-survey/articleshow/109599811.cms?from=mdr

15 AI-powered misinformation is the world's biggest short-term threat, Davos report says. Available at: https://www.thehindu.com/sci-tech/technology/ai-powered-misinformation-is-the-worlds-biggest-short-term-threat-davos-report-says/article67726959.ece

16 AI-powered misinformation is the world's biggest short-term threat, Davos report says. Available at: https://www.denverpost.com/2020/01/08/sunstein-facebook-deepfake-ban/

17 Deepfakes: regulatory challenges for the synthetic society. Available at: https://www.sciencedirect.com/science/article/pii/S0267364922000632#sec0004

18 Survey Finds False Information Risk Highest in India. Available at: https://thewire.in/media/survey-finds-false-information-risk-highest-in-india

19 A study in the USA found that over 40% of people believed false AI-generated content. In a country like India, with low digital literacy and a fragile media system, people are even more likely to accept misleading AI outputs as truth, leading to harmful consequences. Available at: https://link.springer.com/article/10.3758/s13428-023-02124-2?utm_medium=affiliate&utm_source=commission_junction&utm_campaign=CONR_BOOKS_ECOM_GL_PBOK_ALWYS_DEEPLINK&utm_content=textlink&utm_term=PID100095187&CJEVENT=cad2db2f920711ef83b200740a18b8f9

20 Digital Arrest: The Modern-Day Cyber Scam. Available at: https://www.niti.gov.in/sites/default/files/2025-04/Digital_Arrest_The_Modern_Day_Cyber_Scam.pdf

percent increase in synthetic media since 2019.[21] This underscores the dual risk of synthetic media: first, the technology offers a powerful tool for spreading false information, and second, the difficulty of distinguishing fake videos with the naked eye threatens to erode the value of media altogether.[22]

As synthetic media evolve, existing and proposed mitigation strategies require critical analysis. ***Chapter 2*** sets the stage by outlining the report's methodological framework and research approach. Policymakers must balance the harms of malicious synthetic media with the need to protect innovation and legitimate speech. Synthetic media can be used for beneficial purposes, such as satire, artistic expression, and social commentary, which are protected under Indian free speech laws. ***Chapter 3*** analyses these varied contexts to inform more nuanced policy responses. A context-agnostic framework could have unforeseen consequences for protected speech.

Chapter 4 traces the evolution of synthetic media regulation in India, mapping how policymakers grapple with emerging risks while upholding the principles of free speech. Building on this, ***Chapter 5*** surveys regulatory strategies from other jurisdictions, including the EU, USA, and China, offering comparative insights that can guide the development of more context-sensitive approaches in India. Platforms are also developing their moderation strategies for synthetic media. However, the lack of consistency in standards and definitions may undermine these efforts. ***Chapter 6*** discusses the practices of synthetic media moderation across different platforms and how lack of a standardised approach across platforms may create gaps.

Chapter 7 analyses emerging approaches to combat malicious synthetic media, including labelling, transparency, and consent mechanisms. It critically assesses the effectiveness and limitations of these tools in responding to the challenges posed by increasingly sophisticated AI-generated media. Most existing measures to combat malicious synthetic media, such as synthetic media being taken down, remain largely reactive, coming into play only after harm has already occurred, as analysed in ***Chapter 8***. While these are welcome steps, they may fall short due to challenges in detection, labelling, and potential unforeseen risks like violation of privacy and surveillance.

The complexity should not prevent the development of new or updated regulatory frameworks for synthetic media. Policymakers must establish articulated objectives and carefully formulate rules that address potential harms without unduly restricting protected expression and innovation. One of the key challenges regulators faces when governing Gen-AI models is determining how to distribute liability among different

[21] State Of Deepfakes: Realities, Threats, And Impact. Available at: https://www.securityhero.io/state-of-deepfakes/

[22] Deepfake detection with and without content warnings. Available at: https://royalsocietypublishing.org/doi/10.1098/rsos.231214

actors within the ecosystem. This includes deciding the extent of responsibility for the user, developer, and deployers in the event of harm. For instance, accountability must be assigned when an AI model creates problematic media.

In the context of India, the report analyses the applicability of existing liability frameworks that were not designed to address the challenges of AI-generated media in ***Chapter 9***. Competing tensions within the AI ecosystem highlight the need for more proactive governance. In ***Chapter 10***, the report provides context-sensitive governance strategies that can adapt to the evolving technological landscape while maintaining core protections for individual rights and innovation.

2. Methodology and Approach

This report uses a mixed-methods approach combining desk research, policy analysis, and stakeholder consultations. First, we began with a comprehensive literature review in three areas: (1) technical methods for creating strategies, including detection tools, and platform policies. Sources include peer-reviewed articles, white papers, conference proceedings, and industry case studies published mainly between 2017 and 2025, the period in which the term "deepfake" was coined.[23]

Second, we analysed policy documents: legislative bills, ministerial advisories, court judgments, and official press releases from India, as well as regulatory texts from the EU, United States, United Kingdom, and China. We reviewed position papers and consultation drafts produced by multilateral bodies such as OECD, UNESCO, and World Economic Forum to capture emerging global trends. We also examined large-scale surveys, sentiment analysis, and public-opinion datasets on synthetic media produced by research institutes, industry, and other bodies. Finally, using a comparative framework that maps each jurisdiction's approach, voluntary codes, risk-based regulation, or strict labelling, against key policy objectives: user safety, free expression, innovation, and enforceability.

Third, we conducted 14 stakeholder consultations with fact-checking platforms, digital-rights organisations, subject-matter experts, technologists, industry representatives, academics, and other relevant actors. Discussions assessed the effectiveness of existing and proposed rules for curbing misuse of synthetic media, the role of labelling, and the balance between platform accountability, developers, deployers, users, and individual responsibility. To protect their privacy and confidentiality, interviewees were anonymised. This multi-layered approach maps common trends, context-specific gaps, and practical lessons for strengthening India's synthetic media moderation approach. Insights from these consultations were triangulated with secondary research to develop recommendations.

Analysing stakeholders' views involved several key steps. First, the data pre-processing stage involved cleaning the interview transcripts to prepare them for analysis. This included tokenising the data by breaking the text into individual words, removing stop words, and applying lemmatisation to convert words to their base forms. All words

[23] A Brief History of Deepfakes. Available at: https://www.realitydefender.com/insights/history-of-deepfakes

were lowercased for consistency. Second, Latent Dirichlet Allocation (LDA) was used for topic modelling. LDA, which assumes that each document is a mixture of topics and each topic is a mixture of words, was used to extract a defined number of issues from the data. The model's coherence was evaluated to ensure the identified topics were meaningful. The resulting themes were interpreted and analysed.

3. Framing Synthetic Media: Promise, Peril, and Definitional Ambiguity

The chapter explores synthetic media's dual nature, outlining its potential benefits and associated risks. The involvement of diverse technologies in generating synthetic media adds complexity to its definitional scope. Given this expansive scope, regulatory approaches may need to balance mitigating harm and preserving legitimate forms of expression. Stakeholders reflect a range of perspectives, from concerns about harmful applications such as non-consensual synthetic media to support for creative or educational uses. The broad range of tools and contexts used to generate synthetic media makes it difficult to define and regulate effectively.

The growth of synthetic media have largely been driven by advances in Generative Adversarial Networks (GANs). A GAN consists of two neural networks: one generates media based on sample data, while the other evaluates the output, rejecting anything that does not resemble the original. These rejections help the generator improve through a recursive, self-learning cycle. Once confined to expert use, GAN technology is now accessible through user-friendly tools, allowing anyone to create synthetic media.[24] By converting complex audio-visual traits into teachable data points, these models can precisely superimpose an individual's likeness onto source media, producing highly realistic, frame-by-frame audio-visual media that can be quickly generated and distributed at scale.

Politicians, for example, used Gen-AI for voter outreach in the 2024 general election across different countries.[25] Using AI, Indian political parties campaigned in local languages, sending personalised messages to communities. This contributed to India's synthetic-media market, which grew into a US$60mn industry.[26] More than 50 million AI voice-clone calls were conducted two months before voting began. Polymath Synthetic Media Solutions was among many firms working for political clients. It ran five AI-driven campaigns that season, earning about US$55,000, charging roughly

24 An Investigation of the Effectiveness of Deepfake Models and Tools. Available at: https://www.mdpi.com/2224-2708/12/4/61

25 Nilesh Christopher. Available at: https://x.com/NilChristopher/status/1792522178807669072?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1792522178807669072%7Ctwgr%5E9b67acfc8eafd2e546d1a4fa12daf1ca17c760fa%7Ctwcon%5Es1_&ref_url=https%3A%2F%2Fwww.niemanlab.org%2F2024%2F05%2Findian-journalists-are-the-frontline-against-election-deepfakes%2F

26 Indian Voters Are Being Bombarded With Millions of Deepfakes. Political Candidates Approve. Available at: https://www.wired.com/story/indian-elections-ai-deepfakes/

₹125,000 (US$1,500) for a digital avatar and ₹60,000 (US$720) for an audio clone, far less than major consultancies. After collecting basic data, the platform could generate videos or voice calls directly to voters' phones, delivering personalised outreach at scale. Using human call-centre staff costs about ₹4 per call, eight times more than AI-generated calls, making the economic appeal clear.[27]

The technology behind synthetic media is not inherently harmful and can be leveraged for constructive and beneficial uses. These include enhancing educational interactivity, improving customer service, and saving time by automating labour-intensive tasks. For instance, Krisp Accent Conversion is a Gen-AI tool that adjusts an Indian-English accent in real time, making speech easier for native English listeners to understand without altering the speaker's natural voice. This technology could transform customer-service interactions.[28] Similarly, synthetic media has the potential to expand accessibility tools by making them more affordable and tailored to individual needs. These personalised solutions can support people with disabilities in exercising greater agency and achieving independence.

Opportunities	Challenges
• Enhances education through interactive, virtual content (e.g., historical figures brought to life). • Saves time and reduces production costs compared to traditional media creation methods. • Makes synthetic media creation accessible through user-friendly Gen-AI tools. • Supports creativity and entertainment and promotes innovation in areas like digital storytelling and marketing. • Potential for AI tools to increase accessibility (e.g., text-to-speech, localisation, etc.)	• Used for non-consensual sexual media, especially targeting women and marginalised groups. • Democratisation of Gen-AI enables malicious users to create harmful synthetic media. • Fake synthetic media undermines trust and confidence in audio-visual media. • It can be weaponised to manipulate public opinion or incite political and social polarisation and violence. • Psychological harm to victims, especially in cases of pornography, financial fraud, identity theft.

However, synthetic media can also be weaponised to manipulate public opinion, spread misinformation, perpetrate identity fraud, or enable forms of abuse such as

[27] Indian Voters Are Being Bombarded With Millions of Deepfakes. Political Candidates Approve. Available at: https://www.wired.com/story/indian-elections-ai-deepfakes/

[28] Krisp. Available at: https://krisp.ai/ai-accent-conversion/

revenge pornography. For example, synthetic media disproportionately target women and children and can cause psychological and reputational harm. In 2024 alone, over 300 million children globally have been victims of online sexual exploitation and abuse. A growing concern is the rise of AI-generated nudes.

According to Graphika, millions visit over 100 "nudify" websites monthly. With just a few clicks, AI can create disturbingly realistic nude images from ordinary, fully clothed photos.[29] Malicious users no longer need explicit images; any photo can be exploited. A 2019 report analysing over 14,000 deepfake videos found that 96 percent were non-consensual intimate media.[30] In 2020, a report prepared for the United States Department of Homeland Security documented over 100,000 computer-generated fake nude images of women, created without their knowledge or consent.[31]

In 2024, for example, sexually explicit AI-generated images of singer Taylor Swift spread on social media platforms like X, with one post reportedly viewed over 47 million times. The photos, depicting Swift nude in sexual scenarios at a Kansas City Chiefs game, were likely created using a free text-to-image AI tool. This was eventually taken down after a mass-reporting campaign led by Swifties, a term used for fans of the singer-songwriter Taylor Swift. Though aimed at Swift amid misogynistic backlash to her appearances at NFL games, the incident underscored a broader, systemic issue: the creation and circulation of synthetic pornography is a growing threat that extends far beyond celebrities, affecting individuals across society.[32]

Though the incident involving Swift sparked global conversation on malicious use of synthetic media, it was far from an isolated case. Swift's case was an exception because her deepfakes were taken down quickly. Most victims, including public figures, are not as lucky. Their images often stay online, spreading without accountability or removal, leaving it up to individual viewers to decide whether to believe or not.[33]

Individuals often rely on their perceptions to navigate reality and distinguish between genuine and authentic. This extends beyond personal contexts to professional domains, including law enforcement, where officers depend on their sensory

---

29 Schools face a new threat: "nudify" sites that use AI to create realistic, revealing images of classmates. Available at: https://www.cbsnews.com/news/schools-face-new-threat-nudify-sites-use-ai-create-realistic-revealing-images-60-minutes-transcript/

30 Synthetic content and its implications for AI policy: a primer. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000392181

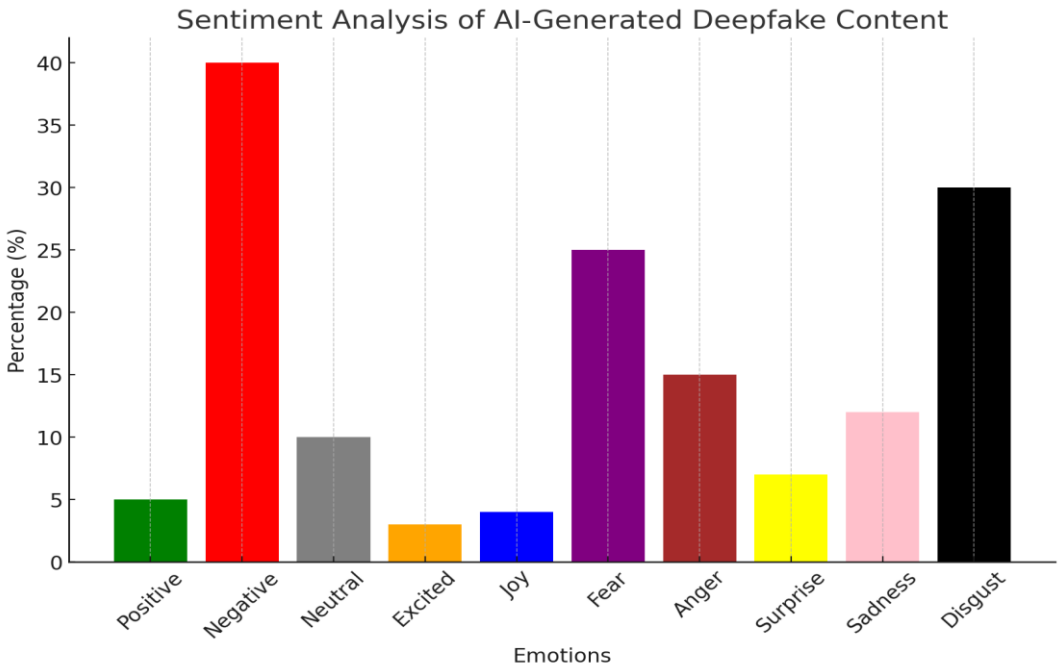
31 Increasing Threats of Deepfake Identities. Available at: https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf

32 Taylor Swift Deepfakes Show What's Coming Next In Gender and Tech – And Advocates Should Be Concerned. Available at: https://www.techpolicy.press/taylor-swift-shows-us-whats-coming-next-in-gender-and-tech-and-advocates-should-be-concerned/

33 Inside the Taylor Swift deepfake scandal: 'It's men telling a powerful woman to get back in her box'. Available at: https://www.theguardian.com/technology/2024/jan/31/inside-the-taylor-swift-deepfake-scandal-its-men-telling-a-powerful-woman-to-get-back-in-her-box

experiences to perform their duties. Primary accounts of events generally accord greater credibility than secondary reports.

