



Centre for Aerospace Power and Strategic Studies



Centre for Aerospace Power and Strategic Studies

ISSUE BRIEF

12/26

08 October 2026

GENERATIVE AI, TRUTH, AND DEMOCRACY

Ms Gowri R

Research Associate, Centre for Aerospace Power and Strategic Studies



Introduction

Generative Artificial Intelligence (GenAI) has transformed contemporary digital information by reducing the time, skill, and cost required to create persuasive synthetic content. While synthetic media offers significant benefits for creativity and communication, it also enables the creation and dissemination of deepfakes and other forms of synthetic media for impersonation, fraud, reputational harm, election interference, and strategic information operations. The proliferation of false information through such content therefore presents both a technical and social challenge. As the content creation is reaching beyond the nominal human ability to validate the veracity of information, mechanisms for verification of AI-enabled synthesis have emerged.

The problem is not limited to identifying whether a particular image, video, or audio clip is authentic. The deeper challenge is the erosion of shared trust and factual basis for democratic deliberation. When citizens cannot reliably distinguish genuine evidence from fabricated material, public trust in institutions, journalism, elections, and even verifiable events can weaken. The resultant “liar’s dividend”¹ allows genuine material to be dismissed as artificial, while fabricated material can be circulated before verification mechanisms respond.

It is especially vulnerable for democratic societies like India, because their commitment to free expression and open information flow can be exploited to disseminate AI-

generated disinformation faster and more persuasively than fact-checking mechanisms can respond. In July 2026, the removal of posts related to India's political leadership² and student protests over examination-paper leaks³ has triggered concerns about freedom of expression and control of information over online social media platforms. This incident exemplifies the growing uncertainties over whether content restraints resulted from state or non-state intervention, platform moderation practices, automated systems, or other external pressures that shaped public discourse.

This brief examines how GenAI-enabled synthetic media challenges the authenticity of digital evidence. Further, it provides recommendations for the Indian context through the lens of existing regulatory responses in the pervasive information environment. It argues that India should respond through a layered whole-of-nation approach comprising technical provenance and forensic verification, independent and multilingual fact-checking, accountable platform governance, proportionate intermediary obligations, international cooperation, and transparent procedural safeguards.

Digital Authenticity and Verification

Synthetic content challenges conventional evidentiary assumptions. Historically, photographs, recorded conversations, and videos have carried substantial persuasive value because they tend to preserve direct traces of real-world events. In the AI era, such material cannot automatically be treated as conclusive proof. Authenticity must increasingly be established through corroboration, source provenance, metadata, trusted institutional verification, and contextual analysis. This growing need for verification is closely linked to the technological maturity of existing GenAI systems and common understanding of their capabilities.

Machine-synthesised or tampered multimedia leaves detectable artefacts.⁴ Original multimedia contains unique signatures due to the natural characteristics of human audio or imagery. Therefore, existing verification techniques need to be improved to identify these unique traits of online media content. Deepfake-forensics techniques can be strengthened with a combination of verifiable accounts of who produced it, when it was created, what it depicts and whether it has been materially altered. Techniques such as watermarking, source fingerprinting,

Deepfake-forensics techniques can be strengthened with a combination of verifiable accounts of who produced it, when it was created, what it depicts and whether it has been materially altered. Techniques such as watermarking, source fingerprinting, content credentials, and embedded metadata can strengthen this process.

content credentials, and embedded metadata can strengthen this process. Tech giants such as Google,⁵ Microsoft,⁶ and lately Anthropic⁷ are actively developing and implementing technical solutions for watermarking GenAI content. However, these techniques may be weakened through re-editing, format conversion, screen capture, or deliberate evasion, while content generated through non-participating systems may carry no provenance signal at all. For this reason, India should treat technical provenance as part of verification, while also considering the institutions and platforms required to verify and communicate findings to users.