Similarly, auditory, and visual recordings are often considered reliable representations of reality. Photographs and videos are crucial in intelligence gathering for police work and are vital evidence in legal proceedings.[34] Gen-AI models capable of generating realistic synthetic media can shift how individuals assess and determine the trustworthiness of media. These risks, if left unaddressed, could undermine public trust in audio-visual media and erode confidence in digital media.[35]



In the consultations, stakeholders were asked to describe their understanding and perception of deepfake technology and its uses. The overwhelming response among stakeholders highlights a critical distinction between different types of synthetic media: while some AI-generated media can be creative or beneficial, deepfakes have been disproportionately weaponised to cause personal harm, with victims predominantly being women and marginalised social groups.[36] This pattern of misuse of Gen-AI including deepfake technologies threatens personal autonomy and psychological well-being. These negative applications[37] have overshadowed any

34 AI-powered misinformation is the world's biggest short-term threat, Davos report says. Available at: https://www.thehindu.com/sci-tech/technology/ai-powered-misinformation-is-the-worlds-biggest-short-term-threat-davos-report-says/article67726959.ece

35‘ Deepfakes' ranked as most serious AI crime threat. Available at: https://www.ucl.ac.uk/news/2020/aug/deepfakes-ranked-most-serious-ai-crime-threat

36 Discussing Deepfakes: The opportunities and challenges of synthetic media technology. Available at: https://www.techuk.org/resource/discussing-deepfakes-how-could-we-respond-to-the-threat-of-new-deepfake-technologies-and-seize-the-potential-of-synthetic-media-technologies.html

37 Synthetic content and its implications for AI policy: a primer. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000392181

constructive potential, leading to widespread alarm about synthetic media and their impact on individual privacy, identity, and personal dignity.

The graph above presents the results of a sentiment analysis conducted on text from stakeholder interviews. Sentiments around AI-generated deepfakes are predominantly negative, with stakeholders frequently expressing emotions such as fear and disgust. Without intervention, such abuse and its negative impact on vulnerable social groups will only grow.[38]

The Complexity of Defining Synthetic Media

As discussed, synthetic media applications constitute diverse contexts, complicating definitional ambiguity and classification. Despite global efforts to develop risk taxonomies, the expanding spectrum of use cases and technological variations challenges regulatory frameworks. Policymakers face challenges in clearly distinguishing between different categories of synthetic media and the technologies used to create them, which limits the development of nuanced governance approaches aimed to address potential harms while supporting beneficial applications.

A central challenge lies in determining whether synthetic media must be deceptive to trigger regulation, or whether factors like intent, harm, and consent should also play a role. Consider a synthetic video of the Prime Minister speaking multiple languages.[39] Though it used Gen-AI to make the speech appear authentic, the purpose was informative, not deceptive, or harmful. Still, it qualifies as synthetic media due to how it was created.

There are multiple interpretations of synthetic media.[40] One view holds that AI must necessarily generate synthetic media. Others argue that the definition should clarify which media types it includes, for instance, whether to include text, given that people engage with text and visual media differently. Some suggest using the term "audio" instead of "voice" to more accurately capture the range of sounds that can be synthetically produced. Another view is that intent should be included in how we define synthetic media, to help separate harmless uses from harmful ones.

---

38 On June 24, 2024, former Reserve Bank of India (RBI) Governor Raghuram Rajan flagged a deepfake video circulating on social media that falsely showed him endorsing investments in specific stocks. In a LinkedIn post, Rajan clarified, "These videos are fake and the perpetrators should be reported to the relevant authorities. I do not give investment advice to the public or promote individual stocks." He advised against following such content, noting that people are generally better off with diversified portfolios, such as bank deposits, bonds, and mutual funds, rather than risking money on individual stocks or options. "You may get lucky, but more likely, you'll end up poorer," he warned. Available at: https://www.medianama.com/2024/06/223-whats-the-takeaway-from-fake-videos-using-raghuram-rajans-identity/

39 Indian language AI tool 'Bhashini' used to translate Prime Minister Narendra Modi's speech. Available at: https://www.thehindu.com/sci-tech/technology/indian-language-ai-tool-bhashini-used-to-translate-prime-minister-narendra-modis-speech/article67649545.ece

40 The Future of Synthetic Media. Available at: https://www.drcf.org.uk/siteassets/drcf/pdf-files/the-future-of-synthetic-media.pdf?v=385978

Some people believe that synthetic media are always misleading or manipulative, so they can never really be considered "good." Others look at how much the fake media looks like a real person. Some define synthetic media more narrowly, arguing they must involve recreating a real person's face or voice, based on how such use cases first started with face-swapping and voice dubbing. Others adopt a broader definition, including fabricated scenes like warzone footage that do not involve real people.

As synthetic media increasingly spans a range of tools, techniques, and contexts, from advanced AI to basic editing software, and from creative use to non-consensual sexual content, developing a more nuanced definition has become essential. This is key for shaping regulatory frameworks that distinguish between benign, beneficial uses (such as education or entertainment) and harmful, deceptive ones.

4. India's Struggle to Mitigate Synthetic Media Harms

The chapter outlines India's evolving approach to synthetic media by examining key executive actions, such as MeitY's advisories on deepfakes, and reviewing legislative initiatives like the proposed Digital India Act, which advocates for risk-based AI regulation and a re-evaluation of safe harbour protections. It also analyses the findings of the MeitY Subcommittee on AI Governance to contextualise the policy discourse around synthetic media. In addition, the chapter explores recent judicial interventions that address the misuse of deepfakes, call for legislative safeguards against AI-related harms, and reaffirm constitutional limits on state control over online speech.

MeitY's Mixed Messaging

The emergence of synthetic media like deepfakes and their capacity to spread false information has pushed governments to conceive appropriate interventions. In India, this has been seen in a series of advisories from MeitY and judicial interventions, showing both urgency and uncertainty in policy responses. The Information and Technology (IT) Minister, in 2024, indicated that the government is working on amendments to the IT Rules to address deepfakes, with plans to increase social media platforms' responsibilities in combating issues related to synthetic media.

The Minister indicated the need to have balanced responsibilities with the positive impact of digital communication.[41] An official press release dated 24 November 2023 announced the government's intent to work with stakeholders to develop actionable strategies for detecting, preventing, reporting, and raising awareness of deepfakes, indicating potential regulatory developments.

[41] Govt bringing laws, taking steps to make social media platforms accountable: IT Minister Ashwini Vaishnaw. Available at: https://indianexpress.com/article/news-today/ashwini-vaishnaw-laws-social-media-platforms-accountable-9153186/

Leaders Sound the Alarm on Deepfakes

In November 2023, Prime Minister Narendra Modi publicly raised concerns about the misuse of artificial intelligence to create deepfakes, calling it a "new crisis" and urging the media to educate the public.[42] He warned that many AI-generated deepfakes appear realistic and could spread harmful disinformation. Citing a viral deepfake of himself dancing Garba, later debunked, PM Modi remarked, "There is a huge section of society which does not have a parallel verification system."[43] He suggested that deepfakes should also come with clear disclosures like cigarette packs carrying health warnings.

Similar warnings have come from other leaders. Echoing these sentiments, in November, IT Minister Ashwini Vaishnaw described deepfakes as a new threat to democracy. He outlined that accountability should rest with both creators and platforms, and noted that the government is considering penalties for both.[44] In December, President Droupadi Murmu acknowledged AI's benefits but flagged deepfakes as a growing threat to society, calling for moral value-based education to address the challenge.[45]

On 26 December 2023, MeitY issued its first advisory focused on intermediary due diligence under the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021.[46] This advisory directed all intermediaries, especially digital platforms, to reinforce compliance with Rule 3(1)(b) of the IT Rules, which prohibits false, defamatory, or harmful content. Platforms were required to inform users about such prohibited content at various stages: registration, login, and while uploading content.

Rule 3(1)(b) of the IT Rules requires intermediaries to "make reasonable efforts to cause users to avoid posting certain types of content. The sub-clauses (i) to (ix) under this rule, which outline the grounds for content removal, have undergone amendments. Notably, Rule 3(1)(b)(v) now includes a provision against "knowingly and intentionally communicating misinformation or information that is patently false, untrue, or misleading in nature."

---

42 PM Modi warns against deepfakes; calls on media to educate people on misinformation. Available at: https://www.thehindu.com/news/national/pm-modi-warns-against-deepfakes-calls-on-media-to-educate-people-on-misinformation/article67543869.ece

43 Saw video where I'm singing Garba': PM flags deepfakes, says matter of concern. Available at: https://www.indiatoday.in/india/story/deepfakes-one-of-the-biggest-threats-facing-indias-system-can-cause-chaos-says-pm-modi-2464032-2023-11-17

44 Centre To Meet Social Media Platforms Over Deepfake Videos: IT Minister. Available at: https://www.ndtv.com/india-news/centre-to-meet-social-media-platforms-over-deepfake-videos-it-minister-4585095

45 Prez Murmu redflags AI misuse to create deepfakes. Available at: https://timesofindia.indiatimes.com/city/nagpur/prez-murmu-redflags-ai-misuse-to-create-deepfakes/articleshow/105704020.cms

46 Advisory. Available at: https://www.meity.gov.in/static/uploads/2024/02/c9f89809b63d22656be38a166ef14949.pdf

The advisory mandated platforms to communicate to users about restricted content through their terms of service, privacy policies, and agreements. A central aim was to educate users about the legal implications of violating these provisions, including potential prosecution under the Indian laws including IT Act, 2000. The advisory also reiterated the responsibility of the platforms to report any such violations to law enforcement agencies. The then Minister, Rajeev Chandrasekhar, stated that misinformation seriously threatens user safety and trust on the Internet. Deepfakes, as AI-powered misinformation, further amplify this risk for our Digital Nagriks.[47]

A second advisory followed on March 1, 2024, to further tighten oversight.[48] This expanded the regulatory scope by directly addressing emerging concerns around AI, including large language models and Gen-AI systems. The advisory required platforms to ensure that AI-driven systems do not facilitate activities that violate Rule 3(1)(b) of the IT Rules or any applicable laws. It introduced strict content labelling mechanisms, such as pop-ups and metadata tagging, to alert users to the fallibility and potential bias in AI-generated content.

It also recommended embedding unique identifiers or metadata in AI-generated or modified media, especially where such content could be weaponised as misinformation, to ensure traceability to both the intermediary and the originator. Moreover, it required users to be explicitly warned about potential legal consequences of violating the rules, ranging from content removal and account suspension to criminal prosecution. However, one of the most contentious provisions of the advisory mandated that under-tested or unreliable Gen-AI models could not be offered to Indian users without government approval. To enforce compliance, the advisory demanded that platforms submit an Action Taken-cum-Status Report within 15 days.

While the intention behind the measures may have been to pre-empt misuse, particularly around electoral interference and misinformation, advisories were criticised as restrictive and knee-jerk responses. Critics argued that requiring prior approval from the government could stifle innovation and negatively impact the potential of emerging technologies. Acknowledging these concerns, the government issued a revised directive on March 15,[49] 2024, which superseded the earlier advisory. This was a notable policy shift. While the core focus on transparency remained intact,

47 MeitY issues advisory to all intermediaries to comply with existing IT rules. Available at: https://pib.gov.in/PressReleasePage.aspx?PRID=1990542

48 Due diligence by Intermediaries / Platforms under the Information Technology Act, 2000 and Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021. Available at: https://regmedia.co.uk/2024/03/04/meity_ai_advisory_1_march.pdf

49 Due diligence by Intermediaries / Platforms under the Information Technology Act, 2000 and Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021. Available at: https://www.meity.gov.in/static/uploads/2024/02/9f6e99572739a3024c9cdaec53a0a0ef.pdf

the contentious aspects, namely, the requirement for prior government approval and the submission of compliance reports, were dropped.

The latest advisory suggested labelling requirements and warned platforms about protecting electoral integrity, but adopted a softer enforcement tone, framing its approach as more suggestive and emphasising voluntary, collaborative efforts. While the advisory mentions labelling AI-generated media, it shifts some responsibility for identifying and curbing misinformation onto users. Moreover, the then State IT Minister clarified on social media that this advisory would only apply to large platforms, not startups, though the criteria for this distinction were not specified.[50] Platforms have also been instructed to deploy AI models that prevent users from posting or sharing unlawful content.[51]

The Ministry's shifting positions have caused confusion and uncertainty among stakeholders including industry and users. They show knee-jerk, surface-level interventions that fail to address the complex challenges of synthetic media. Without meaningful consultation and collaboration with diverse stakeholders, inconsistent advisories risk further regulatory uncertainty for technology platforms and users.

Digital India Act

On 9 March 2023, Rajeev Chandrasekhar, then Minister of State for Electronics and Information Technology, announced in a public consultation that the IT Act of 2000 would be replaced by the upcoming Digital India Act (DIA), a future-ready legislation. A presentation outlining its key features was shared during the session. Since then, MeitY has undertaken consultations to gather feedback from various stakeholders. The government is still drafting the legislation and reassessing its release timeline. The draft has sparked debate on how the internet should be regulated today. The DIA proposes a unified legal framework to regulate high-risk AI systems, with provisions for algorithmic accountability, zero-day threat and vulnerability assessments, and oversight of AI-driven advertising and content moderation.

The proposed core pillars of the DIA include online safety, trust and accountability, an open internet, and regulation of emerging technologies such as AI and blockchain. The Ministry is also reviewing the 'safe harbour' provision, which protects social media platforms from liability for user-generated content. The rationale behind the proposed law is that India's current regulatory framework struggles to keep pace with digital transformation and emerging technologies. With nearly a billion users, the digital landscape has evolved significantly. A range of intermediaries, including e-commerce

[50] MeitY Issued an Advisory Regulating AI. Available at: https://www.cyberpeace.org/resources/blogs/meity-issued-an-advisory-regulating-ai

[51] MeitY Issued an Advisory Regulating AI. Available at: https://www.cyberpeace.org/resources/blogs/meity-issued-an-advisory-regulating-ai

platforms, social media, OTT, online gaming, and AI systems, now operate in the ecosystem, which demands a more nuanced and forward-looking regulatory response.

AI Governance Report

More recently, MeitY undertook stakeholder consultation on the AI Governance Guidelines Development Report.[52] The Subcommittee report highlights existing legal safeguards to address the misuse of AI models in creating malicious synthetic media. The report argues that depending on the context and impact, laws like the Information Technology Act, 2000 and Bhartiya Nyaya Sanhita may apply on AI generated media.[53]

However, it is important to note that the advancement of AI models would often outpaces existing legal responses, compromising user safety. Comprehensive regulation through clear statutes and rules would be crucial to address evolving harms effectively.[54] Existing legal systems, designed before the rise of advanced digital replication technologies, struggle to address the challenges posed by realistic digital replicas, leaving individuals vulnerable to misuse.[55]

For example, in 2023, actor Anil Kapoor sought protection of his personality rights. He obtained a restraining order against the unauthorised use of his name, nicknames like "Lakhan" and "Mr. India," iconic dialogues such as "Jhakaas," and his voice and images. The court ruled in Kapoor's favour, with Justice Pratibha Singh noting that technological advancements, including AI tools, have made exploiting a celebrity's persona easier, infringing on their right to privacy.

Thus, there is a need for legal recognition and remedies for AI-related harms. The subcommittee emphasises that 'minimising harm' should be the central objective of AI regulation. This aligns more with the EU's rights-based approach, such as the AI Act, than the 'pro-innovation' stance of countries like the UK. However, as discussed below, the report does not clarify whether the subcommittee supports a statutory instrument similar to the EU's AI Act to achieve this.

The report aims to establish liability across the AI ecosystem. Taking a lifecycle approach, the report argues that multiple actors are involved at various stages of an AI system's development and deployment. For instance, in the lifecycle of a foundation

---

52 India acknowledges both the potential and challenges of AI. To address these, an Advisory Group chaired by the Principal Scientific Advisor has been established to develop an 'AI for India-Specific Regulatory Framework'. Guided by this group, a Subcommittee on 'AI Governance and Guidelines Development' was formed to provide actionable recommendations. Its report aims to support the creation of a trustworthy and accountable AI ecosystem in India. Available at: https://indiaai.gov.in/article/report-on-ai-governance-guidelines-development

53 Report on AI Governance Guidelines Development. Available at: https://indiaai.s3.ap-south-1.amazonaws.com/docs/subcommittee-report-dec26.pdf

54 No AI FRAUD Act. Available at: https://www.congress.gov/bill/118th-congress/house-bill/6943

55 No AI FRAUD Act. Available at: https://www.congress.gov/bill/118th-congress/house-bill/6943

model, a wide range of stakeholders may participate, including data principals, data providers, AI developers (such as model builders), AI deployers (such as app builders and distributors), and end-users (including both businesses and citizens).

While well-intentioned, it overlooks the distinction between end-users and affected users, those impacted by AI systems but not direct users. In many AI applications, the individuals affected by an AI's decisions or outputs do not interact with the system. For example, a person whose image, video, or voice data has been used by an AI model may not be a system user, but they are nonetheless directly impacted.

Moreover, the Subcommittee Report on AI Governance highlights several challenges and offers commendable recommendations, but lacks clear direction. It continues to emphasise 'voluntary' commitments, which may weaken the impact and intent of its recommendations. The IT Minister has often stated that 'heavy' AI regulation is not suitable for India, favouring lighter and innovation-friendly approaches.[56]

Voluntary commitments are a welcome, but fall short of ensuring the accountability and transparency expected from AI developers and deployers.[57] Given AI's disruptive potential, as acknowledged by the subcommittee, relying solely on such measures fails to create meaningful obligations toward citizens and risks allowing developers and deployers to do the bare minimum, with key information remaining opaque.[58]

Judicial Interventions

Parallel to executive efforts, the judiciary began shaping the synthetic content discourse. On August 28, 2024, the Delhi High Court asked the Union Ministry of Electronics and Information Technology to consider enacting legislation to regulate harmful uses of artificial intelligence, such as deepfakes, on the internet. The court emphasised that these tools pose a significant threat to society, stating, "Everything you see or hear is fake. It cannot be."[59]

The bench noted that some U.S. states have already introduced similar laws and suggested it is time for the Indian Parliament to act. The court added, "As an institution, we have limitations. The government must act; this is going to be a serious menace in society."[60] The court's remarks came against the backdrop of a case

56 Making AI Self-Regulation Work. Available at: https://cerai.iitm.ac.in/docs/selfregulation.pdf
57 India's Advance on AI Regulation. Available at: https://carnegieendowment.org/research/2024/11/indias-advance-on-ai-regulation?lang=en&center=india
58 India's AI Governance Guidelines Report: A Medley of Approaches. Available at: https://www.techpolicy.press/indias-ai-governance-guidelines-report-a-medley-of-approaches/
59 Deepfake technology going to be serious menace, says Delhi HC. Available at: https://www.ptinews.com/story/national/deepfake-technology-going-to-be-serious-menace-says-delhi-hc/1772862
60 Deepfake technology going to be serious menace, says Delhi HC. Available at: https://www.ptinews.com/story/national/deepfake-technology-going-to-be-serious-menace-says-delhi-hc/1772862

involving a renowned television journalist, who had sought an injunction against individuals allegedly circulating deepfake videos misusing his name and likeness to promote pharmaceutical products without consent. In December 2024, the court granted an ex parte injunction, acknowledging the serious reputational harm and public deception caused by such acts.[61]

Recognising the journalist's stature and the public's trust in him, the court underscored the "heightened sense of accountability" that distinguished his case from typical celebrity endorsements. It concluded that unauthorised use of his persona through Gen-AI or deepfake technologies constituted irreparable personal and societal harm. Notably, the injunction barred using the journalist's image, voice, or likeness, now and in the future, through any AI or synthetic technology, setting a precedent that anticipates the misuse.

Developments in India's Regulatory Response to Synthetic Media

Timeline	Summary
9 March 2023	During a consultation, Rajeev Chandrasekhar, then Minister of State for Electronics and Information Technology, announced that the DIA would replace the Information Technology Act of 2000.
26 Dec 2023	An initial advisory was issued, focused on combating misinformation, especially deepfakes. Platforms were directed to inform users clearly about prohibited content under the IT Rules and legal consequences.
01 Mar 2024	Additional advisory. Expanded rules to cover AI/LLM/Generative AI. Government approval was required for under-tested AI models, and Action Taken Reports were demanded from platforms.
15 Mar 2024	Superseding advisory. Withdrawn the government approval requirement and scrapped the action taken report deadline. Took a more advisory tone while retaining core obligations.
28 Aug 2024	Delhi High Court urged MeitY to legislate on AI misuse, calling deepfakes a serious societal threat.

[61] Delhi High Court Grants Injunction Against Using India TV Editor Rajat Sharma's Likeness In Deepfake Videos. Available at: https://www.verdictum.in/court-updates/high-courts/injunction-likeness-personality-infringement-rajat-sharma-v-tamara-doc-1147-of-2024-1561798

Timeline	Summary
21 Nov 2024	The Ministry of Electronics and Information Technology is funding two research projects focused on detecting AI-generated fake videos, audios, and images.[62]
20 Dec 2024	The Court granted an injunction restraining the misuse of the journalist's persona via deepfakes, citing the risk of public deception and personal harm.
06 Jan 2025	An Advisory Group, chaired by the principal scientific advisor, has been established to develop an 'AI for India-Specific Regulatory Framework'. Working under its guidance, a Subcommittee on 'AI Governance and Guidelines Development' was published for public consultation, aiming to shape practical recommendations for AI governance in the Indian context.
31 Jan 2025	MeitY has proposed the establishment of the AI Safety Institute under the Safe and Trusted Pillar of the IndiaAI Mission to address AI risks and safety challenges. The Institute will collaborate with academia, startups, industry, and government bodies to promote safety, security, and trust in AI. It will focus on advancing indigenous research and development using Indian datasets, tailored to the country's diversity.[63]

The above developments highlight a regulatory environment underpinned by uncertainty and reactive responses to emerging concerns about synthetic media. Instead of forward-looking strategies, regulators address crises and criticism as they arise, creating a fragmented regulatory landscape. Additionally, sector-specific requirements, both mandatory and advisory, are spread across multiple regulators, including the Reserve Bank of India,[64] the Securities and Exchange Board of India,[65] the Telecom Regulatory Authority of India,[66] and the Competition Commission of

62 MeitY seeks tools to detect deepfakes, label AI generated content. Available at: http://hindustantimes.com/india-news/meity-seeks-tools-to-detect-deepfakes-label-ai-generated-content-101734410291642.html

63 India Takes the Lead: Establishing the IndiaAI Safety Institute for Responsible AI Innovation. Available at: https://indiaai.gov.in/article/india-takes-the-lead-establishing-the-indiaai-safety-institute-for-responsible-ai-innovation

64 RBI sets up panel to develop a framework on ethical use of AI in financial sector. Available at: https://www.thehindu.com/business/rbi-sets-up-panel-to-develop-a-framework-on-ethical-use-of-ai-in-financial-sector/article69029678.ece

65 SEBI's Proposed New Amendments on Usage of AI Tools by Regulated Entities. Available at: https://corporate.cyrilamarchandblogs.com/2024/12/sebis-proposed-new-amendments-on-usage-of-ai-tools-by-regulated-entities/

66 Consultation Paper on Leveraging Artificial Intelligence and Big Data in Telecommunication Sector. Available at: https://www.trai.gov.in/consultation-paper-leveraging-artificial-intelligence-and-big-data-telecommunication-sector

India.[67] Each applies its institutional lens to AI and its governance, resulting in a fragmented and decentralised approach that can lead to uncertainty.

Moreover, without justification, a one-size-fits-all approach risks capturing benign Gen-AI applications, such as customer care helplines. These reactionary policies could stifle innovation and penalise legitimate uses with minimal or no risk. They fail to distinguish between synthetic media's genuine risks and valuable benefits, preventing the establishment of proportionate interventions. Without clear regulatory direction, India will continue to face synthetic media challenges through a patchwork of disconnected measures, offering inadequate protection and no certainty for responsible AI innovation.

Other Relevant Developments

Fact-Checking Unit

The Fact Check Unit (FCU) under the Press Information Bureau (PIB) was established in November 2019 to deter the creation and dissemination of fake news and misinformation. It was introduced in the government's amendments to the IT Rules, 2021, to flag fake, false, or misleading content. Social media intermediaries were required to remove such content or risk losing their safe harbour protections under the IT Act, 2000. The FCU notification was challenged in the Bombay High Court by the Association of Indian Magazines, the Editors Guild of India, and stand-up political comic Kunal Kamra. The challenge arose from concerns over censorship and content takedowns, as it positioned a government body as the "primary arbiter of online news."

Establishing an FCU at the Union government's discretion to verify online information violates justice and threatens free speech and the right to access information.[68] The case was referred to a third judge after a division bench of the Bombay High Court delivered a split verdict.[69] The court held that the rules violated Articles 14 and 19, which guarantee equality and free speech, noting that mere falsity is not a valid reason to restrict expression. The FCU, appointed by the executive, lacked sufficient checks and was seen as unfair, with no equivalent oversight for print media or state-level content. The rules were also vague and posed a potential chilling effect on free expression, with less intrusive alternatives available to achieve the same objectives.[70]

---

67 CCI engages Management Development Institute Society (MDIS) for Market Study on Artificial Intelligence and Competition. Available at: https://www.pib.gov.in/PressReleasePage.aspx?PRID=2054673

68 Sanity Prevails as Bombay High Court Strikes Down India Government's Fact Check Unit. Available at: https://www.techpolicy.press/sanity-prevails-as-bombay-high-court-strikes-down-india-governments-fact-check-unit/

69 Centre's Bid To Establish Fact-Checking Unit Struck Down By Bombay High Court. Available at: https://www.ndtv.com/india-news/centres-bid-to-establish-fact-checking-unit-struck-down-by-bombay-high-court-6610082

70 Bombay High Court officially strikes down Centre's Fact Check Unit, calls amended IT Rules 'unconstitutional' . Available at: https://www.thehindu.com/news/national/bombay-hc-formally-strikes-down-centres-fact-check-unit-calls-amended-it-rules-unconstitutional/article68684934.ece

Moderating synthetic media faces risks like those encountered with the establishment of the FCU and presents an even more complex challenge. For instance, during the 2024 election, many pieces of so-called 'synthetic media' taken down were not synthetic. An edited video of Home Minister Amit Shah went viral, for example, falsely claiming he promised to end reservations. Although fact-checking revealed the video was doctored, it was labelled a 'deepfake'[71] by several mainstream media outlets.[72] Future synthetic media regulations could be misused similarly to remove legitimate media if checks and balances, like those absent in the FCU case, are not in place.

The Bombay High Court's judgment highlights three key principles for future legislation on online misinformation, which can be applied to synthetic content:

Truth and free speech: Free speech does not guarantee the expression of absolute truth, and the state cannot act as the sole arbiter of what is true.

Unambiguous restrictions: Any restriction on speech must be precisely defined and consistent with constitutional limits on free expression.

Proportionality: Legal measures must be proportionate and narrowly tailored to address specific harms.

Broadcasting Services (Regulation) Bill, 2024

The Indian Ministry of Information and Broadcasting has proposed the Broadcasting Services (Regulation) Bill, 2024, to update the Television Network Act of 1995 and expand oversight to digital content creators.[73] The bill broadly defines 'professional' and 'digital news broadcaster,' covering online creators from YouTube financial advisors to social media news posters. It mandates registration with the government for those providing news, current affairs, or systematic information sharing. It requires content evaluation committees with representatives from diverse groups, including women, child welfare advocates, scheduled castes, tribes, and minorities. Committee membership details must also be submitted to the government.

The bill raises critical questions about its applicability to emerging issues related to synthetic content, which can generate content, provide investment advice, and generate opinions on current affairs. While the proposed legislation aims to ensure content quality, its approach to synthetic content remains unclear. As AI-generated content becomes increasingly sophisticated, the bill's regulatory framework may struggle to monitor and control these new forms of digital communication effectively.

---

71 This means video was doctored using basic editing, not AI-generated; calling it a 'deepfake' may be misleading, as deepfakes specifically involve synthetic media created using AI.