India currently relies on fact-checking institutions to assess online content. Meta's shift from third-party fact-checking to the Community Notes model in the United States (US), aimed at reducing enforcement mistakes, raises doubts about the efficacy of third-party fact-checkers.⁸ This shift also reversed the direction of trust and safety mechanisms against harassment, hate speech, and disinformation.⁹ The reduction of fact-checkers in India is indicative of an asymmetric approach to content verification across the globe.¹⁰ Such changes could weaken timely and credible assessment of manipulated content. Community Notes relies on contributions and agreements among users with different perspectives, which affects the consistency and reliability of content verification.

India's Verification and Information Environment

India's institutional fact-checking initiatives indicate the importance of visual and regional-language verification. The Press Information Bureau (PIB) Fact Check Unit has been functional since 2019.¹¹ During Operation Sindoor in May 2025, it reported blocking over 1,400 Universal Resource Locators (URLs) carrying disinformation propaganda against India and the Indian Armed Forces (IAF).¹² Shakti,¹³ a pan-India joint project launched ahead of the 2024 Indian general elections, connected fact-checkers and news publishers to combat misinformation. DataLEADS,¹⁴ a member of the International Fact-Checking Network (IFCN)¹⁵ initiative, was instrumental in addressing fake news, covering over 60 per cent of content in various regional languages over three months. About 86 per cent of the fact-checked content was visual, of which 2 per cent comprised AI-generated deepfakes. However, the information environment is evolving beyond individual instances of false or manipulated content.

Large Language Models (LLMs) have significant capabilities for generation and coordination of persuasive content at scale. Coordinated Inauthentic Behaviour (CIB) is observed across social media platforms, where AI models generate comments and articles to conduct cyber scams and influence public opinion. OpenAI revealed

activities indicating the coordinated use of GenAI content for gaining engagement and popularity.¹⁶ The persuasive power of AI models depends on post-training prompting rather than scale and personalisation. Conversational LLMs optimised for persuasion tend to produce factually incorrect information and have the potential to falsify information.¹⁷

Online platform defences are improving but remain insufficient to prevent manipulation by threat actors and AI-enabled inauthentic accounts. The likelihood of manipulation remains high due to limited detection of distributed activities and inability to monitor the supporting financial network.

Online platform defences are improving but remain insufficient to prevent manipulation by threat actors and AI-enabled inauthentic accounts. The likelihood of manipulation remains high due to limited detection of distributed activities and inability to monitor the supporting financial network.¹⁸ These capabilities are increasingly reflected in political persuasion and information operations.

The use of LLMs for political persuasion and information operations was evident during the 2024 United States Presidential, 2025 Canadian Federal, and Polish Presidential elections. AI models provide evidence-based supporting facts, making them more effective than psychological persuasion, including social influencing and engineering.¹⁹ For India, this is particularly relevant given its multilingual and diverse information environment. Recognising Foreign Information Manipulation and Interference (FIMI) as a serious threat to freedom of expression, the European Union (EU) has strengthened strategic communication in response to threats and actors targeting EU policies.²⁰ While LLMs can quickly proliferate information manipulation operations in a coordinated fashion, they can also be utilised to analyse data patterns to detect manipulative framings.²¹ EU initiatives such as the Democracy Action Plan and Democracy Shield Initiative, backed by the legal framework of the EU Digital Services Act,²² provide relevant policy approaches for strengthening the resilience of the information environment.