72 The Problem With Wrongly Labelling Real Or Edited Videos As 'Deepfakes'. Available at: https://www.boomlive.in/news/deepfakes-or-real-edited-videos-amit-shah-bjp-congress-rahul-gandhi-lok-sabha-elections-2024-ai-25107

73 Bill 2024 and its impact on digital creators. Available at: https://www.business-standard.com/india-news/explained-new-broadcasting-bill-2024-and-its-impact-on-digital-creators-124080900665_1.html

5. A Comparative Assessment of Regulatory Responses to Synthetic Media

The chapter surveys how major jurisdictions regulate or propose to regulate synthetic media. The European Union's AI Act classifies deepfakes as a "limited-risk" category, requiring transparency disclosures. The United States targets specific issues, such as the DEFIANCE Act for non-consensual explicit deepfakes, the No AI FRAUD Act to protect personal likenesses, and the Take It Down Act aims to ensure removal of non-consensual intimate AI-generated media, particularly to protect minors. The UK's Online Safety Act criminalises intimate deepfakes and imposes obligations on social-media platforms. China's Deep Synthesis mandates watermarking and strict authentication for AI-generated media. These diverse approaches taken by different jurisdictions show a spectrum of regulatory design choices. Understanding their strengths and limitations can help Indian policymakers develop a nuanced framework that mitigates harm without unduly constraining innovation or lawful expression.

The European Union's AI Act

The European Union's EU AI Act has been seen as a significant development in the AI landscape. The AI Act defines a deepfake as AI-generated or altered media that falsely appears authentic or truthful, covering not just people and events but also objects and entities like products or organisations. This broad scope reflects the potential for reputational and financial harm. By distinguishing between authenticity (appearance) and truthfulness (accuracy), the Act sets a higher standard.[74] While it does not ban deepfakes outright, it imposes transparency requirements on both providers and users to mitigate harm through accountability.[75] This includes notifying users when they are interacting with media created or modified by AI, except when it is obvious that the content is artificial.[76]

Within the AI Act's risk classification system, deepfakes are currently categorised as "limited risk" AI systems, subjecting them to fewer regulations compared to "high-

74 Article 50 of the Artificial Intelligence Act, available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689

75 Article 50 of the Artificial Intelligence Act, available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689

76 EU AI Act unpacked #8: New rules on deepfakes. Available at: https://www.lexology.com/library/detail.aspx?g=c25237ee-a37f-4959-837c-f32a624f54ab

risk" systems such as medical AI or facial recognition technologies.[77] The Act relies on monetary penalties, the most severe penalties, up to €35mn or seven percent of global annual turnover, whichever is higher, apply to breaches involving prohibited AI systems. The lowest tier of fines, capped at €7.5mn or one percent of global turnover, applies to operators who provide false, incomplete, or misleading information.[78] However, monetary fines as an accountability mechanism are reactive, addressing harm after the fact rather than preventing the creation or dissemination of malicious synthetic media at the source.

Given the growth of AI technologies and the challenges in tracing the origin of digital content, the AI Act requires platforms to embed technical features that indicate when AI has generated or altered content. These features must allow AI-generated images, audio, or video, especially deepfakes, to be formatted in a machine-readable format, making detection easier.[79] Users of AI systems must clearly label any AI-generated media and disclose its artificial origin in a way that is easy to understand. The user can choose the exact method of disclosure, depending on the context. There are limited exceptions. Disclosure is not required if the content is used for legally permitted activities such as detecting or investigating criminal offences. For artistic, satirical, fictional, or similar works, the law only requires a suitable indication of the presence of AI, provided it does not interfere with the audience's experience.

The EU AI Office was also established in 2024 to promote responsible AI practices across the EU.[80] A key function of this office is to encourage and facilitate the development of codes of practice at the Union level, aimed at effectively implementing obligations related to the detection and labelling of synthetic content.[81] The European Commission is empowered to approve these codes of practice, ensuring they meet specific standards and effectively address the challenges posed by such content. Further, if the Commission deems a particular code of practice inadequate, it can adopt implementing acts to address deficiencies.

Labelling all synthetic media are easier to imagine legally but may fall short in practice. Not all viewers or listeners may notice or understand disclaimers indicating the artificiality of content due to language ambiguities or limited technical literacy. Although watermarking technologies for synthetic media have advanced, they can still

77 EU AI Act: first regulation on artificial intelligence. Available at: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

78 European Union AI Act. Available at: https://artificialintelligenceact.eu/article/99/

79 Techniques like watermarking, metadata, provenance tracking can be used to verify a content's source, history, and modifications, making it easier to assess whether and how it has been altered. This aims to ensure users are informed about the origin of the content they engage with, reducing the risk of deception or misuse.

80 European AI Office. Available at: https://digital-strategy.ec.europa.eu/en/policies/ai-office

81 Artificial Intelligence – Questions and Answers. Available at: https://ec.europa.eu/commission/presscorner/detail/en/qanda_21_1683

be bypassed, limiting their effectiveness as a reliable safeguard. The Act does not currently provide a clear framework for holding developers of synthetic media legally liable. Instead, it focuses on reactive measures such as monetary penalties. Proactive mitigation of the harms of synthetic media could include requiring developers to implement technical safeguards, such as effective watermarking or advanced detection algorithms, to prevent the misuse of synthetic media.

United States Legislative Initiatives

DEFIANCE Act

The introduction of a bipartisan bill in the United States Senate is a move towards addressing the growing concern of non-consensual, sexualised images generated by AI. This legislation, known as the Disrupt Explicit Forged Images and Non-Consensual Edits Act of 2024, or the "Defiance Act," is a direct response to the penetration of AI-generated pornographic images, particularly those involving high-profile individuals such as Taylor Swift on social media platforms.[82]

The legislation aims to prevent the spread of non-consensual, sexually explicit deepfakes and establish a legal framework for victims of such AI-generated content to seek civil remedies. Under the Defiance Act, individuals depicted in nude or sexually explicit "digital forgeries" would be empowered to pursue compensation against those responsible for producing, possessing with intent to distribute.

According to the Act, "Digital forgery" constitutes visual content created using software, machine learning, artificial intelligence, or other technological methods that falsely appear realistic. As per the bill, victims are potentially eligible to receive damages for cases like intimate visual depiction.[83] In more severe instances, where the incident is linked to actual or attempted sexual assault, stalking, or harassment, or is deemed to be the direct and proximate cause of such harms, the compensation could increase to US$250,000.[84]

No AI FRAUD Act

The No AI FRAUD Act aims to establish federal protections for American citizens against AI's unauthorised fakes of their voices and images. This proposed legislation seeks to address the growing concerns surrounding the misuse of personal likenesses in the digital age, particularly in light of advancements in AI-generated synthetic

82 Rep. Ocasio-Cortez Leads Bipartisan, Bicameral Introduction of DEFIANCE Act to Combat Use of Non-Consensual, Sexually-Explicit "Deepfake" Media. Available at: https://ocasio-cortez.house.gov/media/press-releases/rep-ocasio-cortez-leads-bipartisan-bicameral-introduction-defiance-act-combat

83 S.3696 - DEFIANCE Act of 2024. Available at: https://www.congress.gov/bill/118th-congress/senate-bill/3696

84 Left to Shoulder the Worries of AI. Available at: https://www.csis.org/analysis/left-shoulder-worries-ai

media.[85] In other words, the Act aims to provide property rights over an individual's likeness and voice, protecting them from unauthorised use of AI to clone or impersonate, including for harassment, abuse, fraud, or other malicious purposes.[86]

The legislation provides a framework for individuals to pursue action against entities that create, facilitate, or disseminate synthetic media without explicit permission. One of the Act's key provisions is the establishment of punitive damages and attorneys' fees for those whose voices or images are misappropriated. This measure not only acts as a deterrent against potential misuse but also as a means of empowering individuals to enforce their rights effectively.[87] Beyond deterrence and enforcement, this provision underscores the importance of victim restitution. Explicitly recognising the right to compensation affirms that individuals affected by synthetic media misuse are entitled to redress for violations of identity, privacy, and personal dignity.

The Act extends liability to those who distribute, facilitate, transmit or contribute to the public availability of unauthorised AI-generated content, creating a broader scope of accountability. Recognising the complex intersection between individual rights and freedom of expression, the No AI FRAUD Act incorporates a First Amendment[88] defence balancing it against the intellectual property interest. The Act allows a free speech defence in cases where the use of someone's likeness may be justified.

Courts should consider several factors, including whether the content was created for commercial gain, whether the person's likeness is relevant to the primary purpose of the content, and whether the use competes with or harms the market for the original work. This ensures that the law respects both intellectual property and freedom of expression.[89] By doing so, the Act attempts to balance protecting individual rights and safeguarding speech and innovation.

However, the Act has faced criticism for its potentially overbroad scope. The definitions outlined in the legislation extend beyond explicitly fraudulent uses, constituting a wide array of content creation practices. This expansive scope raises concerns about its impact on various forms of creative expression, including parody, historical re-enactments, satirical depictions, and even casual impressions or memes. The Act's broad language could inadvertently affect legitimate forms of expression, potentially stifling creativity, and innovation in media production.[90]

---

85 No AI Fraud Act Bill. Available at: https://www.klemchuk.com/s/Draft-Bill_No-AI-FRAUD-Act.pdf
86 What is the No AI FRAUD Act? Available at: https://www.resemble.ai/no-ai-fraud-act/
87 H.R.6943 - No AI FRAUD Act. Available at: https://www.congress.gov/bill/118th-congress/house-bill/6943
88 First Amendment. Available at: https://www.law.cornell.edu/constitution/first_amendment
89 No AI FRAUD Act Would Create IP Rights to Prevent Voice and Likeness Misappropriation. Available at: https://americanlegaljournal.com/no-ai-fraud-act-would-create-ip-rights-to-prevent-voice-and-likeness-misappropriation/
90 The No AI Fraud Act Creates Far More Problems Than It Solves. Available at: https://www.eff.org/deeplinks/2024/01/no-ai-fraud-act-creates-way-more-problems-it-solves

The implications for content creators, performers, and even users of social media platforms are significant, as the Act's provisions could apply to a wide range of AI-generated media. The legislation also introduces implications for intermediaries and platforms hosting AI-generated media. The Act could fundamentally alter how platforms manage user-generated content by placing these entities directly in the line of potential litigation and excluding them from the protections typically offered by Section 230 immunity[91] for federal intellectual property claims. This may lead to increased caution among service providers, potentially resulting in more stringent content filtering or outright prohibition of certain AI-generated materials.

The potential consequences of the Act extend beyond individual creators to the broader ecosystem of digital content creation and distribution. Platforms and services may become increasingly wary of hosting or supporting user-generated content incorporating AI elements, fearing legal repercussions. This cautious approach could lead to a chilling effect on creative expression and innovation in the digital space.

Take It Down Act

The Take It Down Act is another legislative development in the United States to address the need for the removal of non-consensual, sexually explicit material.[92] The Act seeks to address gaps in protections against non-consensual intimate imagery (NCII), especially as AI heightens the risk, impacting victims including minors. While most U.S. states have laws against NCII, these vary widely in how they define offences and enforce penalties, leading to inconsistent and often ineffective legal outcomes.

Victims also face significant challenges in getting harmful content removed from online platforms, which prolongs exposure and emotional trauma.[93] Although Congress created a civil cause of action in 2022 for victims to sue those who publish NCII, this path is often impractical due to high costs, time demands, and the emotional toll of legal proceedings. Moreover, the anonymous and fast-moving nature of synthetic media makes it challenging to identify and pursue perpetrators.[94]

The Act introduces a federal framework that criminalises the publication or threat to publish NCII, including deepfakes, across state lines to address these issues. At its core, the Act mandates online platforms to remove non-consensual intimate imagery (NCII) within 48 hours of receiving a takedown request from a verified victim or their

---

91 In the U.S., Congress passed Section 230, originally part of the Communications Decency Act, to protect Americans' freedom of expression online by safeguarding the intermediaries users depend on. The law states: "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider." Available at: https://www.eff.org/issues/cda230

92 S.146 - Take It Down Act. Available at: https://www.congress.gov/bill/119th-congress/senate-bill/146/text

93 Inside the First Major U.S. Bill Tackling AI Harms—and Deepfake Abuse. Available at: https://time.com/7277746/ai-deepfakes-take-it-down-act-2025/

94 The TAKE IT DOWN Act. Available at: https://www.young.senate.gov/wp-content/uploads/1-pager_TAKE-IT-DOWN-Act_6.18.2024-FINAL.pdf

authorised representative. This includes content that is AI-generated, altered, or digitally manipulated to falsely depict a person in a sexual context.

The legislation criminalises the knowing publication or threat of publication of NCII in interstate commerce, including deepfakes and realistic, computer-generated pornographic images or videos that falsely depict identifiable individuals. It clarifies that consent to create an authentic image does not imply consent for public distribution. The Act supports good-faith efforts to help victims, such as disclosing NCII to law enforcement or for medical reasons. Platforms must remove reported content and make reasonable efforts to delete duplicates.

Enforcement lies with the Federal Trade Commission (FTC) to ensure oversight and compliance. It is argued that, to uphold constitutional protections, the legislation aims to avoid infringing on lawful expression. Instead, it uses a "reasonable person" standard to assess whether computer-generated content realistically depicts a real person, aligning with First Amendment requirements.

A notable highlight of the Take It Down Act is its strategic circumvention of Section 230 of the Communications Act, that frequently contested shield protecting platforms from civil liability for user content. Instead, the Act derives enforcement authority from the FTC's "deceptive and unfair trade practices" mandate. This sidesteps the contentious Section 230 debate, gaining acceptance from companies concerned about erosion of their legal protections.

Critics, however, question the Act's efficacy, particularly as enforcement rests with an FTC weakened under the current administration, potentially resulting in inconsistent application. More concerning is the Act's expansive takedown provision, which applies to a broader category than the narrower non-consensual intimate imagery definitions, while lacking safeguards against frivolous or bad-faith removal requests. The 48-hour removal requirement provides insufficient time for content verification, likely compelling platforms to default to content removal rather than risk legal exposure.[95]

United Kingdom's Legal Framework

The proposed Criminal Justice Bill makes it a crime to create sexually explicit synthetic media without consent. Under this bill, anyone who uses digital technology to produce intimate images of another person with the intent to cause harm may face prosecution, an unlimited fine, and a criminal record, even if the image is not shared.[96] Previously,

[95] Congress Passes TAKE IT DOWN Act Despite Major Flaws. Available at: https://www.eff.org/deeplinks/2025/04/congress-passes-take-it-down-act-despite-major-flaws

[96] Government cracks down on 'deepfakes' creation. Available at: https://www.gov.uk/government/news/government-cracks-down-on-deepfakes-creation

Section 33 of the Criminal Justice and Courts Act 2015, it was an offence to share sexual media without consent, with the intent to cause alarm, distress, or humiliation.

The Online Safety Act replaced the previous offence with Section 66B, making it a crime to "intentionally share or threaten to share" intimate images or videos of someone, regardless of intent to cause harm. This removes the earlier requirement to prove intent to cause alarm or distress, offering broader protection to victims.[97] This comes in the wake of data showing that deepfake-related abuse has increased by over 400 percent since 2017.[98]

The legislation imposes legal obligations on platforms to identify and mitigate risks associated with prohibited content. Social media platforms are required to proactively detect and address unlawful material, with a particular focus on protecting children from potential harm. Non-compliance with these measures may result in financial penalties, potentially reaching up to 10 percent of global annual revenue or £18 million, whichever is greater, and may expose senior employees to criminal liability. The Act is crucial in holding social media companies accountable for the content shared on their platforms. It places liability on these companies for harmful or illegal synthetic media, making them more responsible for moderating users' posts.

Further, the Act's emphasis on the liability of social media platforms could lead to excessive content moderation. To avoid legal repercussions, social media platforms may err on the side of caution and over-censor content, even when it is not harmful, or fail to act when it is.[99] For instance, a social media platform may remove posts that do not violate any laws but could be deemed controversial, as they want to reduce the risk of facing legal consequences.

Moreover, the Act does not include specific regulations that hold AI model developers and providers accountable for the 'design negligence' in Gen-AI platforms. Users aiming to misuse synthetic media may be unable to produce it if it is designed with appropriate safety measures. Existing law does not criminalise the creation of such malicious media through Gen-AI without consent. This legal gap exposes victims to potential harm, even if the content is never publicised. Just the existence of sexual synthetic media can negatively impact a person's well-being and their reputation.[100]

---

97 Tackling the regulation of sexually explicit deepfakes. Available at: https://www.kingsleynapley.co.uk/insights/blogs/criminal-law-blog/tackling-the-regulation-of-sexually-explicit-deepfakes

98 Britain to make sexually explicit 'deepfakes' a crime. Available at: https://www.reuters.com/world/uk/britain-make-sexually-explicit-deepfakes-crime-2025-01-07/

99 Facebook, Instagram, Snapchat and X failing to remove dangerous suicide and self-harm content. Available at: https://news.sky.com/story/facebook-instagram-snapchat-and-x-failing-to-remove-dangerous-suicide-and-self-harm-content-study-13197195

100 Deepfakes and the Law: Why Britain needs stronger protections against technology-facilitated abuse. Available at: https://www.qmul.ac.uk/media/news/2025/humanities-and-social-sciences/hss/deepfakes-and-the-law-why-britain-needs-stronger-protections-against-technology-facilitated-abuse.html

China's Deep Synthesis Regulations

China recently introduced provisions on Deep Synthesis to regulate AI technologies, particularly those involved in the creation of synthetic media, including deepfakes.[101] This constitutes a broad spectrum of AI-generated media, including text, images, audio, and video, and shows China's ongoing efforts to exert comprehensive control over internet activities within its borders. The Cyberspace Administration of China (CAC), the primary regulatory body overseeing these provisions, has articulated a justification for their implementation.

Central to their rationale is the need to address the misuse of deep synthesis technology, which has been implicated in the production, replication, publication, and dissemination of content deemed illegal or harmful by Chinese authorities. The CAC cites specific concerns such as defamation, identity theft, and fraud, arguing that these AI misuses can potentially disrupt social communication, infringe upon individual rights, and pose threats to national security and social stability.

The scope of these regulations is notably expansive, targeting two primary stakeholder groups: content generation platforms and end-users who utilise these services. According to the provisions, any content created with AI must be marked with a watermark, a text or image overlaid on the content to indicate it has been altered. Platform providers must adhere to several rules, including not processing personal information, authenticating users to ensure creators can be identified, and implementing consumer feedback mechanisms.

Further, platforms are required to verify users' real identities, using national ID numbers, mobile numbers, or public verification systems, before allowing them to publish synthetic media. The idea is to ensure traceability and accountability, particularly in cases of misuse. Providers must also set up accessible channels for users to submit complaints or appeals and publicly disclose the procedures and timeframes for addressing such feedback, ensuring transparency and user protection. Not only content must be clearly labelled, but the regulations also strictly prohibit any attempt to delete, modify, or conceal these AI-generated content labels.[102]

However, the implementation of such wide-ranging regulations raises crucial considerations. First, there are questions about the technical feasibility of effectively watermarking AI-generated content, particularly given the rapid evolution of AI technologies. Broad regulations may unintentionally impact legitimate AI use in creative and professional contexts. Further, mandatory labelling of AI-generated

[101] China: Provisions on Deep Synthesis Technology Enter into Effect. Available at: https://www.loc.gov/item/global-legal-monitor/2023-04-25/china-provisions-on-deep-synthesis-technology-enter-into-effect/

[102] Provisions on the Administration of Deep Synthesis Internet Information Services. Available at: https://www.chinalawtranslate.com/en/deep-synthesis/

media could stigmatise or devalue certain digital media, even when used constructively. For example, AI can be combined with human editorial expertise to enhance a short film. However, labelling such constructive use may hinder innovation due to the negative connotations often associated with synthetic media.

Please see a comparison matrix of global regulatory responses on the next page.

Comparison Matrix

Aspect	EU	US			UK	China	India
Regulatory Framework	AI Act, Digital Services Act	DEFIANCE Act	No AI FRAUD Act	Take It Down Act	Online Safety Act	Deep Synthesis Provisions (CAC)	IT Rules 2021 and MeitY Advisories
Focus	AI-generated image, audio, or video content that resembles persons, objects, places, entities, or events and would falsely appear to be authentic or truthful	Visual depiction created through AI or software to falsely appear authentic with the victim depicted in nude or sexual acts without consent	AI-generated or manipulated content that misuses a person's likeness or voice without their consent.	Knowingly sharing or threatening to share intimate images without a person's consent, including those created using AI, such as deepfakes, is prohibited	Sexually explicit content created using "computer graphics or any other digital technology" to cause alarm, distress, or humiliation without consent	Synthesised media using AI or deep learning to replace a person's likeness in images or videos with another	Synthetic media is aiming to spread misinformation
Regulatory Approach	Labelling and disclosure of synthetic media	Civil remedies for victims and platform account-ability	Civil compensation to victims; assigns legal liability to both creators and disseminators	Criminalises distribute-on and threats to publish NCII; mandates removal within 48 hours of notice; FTC oversight	Criminalisation of non-consensual sexual synthetic media	Mandatory watermarking, content traceability and user verification	Labelling synthetic media
Applicability	Content creators and disseminators AI and social media platforms	Content creators, disseminators, and AI and social media platforms	Content creators and distributors, AI developers and social media platforms	AI and social media platforms	Social media platforms, content creators and disseminators	AI platforms and users	Social media platforms
Enforcement Challenges	Limited AI developer liability, watermark efficacy issues, and risk of bypassing transparency measures	Broad definitions of synthetic media could suppress legitimate expression	Ambiguity in what qualifies as deceptive media; enforcement may chill innovation or fair use	Broad Scope of Takedown Provision; Lack of Safeguard Against Misuse; and Unrealistic 48-Hour Compliance window	Risk of over-censorship by social media platforms, no responsibilities of AI developers or deployers	Negative impact on free speech, online anonymity, and innovation	Voluntary compliance fragmented regulations, potential for over-censorship

6. Platform Approaches to Managing Synthetic Media Concerns

The chapter explores how platforms approach the moderation of synthetic media in the wake of emerging regulatory frameworks. By examining platform strategies to moderate synthetic media, the chapter aims to surface emerging approaches, identify potential gaps in implementation, and underline the possibilities of cohesive content moderation practices.

Integrating AI into everyday applications blurs the line between authentic and AI-generated media, posing unique challenges beyond traditional frameworks. For example, a news article might feature a genuine quote from a public figure, but the surrounding context could be AI-generated or manipulated. Similarly, a video may show a real individual speaking, but with subtle facial or voice alterations made by AI. Such complexities raise critical questions about how platforms should help users make informed decisions about the media they encounter. As online users become more vulnerable to manipulation, stakeholders must develop transparent and effective mechanisms to mitigate these risks.[103]

The evolution of synthetic media content moderation would be hampered by inconsistent definitions and approaches across digital platforms. Each platform independently develops its own classification systems and content moderation policies, creating a fragmented regulatory landscape that confuses users and complicates cross-platform governance efforts. This definitional inconsistency undermines the development of coherent industry standards and makes it difficult to establish clear expectations for creators and consumers alike.

For example, Meta mentions 'AI used in content' and maintains a policy of labelling such material with "AI info" across its platforms.[104] A key part of their label AI-generated content depends on standard indicators used across the industry, which help identify when AI tools were involved. Even minor edits using AI, like retouching, trigger labels like "Made with AI", while the "AI info" label for content is shown on content generated by an AI tool. Further, Meta's election-related policies focus on "preventing interference, fighting misinformation, and increasing transparency".

[103] AI or Your Lying Eyes: Some Shortcomings of Artificially Intelligent Deepfake Detectors. Available at: https://link.springer.com/article/10.1007/s13347-024-00700-8

[104] Our Approach to Labeling AI-Generated Content and Manipulated Media. Available at: https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/

However, Platforms like Meta aim to distinguish between included and excluded content, often based on the deceptive versus non-deceptive divide.[105]

YouTube has a policy that covers fake or altered content, including videos made using generative AI tools.[106] It requires creators to tell viewers when realistic-looking content, which could be mistaken for a real person, place, scene, or event, has been altered or made using AI tools. This includes swapping a person's face in a video, using AI to mimic someone's voice, faking real-world events like a fire in a building, or showing realistic but fictional disasters like a tornado hitting a real town. However, YouTube does not ask for disclosure if AI is used behind the scenes, like writing scripts or generating ideas, or if the content is fake or the changes are minor. That means no need to label content such as animations, fantasy scenes, colour filters, background blurs, or beauty effects.

X also mentions "Synthetic and Manipulated Media" in its policy, but does not call out generative AI-created media. However, it bars the use of fake personas for accounts, including those using AI-generated profile photos or bios.[107] It also bans sharing fake or altered content that could mislead people, cause confusion, or harm. This includes manipulated visuals or audio, out-of-context media, and fake profiles using AI-generated or stolen images. In contrast, TikTok explicitly uses the term "AI-generated content (AIGC)" and automatically labels content created when it is uploaded from other platforms.[108] It also requires creators to label realistic AI-generated content to ensure transparency with viewers.

Snapchat, on the other hand, prohibits sharing false or harmful information, including manipulating content made through generative AI.[109] AI-generated content through the Snap app is indicated through several ways, like use of sparkle icon, a disclaimer, a Context Card, or a tool tip.[110]

While above-mentioned approaches indicate industry progress, challenges around consistency, standardisation, and external accountability persist in moderating synthetic media. For example, the fragmentation in approaches creates challenges for user trust, and transparency, and standards. Synthetic media might be fully labelled on one platform, partially labelled on another, and completely unlabelled on a third

[105] How to identify AI content on Meta Products. Available at: https://www.meta.com/help/artificial-intelligence/1783222608822690/
[106] How we're helping creators disclose altered or synthetic content. Available at: https://blog.youtube/news-and-events/disclosing-ai-generated-content
[107] Authenticity. Available at: http://help.x.com/en/rules-and-policies/authenticity
[108] Partnering with our industry to advance AI transparency and literacy. Available at: https://newsroom.tiktok.com/en-us/partnering-with-our-industry-to-advance-ai-transparency-and-literacy
[109] Harmful False or Deceptive Practices. Available at: https://values.snap.com/policy/policy-community-guidelines/harmful-false-deceptive-information
[110] Generative AI on Snapchat. Available at: https://help.snapchat.com/hc/en-us/articles/25494876770580-Generative-AI-on-Snapchat

due to a lack of standards and interoperability in detection and labelling processes. This would lead to confusion and may weaken the effectiveness of ongoing efforts.[111]

Further, AI-altered media may not always be deceptive, and context-agnostic labelling can diminish the value of labels as well as content. For example, good-faith actors are likely to label their content without a standard, while malicious actors often do not. As a result, users may become more suspicious of labelled content, even when it is legitimate, and less suspicious of unlabelled content, even when it is harmful. This undermines users' ability to make informed decisions. The key distinction lies not in the technology itself, but in its intended purpose, whether creative and constructive, or deceptive and harmful. In essence, the binary identification of synthetic media may not always align with public perception.

Moreover, users may not perceive basic AI applications such as colour correction as manipulation, while opinions vary when it comes to edits like facial adjustments or the use of filters. There is no clear consensus on when AI-generated media is altered enough to justify a label, which leads to inconsistent labelling and user confusion. For example, if AI is used only to change the colour of an image, it remains unclear whether this modification requires a label or how it differs from traditional editing. In addition, labelling content as synthetic may unfairly suggest deception in cases such as clear parodies or artistic works, while labelling something as 'authentic' might wrongly imply trustworthiness even when manipulative framing is used to distort its meaning.

Further, effective large-scale risk mitigation strategies require interoperability operationalised through thoughtfully developed codes to achieve meaningful impact. Without consistent implementation across diverse Gen-AI models and social media platforms, these safeguards would fall short in establishing the identification and verification framework necessary for public trust.[112] Effectively implementing these techniques at scale requires coordination among diverse actors and systems, which often have conflicting priorities. For instance, hashing and filtering malicious synthetic media like CSAM demands collaboration between platforms, AI developers, ISPS, law enforcement, and advocates.[113]

Another challenge lies in determining what content qualifies for labelling. Current detection methods rely heavily on voluntary self-disclosures from creators, fact-checkers, crowd sourced assessments, and media forensics. Self-reporting mechanisms are often flawed, especially since malicious actors are the least likely to disclose AI use. The diversity of user bases could further complicate such flawed

111 Platforms fail to label and remove AI-generated and manipulated election content. Available at: https://www.isdglobal.org/digital_dispatches/platforms-fail-to-label-and-remove-ai-generated-and-manipulated-election-content/

112 Synthetic Content: Exploring the Risks, Technical Approaches, and Regulatory Responses. Available at: https://fpf.org/wp-content/uploads/2024/10/Synthetic_Content_Report_October_2024.pdf

113 Synthetic Content: Exploring the Risks, Technical Approaches, and Regulatory Responses. Available at: https://fpf.org/wp-content/uploads/2024/10/Synthetic_Content_Report_October_2024.pdf

labelling. Different cultural and linguistic contexts influence how terms like "artificial" or "synthetic" may be interpreted.[114]

In addition, major AI platforms are also increasingly developing principles, such as risk mitigation strategies in their AI models and applications.[115] However, there is currently little clarity in how these efforts are reported. The specifics of company initiatives vary widely, and there is no consensus on the reporting standard, making it difficult to assess their consistency or effectiveness. Moreover, the extent of meaningful scrutiny remains unclear, and it is challenging to determine how companies are putting their principles into practice, measuring the effectiveness of their risk mitigation strategies, or ensuring accountability.[116]

Platform	Risk Mitigation
Meta	Restricting access and delaying release for high-risk models until mitigations are in place, while suspending critical-risk models and enforcing security to prevent data theft.[117]
Microsoft	Integrates AI risk management into its enterprise risk framework while prioritising open-access tools for risk mapping and measurement.[118]
Google	Collaborates with external partners to establish safeguards, introducing frontier safety frameworks and Gen-AI toolkits to guide responsible development.[119]
OpenAI	Focuses on regulatory-informed research, particularly for assessing dangerous capabilities, with transparent red-teaming and safety protocols.[120]
Mistral	Implements strict prohibitions against illegal use, hate speech, misinformation, and CSAM, alongside robust data security measures.[121]
Anthropic	Deploys AI Safety Level standards to address catastrophic risks like misuse or unintended autonomous behaviour, requiring intensive pre-deployment testing.[122]

114 Labeling AI-Generated Content: Promises, Perils, and Future Directions. Available at: https://mit-genai.pubpub.org/pub/hu71se89/release/1
115 State of AI Governance, 2024. https://takshashila.org.in/research/state-of-ai-governance
116 State of AI Governance, 2024. https://takshashila.org.in/research/state-of-ai-governance
117 Frontier AI Framework. https://ai.meta.com/static-resource/meta-frontier-ai-framework/?utm_source=newsroom&utm_medium=web&utm_content=Frontier_AI_Framework_PDF&utm_campaign=Our_Approach_to_Frontier_AI_blog
118 Responsible AI at Microsoft. https://www.microsoft.com/en-us/ai/responsible-ai
119 Our AI Principles. https://ai.google/responsibility/principles/
120 Announcing OpenAI's Bug Bounty Program. https://openai.com/index/bug-bounty-program/
121 Legal terms and conditions. https://mistral.ai/terms#terms-of-service
122 Anthropic's Responsible Scaling Policy. https://www.anthropic.com/news/anthropics-responsible-scaling-policy

7. Detection and Labelling of Synthetic Media: Challenges of Accuracy, Trust, and Context

As synthetic media advances, regulatory approaches increasingly focus on detection and labelling mechanisms. While these are positive developments, they may fall short as standalone solutions due to fundamental limitations: inadequate diversity in incidental databases, user literacy levels, and resource requirements for system development. The potential unintended consequences are more concerning, including privacy violations and expanded surveillance capabilities. These effects likely disproportionately impact vulnerable populations, creating a tension between protection and harm. This chapter explores key mitigation strategies like detection, labelling and transparency for synthetic media harms and unintended consequences.