India's Regulatory Response to Synthetic Media

India has addressed online-content governance primarily through the Information Technology (IT) Act, 2000 and IT (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021. The February 2026 amendments introduced the category of “synthetically generated information” (SGI), which includes all forms of audio, visual, and audio-visual information, including modification or alteration. It necessitates due diligence obligations on relevant intermediaries to include technical measures to prevent unlawful SGI.²³ It also mandates social-media intermediaries to remove or

disable access to unlawful content within three hours of receiving a court order or a valid, reasoned intimation from the government or its authorised agency, instead of the earlier 36-hour period. The rules also reduce the timeline for resolving user grievances about prohibited content from 72 hours to 36 hours. Further, the time limit to resolve user grievances has been reduced from 15 days to seven days.²⁴ This accelerated requirement aims to limit the rapid dissemination of unlawful synthetic media, including deepfakes and AI-enabled impersonation. Any online content can achieve its intended impact within a short period after upload.²⁵ However, faster content removal has been criticised for infringing freedom of expression and democratic principles.²⁶ Due to reduced timelines for content removal and user complaints, intermediaries tend to automate these processes, which may undermine freedom of expression through undue censorship without due regard for content creators. The legitimacy of rapid takedown obligations depends on procedural safeguards. Takedown directions should be written, reasoned, and based on clearly defined legal grounds. Platforms should notify affected users, explain the reasons for removal, and establish accessible appeal procedures.

Recommendations: Building Democratic Resilience Against Synthetic Media

India should respond to GenAI-enabled disinformation through a layered framework of democratic resilience. Individual measures, such as technical verification, algorithmic detection, content removal, legal action, and fact-checking, are insufficient when implemented in isolation. A sustainable response should therefore require a whole-of-nation approach that combines technological, institutional, regulatory, and societal measures, while promoting citizen awareness, responsible digital behaviour, and protection of lawful freedom of expression.

- **Technological Innovation and Coordinated Technical Defence.** AI and Machine Learning (ML) should be utilised for deepfake detection, source identification and attribution. India should strengthen watermarking, fingerprinting, embedded metadata, and other persistent provenance mechanisms to establish the authenticity and origin of digital content. Coordinated defences should be developed through threat-intelligence sharing, behavioural-pattern analysis, and detection of Coordinated Inauthentic

India should respond to GenAI-enabled disinformation through a layered framework of democratic resilience. Individual measures, such as technical verification, algorithmic detection, content removal, legal action, and fact-checking, are insufficient when implemented in isolation. A sustainable response should therefore require a whole-of-nation approach.

Behaviour (CIB). LLMs and other AI systems should also be used defensively to identify coordinated campaigns, detect manipulative framing, and support rapid verification of emerging information threats.

- **Policy, Regulation, and Multi-Stakeholder Accountability.** Policy and regulatory frameworks should facilitate the responsible adoption of technological solutions through a multi-stakeholder model involving government, intermediaries, researchers, journalists, civil society and technology developers. The framework should establish clear legal safeguards and accountability mechanisms for intermediaries. Intermediary obligations should include transparent procedures, reasoned decisions and effective accountability mechanisms to prevent arbitrary or excessive removal of lawful content. Synthetic-content governance should maintain an appropriate balance between privacy, freedom of expression and the security of individuals and the state.

- **Institutional Capacity and Public Resilience.** Governance at all levels should strengthen institutional capacity for effective oversight and information verification while balancing individual privacy and national security. Institutional fact-checking capabilities should be strengthened, particularly for regional languages, visual content, and rapidly evolving deepfake techniques. Media literacy, digital literacy, and public awareness should be promoted to help citizens identify manipulated content, verify sources and resist emotionally manipulative narratives. Responsible digital behaviour and citizen participation should be integral to the whole-of-nation approach.

- **International Cooperation and Common Standards.** India should strengthen international cooperation to address the cross-border nature of AI-enabled information manipulation and Foreign Information Manipulation and Interference (FIMI). This should include developing common standards for interoperable technical systems, technology transfer and cooperation, threat intelligence sharing, legal cooperation, information sharing and the exchange of best practices. This also supports coordinated responses to cross-border influence operations and strengthens the interoperability of national verification and attribution capabilities.

Notes:

¹ Kaylyn Jackson Schiff, Daniel S. Schiff and Natalia S. Bueno, "The Liar's Dividend: The Impact of Deepfakes and Fake News on Trust in Political Discourse," Center for Open Science, 2023, pp.39-42, <https://ideas.repec.org/p/osf/socarx/x43ph.html>. Accessed on August 12, 2026.