Identifying synthetic media is becoming increasingly complex. In response, stakeholders are developing different strategies to counter malicious synthetic media.[123] These efforts include both supply and demand-side measures, like media literacy, fact-checking initiatives, detection tools[124], and regulatory interventions.[125] Tech companies are also developing tools to counter malicious synthetic media. Microsoft's Video Authenticator, for example, targets synthetic media using public training datasets, while Meta has launched the collaborative 'Deepfake Detection Challenge' to accelerate innovation through industry-academic partnerships.[126] Similarly, Coalition for Content Provenance and Authenticity (C2PA) is a joint initiative by Adobe, Arm, Intel, Microsoft, and Truepic to develop open standards for certifying the origin and history of media content.

Detection and labelling are a broad term that constitutes tools and methods used to determine whether a piece of content is synthetic.[127] Despite being seen as a solution, these tools have yet to prove effective in addressing the risks posed by synthetic media. For instance, detection technologies for synthetic media often provide authenticity percentages rather than a definitive judgement, such as indicating the likelihood that

123 The Era of AI-Generated Election Campaigning is Underway in India. Available at: https://www.techpolicy.press/the-era-of-aigenerated-election-campaigning-is-underway-in-india/
124 AI Startups and the Fight Against Mis/Disinformation Online: An Update. Available at: https://www.gmfus.org/news/ai-startups-and-fight-against-misdisinformation-online-update
125 AI Startups and the Fight Against Mis/Disinformation Online: An Update. Available at: https://www.gmfus.org/news/ai-startups-and-fight-against-misdisinformation-online-update
126 https://www.chase-india.com/media/faengjxy/ai-and-content-moderation.pdf
127 Synthetic Content: Exploring the Risks, Technical Approaches, and Regulatory Responses. Available at: https://fpf.org/wp-content/uploads/2024/10/Synthetic_Content_Report_October_2024.pdf

a video is authentic and not manipulated. This probabilistic nature of AI raises questions across different fields.[128] They may not reveal specific manipulations or their relevance within the content's context, further complicating the assessment process.

For instance, it would be difficult for media organisations and legal systems to determine acceptable thresholds for considering content reliable.[129] Stakeholders suggested that these thresholds can vary by context. Public broadcasters may follow stricter standards than private platforms. As a result, deciding where to set the threshold for actions like labelling remains crucial and could impact the effectiveness of regulatory strategies.[130] Currently, sensitive investigations rely on labs and forensic tools to verify authenticity, which may not be sustainable given the scale at which synthetic content would be produced.

Detection software analyses signs of manipulation and supports reviewers by generating explainable reports. It can detect lighting inconsistencies between foreground and background elements to identify potentially fake content. However, while such labelling can be helpful, it also carries the risk of misclassifying authentic content, especially if it was captured under unusual lighting conditions, thereby unintentionally promoting distrust in authentic media.[131] Stakeholders underlined that larger platforms may have the resources to develop robust labelling systems. However, smaller platforms, face significant challenges due to limited investment and capacity. These platforms often rely on user declarations for content authenticity.

In addition, GANS used to create synthetic media can be modified to evade detection systems. When detection models learn to flag specific patterns in synthetic media, GAN can adjust to stop producing those detectable features.[132] This adaptive process allows GANS to generate more sophisticated synthetic media that can evade detection tools.[133] AI-created audio files are even more complicated because fake audio usually lacks a clear source, unlike manipulated videos, where the original footage can often be traced and compared. Without an original reference, it becomes much harder to verify authenticity or prove manipulation.[134]

---

128 AI Detectors Don't Work. Here's What to Do Instead. Available at: https://mitsloanedtech.mit.edu/ai/teach/ai-detectors-dont-work/

129 Deepfake video detection: challenges and opportunities. Available at: https://link.springer.com/article/10.1007/s10462-024-10810-6

130 Deepfake video detection: challenges and opportunities. Available at: https://link.springer.com/article/10.1007/s10462-024-10810-6

131 The DeepFake Detection Challenge (DFDC) Dataset. Available at: https://arxiv.org/abs/2006.07397

132 GANs are a type of AI model that consists of two neural networks—a generator and a discriminator—that work against each other. The generator creates synthetic data (like images or audio), while the discriminator evaluates how real or fake the data is. Through this adversarial process, GANs learn to produce highly realistic content.

133 Facing reality? Law enforcement and the challenge of deepfakes. Available at: https://www.europol.europa.eu/cms/sites/default/files/documents/Europol_Innovation_Lab_Facing_Reality_Law_Enforcement_And_The_Challenge_Of_Deepfakes.pdf

134 Indian Voters Are Being Bombarded With Millions of Deepfakes. Political Candidates Approve. Available at: https://www.wired.com/story/indian-elections-ai-deepfakes/

Detection models are often trained on databases of existing synthetic media to identify manipulation patterns.[135] This reliance on existing data creates challenges for the models' effectiveness against synthetic media generated by updated technologies.[136] The absence or fragmentation of existing synthetic media datasets poses a challenge for developing reliable detection methods.[137] Lack of robust data explains why such detection tools and classifiers often prove unreliable and inaccurate, frequently over- or under-detecting synthetic content. One study found that such tools can be biased against non-native English speakers, who are more likely to have their submissions detected as synthetic media due to training on inconsistent datasets.[138]

India's linguistic and geographical diversity presents a significant challenge in identifying malicious synthetic media. While news platforms can monitor misinformation in widely spoken languages like Hindi or English, they often lack the resources across all official languages and multiple regional languages spoken nationwide. Further, media platforms generally rely on forensic teams to confidently identify suspicious media. This has worked so far due to the difficulty of producing fabricated media at scale. However, with generative AI enabling rapid and large-scale creation, forensic methods may struggle to keep up. Therefore, regulations to address synthetic media harms must account for these complexities.

As AI becomes more sophisticated, identifying synthetic media is becoming increasingly difficult.[139] For example, AI might alter the background or context of a statement while the core statement remains authentic, or vice versa. This raises questions about how detection systems should be implemented to help users make informed decisions about the media they encounter. At the heart of these issues are concerns about intent, fairness, and deception. As online users become more vulnerable to manipulation, it is essential to collaboratively develop transparent and effective mechanisms to mitigate these risks.[140] Relying on fallible tools can create a false sense of credibility for malicious synthetic media that lacks adequate labels.[141]

---

135 Deepfakes And Policy Considerations. Available at: https://www.dsci.in/files/content/knowledge-centre/2023/DSCI%20Exploratory%20Note%20on%20Deepfakes_0.pdf

136 The Creation and Detection of Deepfakes: A Survey. ACM Computing Surveys, Available at: https://doi.org/10.1145/3425780

137 A survey on deepfake video detection datasets. Available at: https://www.researchgate.net/profile/Md-Fahimuzzman-Sohan/publication/374142887_A_survey_on_deepfake_video_detection_datasets/links/650f9297c05e6d1b1c2af258/A-survey-on-deepfake-video-detection-datasets.pdf

138 GPT detectors are biassed against non-native English writers. Available at: https://arxiv.org/abs/2304.02819

139 Understanding the Impact of AI-Generated Deepfakes on Public Opinion, Political Discourse, and Personal Security in Social Media. Available at: https://www.computer.org/csdl/magazine/sp/2024/04/10552098/1XApkaTs5l6

140 AI or Your Lying Eyes: Some Shortcomings of Artificially Intelligent Deepfake Detectors. Available at: https://link.springer.com/article/10.1007/s13347-024-00700-8

141 MULTITuDE: Large-Scale Multilingual Machine-Generated Text Detection Benchmark. Available at: https://aclanthology.org/2023.emnlp-main.616/

The increased scepticism around media may fuel the 'liar's dividend' where genuine media are dismissed as synthetic to evade accountability. This can destabilise the information system and threaten the notion of shared factual reality. The case of BJP MP Dinesh Lal Yadav "Nirahua" shows this issue clearly. After a video emerged showing him claiming that Prime Minister Modi and Chief Minister Adityanath addressed unemployment by not having children, while blaming those with multiple children for the problem, Yadav immediately dismissed the footage as AI-manipulated. "Anyone who sees it will understand that the lips are saying something else. What do you want to prove by cloning the sound with AI?" he argued. However, the complete interview remained on the channel, with the statement visible at 11:27.[142]

Tensions Between Identification and Privacy in Synthetic Media

Techniques such as watermarking, provenance tracking, metadata recording, and content hashing would likely to play a critical role in moderating synthetic media.[143] Watermarking for example, involves embedding information within content to confirm its authenticity and identify its source. Similarly, provenance tracking and metadata recording help trace content's origin and history by capturing higher-level information that supports authentication. Additionally, generating unique hashes of harmful content enables platforms to detect and filter such material from datasets, enhancing efforts to mitigate misuse while maintaining content integrity.[144] These techniques serve similar objectives by aiming to trace the sources of content.

Approach	Description
Watermarking	Embedding information within content to verify its authenticity and identify its source.
Provenance tracking	Tracking the provenance of content to verify its origin, history, authenticity, or quality.
Metadata recording	Used to authenticate the origin and history of content.
Hashing	Generating unique hashes of content to detect and filter it from datasets or platforms.

[142] BJP MP Nirahua said, 'Modiji & Yogiji curbed unemployment by not having children.' The video is NOT a deepfake. Available at: https://www.altnews.in/bjp-mp-nirahua-said-modiji-yogiji-curbed-unemployment-by-not-having-children-the-video-is-not-a-deepfake/

[143] Synthetic Content: Exploring the Risks, Technical Approaches, and Regulatory Responses. Available at: https://fpf.org/wp-content/uploads/2024/10/Synthetic_Content_Report_October_2024.pdf

[144] AI watermarking. Available at: https://www.techtarget.com/searchenterpriseai/definition/AI-watermarking

However, embedding them in synthetic media may raise concerns about the potential exposure of sensitive information, particularly when tracking is involved.[145] For example, a human rights advocate using Gen-AI to expose abuse of government power may face risks if an embedded watermark reveals their identity under an oppressive regime.[146] This contradicts the privacy-by-design principles.[147] Moreover, online anonymity is essential for multiple reasons and should not be equated with malicious activity. For example, anonymity provides a safe space for survivors of domestic violence, sexual abuse, or LGBTQ+ individuals in hostile environments to seek support and share experiences without exposing their identities.[148]

By monitoring everyone, the regulatory mechanism risks treating all users as potential offenders, shifting the burden of proof onto individuals. Assigning identifiable and unique identities does not always ensure accountability, and anonymity does not imply non-compliance with the law. These risks vary by use case, requiring tailored safeguards. In sensitive contexts like healthcare, techniques such as watermarking or metadata tracking may reveal patient identities if not properly de-identified, risks that differ significantly from those in social media or gaming.[149]

Techniques required for the detection of synthetic media often introduce privacy concerns. For instance, solutions to AI-generated scam calls include real-time detection tools that would alert recipients to potential AI impersonation.[150] While such initiatives could protect individuals from AI-driven scams, they may necessarily require real-time collection and analysis of personal, potentially sensitive conversations.[151] Third-party monitoring of private communications between individuals creates privacy risks. Similarly, while the traceability of synthetic media can help prevent emerging harms, it will continue to pose challenges for privacy, including its effectiveness when applied to encrypted platforms.

---

145 Detecting AI fingerprints: A guide to watermarking and beyond. Available at: https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/

146 AI Watermarking: How It Works, Applications, Challenges. Available at: https://www.datacamp.com/blog/ai-watermarking

147 Data Empowerment And Protection Architecture. Available at: https://www.niti.gov.in/sites/default/files/2023-03/Data-Empowerment-and-Protection-Architecture-A-Secure-Consent-Based.pdf

148 Why Online Anonymity is Critical for Women. Available at: https://womensmediacenter.com/speech-project/why-online-anonymity-is-critical-for-women

149 Comment on FR Doc # 2024-09824. Available at: https://www.regulations.gov/comment/NIST-2024-0001-0027

150 Comments Of Microsoft Corporation. Available at: https://www.fcc.gov/ecfs/document/1219842001792/1

151 Implications of Artificial Intelligence Technologies on Protecting Consumers from Unwanted Robocalls and Robotexts. Available at: https://docs.fcc.gov/public/attachments/FCC-24-84A1.pdf

The Limits of Labelling: Deception, Literacy, and Trust

Labelling has emerged as a commonly proposed strategy for mitigating the risks of synthetic media.[152] However, stakeholders noted that the effectiveness of such labelling remains uncertain, as even the labelling process itself can be deceptive, as observed in various industries. For example, Hindustan Unilever Limited, which owns brands like Horlicks and Boost, has rebranded its 'health drinks' category. It renamed the functional nutritional drinks segment and removed the 'health' label from Horlicks. This change follows a directive from the Minister of Commerce and Industry, asking e-commerce platforms to remove drinks and beverages from the 'healthy drinks' category. Similar deception tricks can be used in synthetic media.

While labelling can raise awareness, it is not a comprehensive solution. The effectiveness of measures like labelling synthetic media depends on users' ability to interpret those labels and recognise manipulated media, skills that are often limited, particularly in the Indian context. This is because users need to understand how these tools work, which requires increased investment in digital literacy and public education.[153] At the same time, when systems for labelling misinformation are implemented, any misinformation that remains unlabelled tends to be perceived as more accurate. This phenomenon is known as the "implied truth effect".[154]

Meanwhile, legal recourse for victims can be convoluted due to the challenge of identifying the creator of synthetic content, mainly when they operate anonymously. The core issue is attribution. Thus, even if legal provisions provide compensation and reparation for harm caused by synthetic media in several jurisdictions like the USA, enforcement would continue to be a challenge. Moreover, the metadata needed to trace the origin of synthetic content may be insufficient for identification. This problem persists when the synthetic content is shared by someone other than the creator on social media. Technologies like Tor, which enable anonymous content sharing, make it even more challenging to identify the individual responsible.[155]

---

[152] Social media networks have faced criticism for inadequate efforts to combat disinformation. However, the social media platform Meta will begin applying "Made with AI" labels in May to AI-generated videos, images, and audio shared on its platforms, expanding its policy beyond the limited scope of previously addressed doctored videos. Additionally, it will use separate, more prominent labels for digitally altered media that poses a "particularly high risk of materially deceiving the public on important matters," regardless of whether AI or other tools were used to create it. While it is too early to assess this approach's effectiveness, it opens many questions i.e., to what extent will labelling AI-generated content prevent creation and spread of the malicious synthetic content. Available at: https://www.reuters.com/technology/cybersecurity/meta-overhauls-rules-deepfakes-other-altered-media-2024-04-05/

[153] Comment on FR Doc # 2024-09824. Available at: https://www.regulations.gov/comment/NIST-2024-0001-0046

[154] The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. Available at: https://doi.org/10.1287/mnsc.2019.3478

[155] Tore. Available at: https://www.torproject.org/

AI-Generated Video of Ram Setu
A video circulating on social media claims to show scuba divers exploring the submerged ruins of the Ram Setu, a chain of limestone shoals between Rameshwaram in Tamil Nadu and Mannar Island in Sri Lanka, believed by many to hold religious value. The video shows divers examining inscribed stone boulders, tools, and sculptures resembling Hindu deities like Hanuman and Ram. At first glance, the clarity and visual richness of the footage led many to believe it was real. However, a fact-check revealed that the footage was created using deepfake technologies.[156] The original creator, who identifies as an AI artist, had posted the video with a clear disclaimer: "Note: All visuals shown here are AI-generated and meant purely for artistic and educational purposes. We deeply respect all beliefs and traditions; this reel is intended to inspire wonder, not offend." Despite this disclaimer, the video was widely reshared across social media platforms without any AI label or context, misleading viewers.

Even when content is clearly labelled as synthetic, it can still cause real harm. For instance, a woman featured in a non-consensual AI-generated pornographic video may still experience feelings of shame, sexual objectification, and distress, even if the video includes a disclaimer and is eventually removed. Labelling alone is insufficient because such media can be viewed or shared thousands of times before detection.

Careful consideration must also be given to how labels are displayed and what level of detail they should provide. Should a basic disclaimer suffice, or must the label explain which parts of the media were manipulated and how? Should the intended audience for the label be the public, the person depicted, or the platform hosting the synthetic media? How should the format and placement of the label be determined, and what level of detail should be disclosed? For example, is a simple disclaimer that the content is artificially generated sufficient, or should the label specify which elements have been manipulated and how? While labelling is a step forward, stakeholders noted these questions must be answered to make interventions more effective.

Stakeholder also noted that most content moderation is now automated, amplifying commercial interests and existing biases. AI systems can reinforce these structural issues or favour certain ideological views. For example, algorithms may keep recommending content aligned with a user's political preferences, filtering out opposing views and creating echo chambers that limit exposure to diverse

[156] Viral videos of scuba divers' 'Ram Setu exploration' are AI-generated. Available at: https://www.altnews.in/viral-videos-of-scuba-divers-ram-setu-exploration-are-ai-generated/

perspectives.[157] Recommendation algorithms are critical in determining the reach and impact of any content, including AI-generated media, as they dictate visibility. Malicious synthetic media may have little effect without a social media platform's algorithm recommendation. Platforms could manage visibility through transparent shadow banning on fairly labelled synthetic media.

The Promise and Limits of AI Disclosure

Transparency is increasingly being used to inform users about synthetic media's potential risks and limitations. For example, the EU AI Act includes provisions for such transparency measures. This is intended to educate individuals that content is synthetic or that they are interacting with a Gen-AI system. These may consist of content labels, notices during interactions with AI chatbots, or documentation such as model cards and system cards detailing the AI system used, its training data, performance, and capabilities. In principle, this is a good idea in maintaining the integrity of public discourse and preventing the inadvertent spread of misinformation.

However, the efficacy of such measures must be carefully considered within the broader context of information overload in the digital age. Studies reveal that almost no one reads online contracts, terms of service, or privacy policies. A Deloitte survey found that 91 percent of consumers agree to terms without reading them, with 97 percent among those aged 18 to 34. Even if someone tried to read all notices and disclosure documents, it would take about 76 workdays to complete.[158]

Moreover, information overload can have detrimental effects on decision-making quality, user convenience, and overall satisfaction with digital services. The cognitive burden imposed by excessive information can lead to decision fatigue, potentially compromising the user's ability to make informed choices about their content. Implementing frequent reminders, while well-intentioned, may paradoxically contribute to this information overload. Users, overwhelmed by the constant barrage of notifications and alerts, may develop a tendency to perfunctorily agree to terms and conditions without fully comprehending their implications.[159]

Furthermore, while notices are intended to inform users, they are often seen as interruptions to seamless digital interaction and may negatively affect the overall user experience. Many users engage with such notices superficially, typically clicking "accept" without fully understanding their implications, remarked stakeholders. This

[157] Content Moderation in a New Era for AI and Automation. Available at: https://www.oversightboard.com/news/content-moderation-in-a-new-era-for-ai-and-automation/

[158] You're not alone, no one reads terms of service agreements. Available at: https://www.businessinsider.in/tech/youre-not-alone-no-one-reads-terms-of-service-agreements/articleshow/61659553.cms

[159] Enlarging the Small Print: A Study on Designing Effective Privacy Notices for Mobile Applications. Available at: https://cis-india.org/internet-governance/blog/enlarging-the-small-print

highlights the need for sustained consumer education efforts. Awareness campaigns focusing on misinformation and AI-generated content are essential and should complement, rather than be replaced by, legal notices or disclosure mechanisms. However, these challenges are compounded by deeper structural gaps in the governance of synthetic media.

8. Uncovering Structural Gaps in Synthetic Media Regulation

Following the discussion on detection and labelling, this chapter explores the structural challenges of moderating AI-generated media. It considers how a focus on labelling may not sufficiently address the broader conditions that enable harmful uses of Gen-AI. These include concerns around data rights, privacy, consent, and the complexities of cross-border content moderation, influencing how synthetic media is produced and disseminated.

Content moderation challenges are set to intensify with the rise of generative AI. Stakeholders noted that regulating synthetic media in isolation tackles only the symptoms, not the root causes. They underscored that a more helpful approach would be to focus on core issues like data rights, privacy, content moderation, and consent, which may allow the misuse of AI tools. For example, a stakeholder noted that clear data ownership laws could give individuals greater control over their data. Generative AI models often scrape diverse datasets such as text, images, videos, and audio to generate synthetic media. Clear ownership rights would discourage scraping public data, curbing synthetic media at its source.

India's Digital Personal Data Protection Act does not define data ownership. Instead, it encourages AI systems to scrape and use it freely for training. Without clear ownership of data, privacy and data protection rights may have limited effectiveness in combating harms related to synthetic media, noted stakeholders. Data ownership is not tied to identity, and for digital footprints, images, videos, and text created by user actions online. The Act only protect data deemed private. For AI platforms, both personal and non-personal data are crucial. These needs must be balanced with the requirement for such data to support AI innovation, particularly in a country like India, which is striving to develop indigenous AI models.[160]

Synthetic media can violate privacy and reputation by creating explicit or damaging media using an individual's likeness without consent. This misuse may begin with unauthorised access or by using publicly available data, such as personal photos or videos, constituting a privacy invasion. The concerns extend beyond initial collection to include secondary use of personal information, where individuals lose control over

[160] Centre selects start-up Sarvam to build country's first homegrown AI model. Available at: https://indianexpress.com/article/business/centre-selects-start-up-sarvam-to-build-country-first-homegrown-ai-model-9967243/

how synthetic media creators and platforms manipulate and distribute their data. For instance, actress Rashmika Mandanna was targeted using synthetic media that digitally altered her face without consent. "It's terrifying, not just for me, but for all of us, as we're all vulnerable to such harm today," Rashmika wrote in a tweet.[161]

In addition to privacy, the risks associated with synthetic media constitutes multiple dimensions, including defamation and violations of personality rights. These rights generally allow individuals, including public figures, to control the use of their name, image, likeness, or identity, particularly in relation to their reputation and commercial exploitation. Personality rights are recognised as legal rights across various jurisdictions. When Gen-AI superimposes a person's face onto another body or creates synthetic media featuring their likeness, it may infringe upon these rights, resulting in actionable harm even if the content is not explicitly sexual or defamatory.

Stakeholders also noted that current responses to malicious synthetic media like copyright claims, personality rights or as a financial crime involving transactions or digital impersonation, are largely reactive. They address reputational, economic, or societal harm after it has already occurred. They do not prevent or reduce the creation of malicious synthetic media or deter AI platforms that facilitate its generation.

In India, for example, when dealing with deceased individuals, the legal approach typically involves obtaining copyright through a power of attorney or residual rights from the family or representatives, such as a widow. Even with family permission, this remains a serious issue, as it affects the dignity and autonomy of the deceased. For instance, AI-generated videos of deceased Indian politicians have been used to issue posthumous election endorsements. In 2024, Wired reported that over 50 million AI-generated voice clone calls were made in India in the two months before the elections, showing the rise of a deepfake-driven cottage industry for voter outreach.[162]

The concern is further compounded by the fact that deceased individuals cannot respond to or be held accountable for such content, and audiences may be unaware of the individual's death, potentially leading to confusion or the spread of misleading information. To further complicate matters, AI developers frequently use this content without securing permission from rights holders, leading to increasing calls for greater transparency about what data is used to train AI models. Moreover, stakeholders noted that regulating synthetic content is challenging because it targets specific AI applications rather than addressing the broader technology ecosystem.

---

161 Spectre of deepfakes looms large as fake video of Rashmika Mandanna goes viral. Available at: https://www.thehindubusinessline.com/info-tech/spectre-of-deepfakes-looms-large-as-fake-video-of-rashmika-mandanna-goes-viral/article67508628.ece

162 Indian Voters Are Being Bombarded With Millions of Deepfakes. Political Candidates Approve. Available at: https://www.wired.com/story/indian-elections-ai-deepfakes/

Stakeholders noted that frameworks like the IT Act or BNS may be inadequate for addressing AI-generated media concerns.[163] They require court filings that fail to adequately address the harm to reputation and dignity. These legal procedures face practical challenges: identifying uploaders can be difficult, and the time between initiating legal proceedings and obtaining court orders is often too long when synthetic media spreads virally. These mismatch between legal remedies and synthetic media harms comes from the technology's inherent characteristics, remarkable speed and amplification capabilities that allow malicious content to achieve widespread damage before legal systems can respond.[164]

The gap underscores the need for regulatory refinement that considers the specific characteristics of synthetic media, rather than relying on traditional legal mechanisms built for slower-moving forms of content. For example, a synthetic media piece targeting a political leader could be created in one country, go viral in another, and be hosted on servers in a third, delaying removal and amplifying harm. Content moderation across jurisdictions is already complex, and by the time takedown requests are processed, content may have been widely circulated. The effectiveness of these requests also varies significantly between countries. These challenges highlight the need for a shift in governance strategies.

[163] India's Advance on AI Regulation. Available at: https://carnegieendowment.org/research/2024/11/indias-advance-on-ai-regulation?lang=en&center=india

[164] Regulating Deepfakes: An Indian perspective. https://digitalcommons.usf.edu/cgi/viewcontent.cgi?article=2245&context=jss

9. Safe Harbour in Question: Synthetic Media and Legal Ambiguities

Building upon the previous chapter, this chapter explores intermediary liability in the Indian context. It considers how existing legal frameworks, particularly the safe harbour, under Information Technology (IT) Act, 2000 and its judicial interpretations, engage with the emerging challenges of AI-generated media. The discussion raises questions around attribution and responsibility, including the roles of users, social media platforms, AI developers and deployers. Based on case examples and stakeholder perspectives, the chapter examines how current laws may contend with synthetic media's speed, scale, and decentralised nature.

Moderating synthetic media might have implications for intermediaries. The latest advisory requires intermediaries to take swift action against malicious synthetic media, mandating their removal within 36 hours of receiving a report from either a user or a government authority, as specified in the IT Rules. Failure to do so can lead to the loss of the immunity provided by Section 79(1) of the IT Act.[165] This could expose them to legal action or prosecution under various laws, including the IT Act and the BNS, particularly Section 469.[166]

The Bombay High Court noted that requiring intermediaries to ensure content accuracy effectively censors user-generated content by shifting responsibility from the creator to the intermediary, which has no control over it. The court observed that even if content is identified as fake or misleading, complete removal is unlikely due to the replicable nature of digital content. Thus, intermediaries risk losing safe harbour protection despite making 'reasonable efforts.'

To contextualise the discussion, Section 79(3) of the IT Act holds intermediaries liable in two cases: (a) if they conspire, abet, aid, or induce the commission of an unlawful act, or (b) if they fail to restrict access to data used for unlawful acts promptly. However, these conditions require substantial proof from the injured party, making it easier for social media platforms to evade compliance.[167]

---

165 The Information Technology Act, 2000. Available at: https://www.meity.gov.in/content/information-technology-act-2000

166 Forgery for the purpose of harming reputation. Available at: https://devgan.in/ipc/section/469/#:~:text=Whoever%20commits%20forgery%2C%20intending%20that,also%20be%20liable%20to%20fine.

167 The Information Technology Act, 2000. Available at: https://www.meity.gov.in/content/information-technology-act-2000

Failure to block malicious synthetic media when directed by a competent authority can result in imprisonment for up to seven years, a fine, or both under the IT Act. In cases where synthetic media threatens national sovereignty,[168] intermediaries could face charges, including cyber terrorism or waging war against India.[169] Intermediaries may be penalised for publishing obscene material in electronic form. They may also face charges under provisions for criminal breach of trust, criminal conspiracy, and cheating under the BNS, depending on the case's specific circumstances.