² Ajith Athrady, "Big Brother Watching? Political Leaders, Digital Rights Advocates Raise Alarm over Takedown of CJP Protest Reels," *Deccan Herald*, August 07, 2026, <https://www.deccanherald.com/india/big-brother-watching-political-leaders-digital-rights-advocates-raise-alarm-over-takedown-of-cjp-protest-reels-4103389>. Accessed on August 12, 2026.

³ Ibid.

⁴ Shavez Mushtaq Qureshi, et. al., "Deepfake Forensics: A Survey of Digital Forensic Methods for Multimodal Deepfake Identification on Social Media," *PeerJ Computer Science*, May 27, 2024, 10.7717/peerj-cs.2037. Accessed on September 10, 2026.

⁵ "A Tool to Watermark and Identify Content Generated through AI," *Google DeepMind*, <https://deepmind.google/models/synthid/>. Accessed on March 04, 2026.

⁶ "Add Watermarks to Content Generated or Altered by Using AI in Microsoft 365," *Microsoft Learn*, November 17-20, 2026, <https://learn.microsoft.com/en-us/copilot/microsoft-365/watermarks>. Accessed on March 04, 2026.

⁷ "How Claude's Text Watermark Works," *Anthropic*, August 14, 2026, <https://www.anthropic.com/news/claude-text-watermark>. Accessed on August 17, 2026.

⁸ Joel Kaplan, "More Speech and Fewer Mistakes," *Meta*, January 07, 2025, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>. Accessed on March 04, 2026.

⁹ Liv McMahon, et al., "Facebook and Instagram get Rid of Fact Checkers," *BBC News*, January 07, 2025, <https://www.bbc.co.uk/news/articles/cly74mpy8klo>. Accessed on March 04, 2026.

¹⁰ Aroon Deep, "Meta to Cut Pay Outs to Fact-Checking Partners in India," *The Hindu*, February 15, 2026, <https://www.thehindu.com/sci-tech/technology/meta-to-cut-pay-outs-to-fact-checking-partners-in-india/article70631703.ece>. Accessed on March 04, 2026.

¹¹ Press Information Bureau, Government of India, "PIB Fact Check Unit," <https://www.pib.gov.in/aboutfactcheck.aspx?reg=3&lang=2>. Accessed on March 04, 2026.

¹² Press Information Bureau, Government of India, "Government Debunks Pakistani Propaganda Against India and Armed Forces via Official Fact-Check Unit," July 30, 2025, <https://www.pib.gov.in/PressReleasePage.aspx?PRID=2150213®=3&lang=2>. Accessed on March 04, 2026.

¹³ Durga Raghunath and Surabhi Malik, "Partnering with News Publishers and Fact Checkers to Tackle Misinformation Ahead of the Indian General Elections 2024," *Google India Blog*, March 01, 2024, <https://blog.google/intl/en-in/company-news/partnering-with-news-publishers-and-fact-checkers-to-tackle-misinformation-ahead-of-the-indian-general-elections-2024/>. Accessed on March 04, 2026.

¹⁴ Enock Nyariki, “Q&A: How India’s Shakti Project Fact-Checked the Largest Election in History at Scale,” August 20, 2024, <https://dataleads.co.in/press-release/qa-how-indias-shakti-project-fact-checked-the-largest-election-in-history-at-scale/>. Accessed on March 04, 2026.

¹⁵ “Commit to Transparency,” IFCN Code of Principles, <https://www.ifcncodeofprinciples.poynter.org/>. Accessed on March 04, 2026.

¹⁶ “Disrupting Malicious Uses of Our Models,” *Open AI*, February, 2026, pp. 1-35, <https://cdn.openai.com/pdf/df438d70-e3fe-4a6c-a403-ff632def8f79/disrupting-malicious-uses-of-ai.pdf>. Accessed on March 05, 2026.