The above discussion underscores that measures against malicious synthetic media must be carefully crafted to avoid unintended consequences for users and platforms. Without a clear direction, a broad range of online intermediaries could face legal risks. Addressing this challenge requires two things: first, determining whether the content is synthetic, and second, assessing whether it is problematic. This differentiation could lead to overly cautious approaches by platforms seeking to avoid liability. As a result, they might implement aggressive content filtering measures or, in extreme cases, cease supporting user-generated content altogether to avoid getting into trouble by segregating synthetic from human-generated content.

Moreover, given the potential consequences of placing the responsibility of determining permissible speech on private entities, such an approach risks contravening the Supreme Court's guidance in Shreya Singhal vs Union of India. It may lead to the over-removal of legitimate content due to risk-averse moderation practices, while also discouraging the growth of user-generated content platforms—thereby narrowing the space for online expression, creativity, and democratic participation. The Supreme Court narrowed the scope of Section 79 and its Rules, ruling that online intermediaries are required to remove content only upon receiving an order from a court or government authority.[170]

The distinction between human and synthetic media is crucial for determining liability when content is deemed violative. A key question arises: Should a company developing or deploying an AI-powered chatbot be held responsible for the content it generates, or does liability lie with the user who requested it? This issue becomes particularly complex in cases where LLMs generate potentially defamatory content. Under the current safe harbour provisions of the IT Act[171], platforms displaying third-party content are not considered publishers, and liability typically rests with the users who post the content. This framework protects online services that host user-generated

168 The scope of terms such as national sovereignty, national security, and public order is not defined in the law.
169 Condemned to Be Free: Balancing Free Speech and Security Online. Available at: https://www.orfonline.org/expert-speak/condemned-to-be-free-balancing-free-speech-and-security-online
170 Shreya Singhal v. Union of India. Available at: https://globalfreedomofexpression.columbia.edu/cases/shreya-singhal-v-union-of-india/
171 The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021. Available at: https://www.meity.gov.in/content/information-technology-intermediary-guidelines-and-digital-media-ethics-code-rules-2021

content. However, the balance of responsibility between the user and the AI developer must be carefully assessed when it comes to synthetic media.

One argument is that AI-generated media should be covered under safe harbour provisions, as it is driven by users' prompts and, based on the training datasets, which are being scraped from publicly available digital footprints. Such a perspective places responsibility entirely on the users rather than the AI developers or deployers. However, unlike social media platforms, which are considered neutral hosts of user-generated content, Gen-AI generates content from trained datasets. Because Gen-AI generates content rather than merely hosting or suggesting it, it may not be viewed as a neutral entity. In this case, the AI developer should bear responsibility for ensuring due diligence for the content it produces.

Moreover, social platforms have simplified content creation by incorporating generative AI tools into their services. A key question is whether Gen-AI tools must comply with the restrictions under the IT Rules, 2021. Rule 3(1)(b), for instance, requires intermediaries to make reasonable efforts to prevent users from sharing prohibited content. When Gen-AI is developed by a social media platform's parent company and deployed on that platform, does it become part of the intermediary service? If so, is it obligated to reject queries that violate the IT Rules or platform guidelines? The issue becomes more complex when a Gen-AI tool, developed by a third party, is integrated into the platform. Would its liability and compliance obligations differ, since the same platform does not own it?

If Gen-AI tools create malicious synthetic media, regulations must clearly define who is liable. One of Germany's Federal Court decisions can guide this discussion. The court found that Google's autocomplete search suggestions can constitute defamation when they generate false implications about an individual. The court ruled that search suggestions linking a person's name to terms like "fraud" or "Scientology" without evidence can cause reputational damage. The court said Google must block libellous terms associated with a name entered in its search bar if alerted. The case was filed by an unnamed food supplements company and its CEO, identified as R.S. Upon entering their name into the search bar, the auto-complete function suggested results including the terms "scientology" and "fraud" without any basis for such linkage.[172]

By holding Google accountable for automatic search prompts, the court recognised that seemingly neutral technological features can nonetheless cause personal harm.[173] In Gen-AI, users provide input, but the AI generates content without understanding

[172] German court orders Google to remove auto-complete entries. Available at: https://techmonitor.ai/technology/german-court-orders-google-to-remove-auto-complete-entries-150513

[173] Google loses case in Germany over autocomplete search suggestions. Available at: https://www.cnet.com/tech/services-and-software/google-loses-case-in-germany-over-autocomplete-search-suggestions/

its implications. To address this, platforms have begun implementing safeguards to limit harmful outputs. For example, ChatGPT includes content moderation systems that flag sensitive topics such as violence or abuse.[174] Similarly, it also restricts users from generating images of celebrities to avoid violating their personality rights. While AI cannot bear liability due to its lack of moral understanding, stakeholders argued that Gen-AI developers and providers should be held accountable for their outputs when they fail to ensure transparency and/ or integrate necessary safeguards.

Stakeholders emphasised that liability should arise if malicious content is produced due to negligence in the design or development of their models. Liability for AI-generated content involves both civil and criminal dimensions, addressing distinct harms. Civil liability may cover harms related to intellectual property violations, consumer deception, or misleading advertisements. For example, generating synthetic images mimicking a brand could lead to trademark infringement or false advertising claims. They also noted that criminal liability, by contrast, pertains to offences like cybercrimes, identity theft, or malicious deepfakes, which can cause harm to individuals. Civil remedies can focus on compensation and restitution for affected parties, while criminal penalties address intentional harm and serve as deterrents.

[174] Usage Policy Content Violation. Available at: https://community.openai.com/t/usage-policy-content-violation/974993

10. Conclusion and Recommendations

The growing sophistication of synthetic media raises critical concerns about its widespread proliferation and the resulting impact on individuals, society, politics, and economic systems. Addressing these issues is complex and requires a multi-layered approach. Discussions focus on implementing labelling, requiring social media platforms to moderate malicious synthetic media, raising public awareness about synthetic media, and establishing grievance redressal and reporting mechanisms.

Many jurisdictions are making progress by enacting laws that criminalise the creation and/or distribution of synthetic media with malicious intent and providing recourse for those affected. However, it is crucial that these laws are carefully crafted to avoid infringing on legitimate speech rights and online anonymity and should not negatively impact the innovations in the AI space.

Moreover, the right to freedom of expression extends protection to seeking, receiving, and imparting information across all media, including AI. This protection constitutes not only welcome or innocuous ideas but also those that might offend, shock, or disturb, a principle crucial for maintaining the pluralism, tolerance, and openness that societies require. Regulatory approaches driven by panic rather than evidence tend to obscure the potential of AI technologies to enrich human communication, nurture creative expression, and contribute meaningfully to expanding human knowledge.

As discussions progress, factors like developing detection tools and establishing clear accountability frameworks will likely shape long-term policy decisions. Regulation must address multiple moving elements, including the role of developers, deployers, and users, while considering the effectiveness of technological solutions.

The above discussion makes it clear that there is no simple solution to combating the misuse of synthetic media. The knee-jerk response could prove prohibitively expensive for small to medium-sized platforms, effectively restricting participation to large corporations with the necessary resources. This scenario risks increasing market concentration, a trend in other sectors with similar regulatory measures.

Reactionary policies risk stifling innovation and penalising legitimate use cases. Effective regulation should be grounded in evidence and precise articulation. Without these elements, policies risk undermining freedom of expression, creativity, and satire,

essential components of healthy discourse. The challenge is to craft proportionate measures that guard against real and evolving threats without stifling AI innovation or lawful expression. The real thrill lies not just in regulating AI, but in doing so thoughtfully and effectively.

Recommendations

Reimagining Liability: Synthetic media governance must be context-specific, as the legality and appropriateness of AI-generated content often depend on how and where it is used. A precedent from one case cannot be indiscriminately applied across all situations. For example, a teacher using Gen-AI to create synthetic media for an educational history lesson involving hate speech might not be acting unlawfully, given the context and intent. However, if someone uses Gen-AI to disseminate speech that incites violence publicly, the implications are far more serious. A risk-based regulatory framework would calibrate oversight intensity proportional to potential impact. This requires developing evidence-grounded risk classifications and harm principles through collaborative processes.

Based on collaboratively developed risk classification and harm principles, codes and standards should be created and made mandatory for integration across the AI system lifecycle. For example, AI systems should incorporate safety codes, security standards and privacy-by-design principles as non-negotiable baseline requirements, regardless of their apparent risk profile. Developers should implement progressively stronger protective measures as potential harm increases across dimensions such as scope, severity, and probability. This risk-calibrated approach ensures proportional safeguards: high-risk applications require comprehensive guardrails, including rigorous testing and transparent accountability mechanisms, while excessive constraints do not burden lower-risk innovations.

Liability should trigger upon non-compliance with foundational and subsequent necessary safeguards. This is important because one-third of Gen-AI tools enable pornographic content creation, with technology advancing to produce 60-second synthetic videos from a single image in under half an hour.[175] This requires training datasets to be meticulously curated to exclude harmful content such as child sexual abuse material throughout the development lifecycle. Compliance monitoring should fall to independent oversight bodies comprising government officials, civil society representatives, academics, and subject matter experts. These bodies would produce annual reports that would be available for public scrutiny. Such oversight mechanisms would enhance transparency, build user trust, and provide a more precise understanding of generative AI's capabilities.

Building A Collaborative Framework: The proposed AI Safety Institute of India (AISI)[176], under MeitY, could play a key role in facilitating an iterative, collaborative process involving civil society, academics, and subject matter experts to set

[175] State Of Deepfakes. Available at: https://www.securityhero.io/state-of-deepfakes/

[176] India's Artificial Intelligence Safety Institute to work virtually across institutions, officials say. Available at: https://www.thehindu.com/sci-tech/technology/indias-artificial-intelligence-safety-institute-to-work-virtually-across-institution-officials-say/article69191949.ece

benchmarks, establish transparency standards, and guide AI actors on complying with co-developed codes. By designing and implementing collaboratively developed codes and standards, stakeholders can create a practical framework that promotes accountability, protects individual rights, and mitigates risks.

Relying only on voluntary commitments creates critical problems, failing to establish binding obligations and leaving AI systems' operations opaque. The AISI could partner with social media platforms, Gen-AI firms, and grassroots consumer groups to improve synthetic media detection and labelling. Such collaboration would leverage each stakeholder's strengths without imposing rigid rules that negatively impact innovation. Insights from these efforts could guide a more nuanced regulatory framework that protects users while supporting technological progress.

Given the diverse applications of synthetic media, from entertainment to political misinformation, clearly defining permissible and impermissible uses is vital. The evolving definition of synthetic media, now encompassing voice cloning, full-body manipulation, and text-to-image generation, presents challenges for effective regulation. Without clear boundaries, there is a risk of stifling legitimate creative expression and inadequately addressing harmful uses, leading to regulatory inconsistencies across jurisdictions.

Restricting Non-Consensual Synthetic Media: The creation, distribution, and hosting of non-consensual synthetic media should be strictly prohibited unless the individual depicted has given explicit, informed consent. To effectively curb such practices, penalties should be established based on the harm caused and the creator's intent. Individuals who deliberately share harmful media should be subject to legal action. Social media platforms should be required to verify consent for synthetic media at the time of upload for all the people. Platforms that fail to prevent the spread of non-consensual media should be held liable, especially since their algorithms can detect and limiting the reach of malicious synthetic media.

The framework for penalties should recognise the power imbalances between platforms and users. Civil society should be legally permitted to support and represent affected individuals in claiming and receiving adequate compensation or reparation. Policymakers should also promote specialised risk mitigation mechanisms and insurance pools for AI systems. These market-based solutions would establish dedicated financial reserves for AI-related claims where developers and deployers create shared reserves. Such mechanisms distribute risk proportionally across the ecosystem and ensure compensation availability even when individual entities lack sufficient resources to address large-scale harms.

In addition, social media intermediaries and AI platforms should proactively educate policymakers and the public about synthetic media moderation, sharing information on media removal practices and providing clear explanations for their decisions. To ensure regulatory frameworks remain responsive to technological developments, policymakers should create structured feedback channels through which civil society insights directly inform policy evolution. This would help build a more informed digital ecosystem where all stakeholders understand the challenges and responsibilities of managing synthetic media harms. media

Empowering Users: Top-down interventions matter, but grassroots action is essential to closing the gaps. Stakeholders must prioritise reliable and accessible AI detection and verification tools to protect users from the impact of harmful synthetic media. MeitY is developing real-time synthetic media detection tools, watermarking, and labelling AI-generated media, among other things. Independence must be safeguarded through appropriate checks and balances, including regular testing and public transparency reports.[177]

A similar model is being developed by innovators like Itisaar, which, in partnership with Indian startup DigitID, allows users to upload files directly and automates parts of the verification process. Despite its recent launch, the platform already has over 80 approved users.[178] Rather than imposing government-mandated detection tools, India should facilitate market innovation of technical safeguards while strengthening independent fact-checking networks.

Drawing inspiration from Taiwan's "nerd immunity" approach, which integrates professional fact-checkers with public participation, India should develop a collaborative ecosystem where social media platforms incorporate fact-checking tools within their applications.[179] This would address malicious synthetic media without resorting to dipropionate regulations.[180] India already hosts similar models like the Deepfakes Analysis Unit, created by fact-checking organisations.

Government and industry support for such citizen-centred mechanisms would enhance nationwide synthetic media verification capabilities. The goal goes beyond identifying synthetic media to include mechanisms for recourse and remediation when harm occurs. The government should establish dedicated funding mechanisms to

177 MeitY seeks tools to detect deepfakes, label AI generated content. Available at: https://www.hindustantimes.com/india-news/meity-seeks-tools-to-detect-deepfakes-label-ai-generated-content-101734410291642.html

178 Itisaar - Deepfake Platform. Available at: https://www.behance.net/gallery/196394089/itisaar-Deepfake-Platform

179 Striking the balance: Navigating deepfakes and free speec. Available at: https://www.amsshardul.com/insight/striking-the-balance-navigating-deepfakes-and-free-speech/

180 Deep Div Deep Dive into Deepfak o Deepfakes—Safeguarding Our Digital Identity ding Our Digital Identity. Available at: https://brooklynworks.brooklaw.edu/cgi/viewcontent.cgi?article=2012&context=bjil

support civil society's capacity-building efforts, particularly in developing accessible educational resources that enhance public literacy around synthetic media.

CUTS International

D-217, Bhaskar Marg, Bani Park, Jaipur 302 016, India
Ph: 91.141.228 2821, Fax: 91.141.228 2485
Email: cuts1@cuts.org, Website: www.cuts-international.org

Also at Delhi, Kolkata and Chittorgarh (India); Lusaka (Zambia); Nairobi (Kenya); Accra (Ghana); Hanoi (Vietnam); Geneva (Switzerland) and Washington DC (USA).