¹⁷ Kobi Hackenburg, et al., “The Levers of Political Persuasion with Conversational Artificial Intelligence,” *Science* 390, no. 6777 (2025): 7-8, <https://doi.org/10.1126/science.aea3884>. Accessed on March 05, 2026.

¹⁸ Gundars Bergmanis-Korāts, et al., “Addressing the Scale of the Market: Assessment by the Financial Intelligence Unit of Latvia,” in the *Social Media Manipulation for Sale: 2025 Experiment on Platform Capabilities to Detect and Counter Inauthentic Social Media Engagement*, NATO StratCom, January 2026, p. 27, <https://stratcomcoe.org/pdfjs/?file=/publications/download/Social-Media-Manipulation-FINAL-FILE.pdf?zoom=page-fit>. Accessed on March 05, 2026.

¹⁹ Lukasz Olejnik, “Weaponising AI for Political Persuasion - Information Operations on an Industrial Scale,” *Security, Privacy & Tech Inquiries*, December 07, 2025, <https://blog.lukaszolejnik.com/weaponising-ai-for-political-persuasion-information-operations-on-an-industrial-scale/>. Accessed on March 05, 2026.

²⁰ “Strategic Communication and Countering Foreign Information Manipulation and Interference,” European Union, https://commission.europa.eu/topics/countering-information-manipulation_en. Accessed on March 04, 2026.

²¹ “AI: Friend or Foe of Disinformation?,” European Union, September 24, 2025, https://joint-research-centre.ec.europa.eu/jrc-explains/ai-friend-or-foe-disinformation_en. Accessed on March 05, 2026.

²² “The Digital Services Act,” European Union, <https://digital-strategy.ec.europa.eu/en/policies/digital-services-act>. Accessed on March 05, 2026.

²³ Ministry of Electronics and Information Technology, Government of India, “Frequently Asked Questions on The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026,” February 10, 2026, p. 2, <https://www.meity.gov.in/static/uploads/2025/10/065b6deb585441b5ccdf8be42502a49c.pdf>. Accessed on August 11, 2026.

²⁴ Ibid. p.10.

²⁵ Arvind Narayanan and Sayash Kapoor, *AI Snake Oil: What Artificial Intelligence Can Do, What It Can’t, and How to Tell the Difference* (Princeton: Princeton University Press, 2024), p. 172.

²⁶ Rahul Unnikrishnan, "A Fact Check Unit That Is Unconstitutional," *The Hindu*, November 21, 2023, <https://www.thehindu.com/opinion/op-ed/a-fact-check-unit-that-is-unconstitutional/article67555398.ece>;

"Digital Media Regulation in India: Balancing Press Freedom," Intellectual Property and Legal Filings (IPLF), March 02, 2026, <https://www.ipandlegalfilings.com/journalism-and-digital-media-misinformation-content-regulation-and-censorship-under-it-rules-in-india/>. Accessed on March 04, 2026.



Centre for Aerospace Power
and Strategic Studies

The Centre for Aerospace Power and Strategic Studies (CAPSS) is an independent, non-profit think tank that undertakes and promotes policy related research, study and discussion on defence and military issues, trends, and development in air power and space for civil and military purposes, as also related issues of national security. The Centre is headed by Air Marshal Nagesh Kapoor, SYSM PVSM, AVSM, VM (Retd).

Centre for Aerospace Power and Strategic Studies
P-284, Arjan Path, Subroto Park, New Delhi 110010
Tel: +91 11 25699130/32, Fax: +91 11 25682533

Editor: Dr Shalini Chawla e-mail: shaluchawla@yahoo.com

Formatting and Assistance: Ms Radhey Tambi, Ms Priyadarshini Baruah and Mr Rohit Singh

The views expressed in this brief are those of the author and not necessarily of the Centre or any other organisation.