myresearchgo Volume 2, Issue 5, May 2026, ISSN: 3107-3816 (Online)

Detection and Social Impact of AI-Generated Deepfake Videos on Digital Trust Among Social Media Users

Author Name: Pranali Vishnu Mane

College Name: Shree Ram College of Commerce, Bhandup.

Subject Area: Artificial Intelligence

Year: 2026

Abstract

The emergence of AI-generated deepfake videos — synthetic media fabricated using Generative Adversarial Networks (GANs), diffusion models, and related deep learning techniques — represents one of the most consequential technological developments in the contemporary information ecosystem. This paper investigates two interlocking dimensions of the deepfake phenomenon: the technical landscape of deepfake detection methodologies, and the social and psychological impact of deepfake proliferation on digital trust among social media users.

Drawing on interdisciplinary literature spanning artificial intelligence, media studies, social psychology, digital sociology, and cybersecurity, this paper provides a comprehensive analysis of deepfake technologies and their societal impact. Key findings indicate that current detection methods remain in an adversarial arms race with generation capabilities. At the societal level, deepfake proliferation contributes to epistemic confusion, political polarisation, erosion of journalistic credibility, and a generalised crisis of digital trust. The paper concludes with a framework of technical, regulatory, platform-level, and media literacy interventions.

Keywords: Deepfake Detection, AI-Generated Media, Digital Trust, Social Media Misinformation, GAN, Synthetic Video, Media Authenticity

Research Scope & Methodology

This paper synthesises peer-reviewed literature from 2018–2025, including computational AI/ML research, social psychology experiments, platform policy analyses, legal scholarship, and survey-based studies across more than 15 countries. The deepfake detection landscape, social trust frameworks, and policy ecosystem are examined as integrated components of a unified challenge to information integrity.

myresearchgo Volume 2, Issue 5, May 2026, ISSN: 3107-3816 (Online)

1. Introduction

In 2018, a video surfaced appearing to show former U.S. President Barack Obama delivering fabricated remarks — a deliberately disclosed warning of what researchers were calling a "deepfake": AI-generated video rendering a real person performing actions or speaking words they never produced. Since that demonstration, the technology has advanced with startling speed, now accessible to ordinary users on consumer hardware through freely available applications.

The proliferation of deepfake technology introduces what scholars term the "liar's dividend" — the capacity for any real video to be dismissed as a potential deepfake, creating epistemological uncertainty in which video evidence loses its evidentiary value. This has profound consequences for journalism, law, democratic processes, and the social fabric of trust underlying online communities.

1.1 Problem Statement

The central problem addressed in this paper is dual in nature. First, existing deepfake detection technologies are engaged in an asymmetric arms race with generation capabilities — improvements in detection have consistently been matched or exceeded by improvements in generation, creating persistent detection gaps. Second, the social and psychological consequences of deepfake proliferation on digital trust among social media users are profound and insufficiently understood.

1.2 Research Questions

What are the technical characteristics of current deepfake generation and detection methods, and what are their fundamental limitations?

How does exposure to deepfake content affect digital trust at individual, institutional, and societal levels?

What demographic, psychological, and contextual factors moderate the impact of deepfake exposure on trust?

What integrated framework of interventions is required to address both the technical and social dimensions of the deepfake challenge?

1.3 Significance

myresearchgo Volume 2, Issue 5, May 2026, ISSN: 3107-3816 (Online)

Research on deepfakes and digital trust sits at the nexus of pressing challenges in contemporary information society. As social media platforms become primary information environments for billions of people, the integrity of video content on those platforms is directly relevant to public health communication, political participation, legal processes, commercial activity, and interpersonal relationships.

2. Understanding Deepfake Technology

2.1 Historical Development

The term "deepfake" is a portmanteau of "deep learning" and "fake," originating in late 2017 on Reddit. While face-swapping and digital video manipulation predate this moment, the incident marked the democratisation of high-quality face synthesis, leveraging advances in deep learning. The underlying technology draws from generative artificial intelligence, revolutionised by the introduction of Generative Adversarial Networks (GANs) by Ian Goodfellow and colleagues in 2014.

2.2 Technical Architecture of Deepfake Generation

2.2.1 Generative Adversarial Networks (GANs)

The GAN framework remains the most widely used architecture for deepfake video generation. Key variants include FaceSwap-GAN (early open-source face swaps), StyleGAN/StyleGAN2 (high-fidelity photorealistic face synthesis), First Order Motion Model (animation of static images), and SimSwap/HifiFace (improved identity preservation).

2.2.2 Diffusion Models

Since 2022, diffusion-based generative models (Stable Diffusion, DALL-E, and video adaptations) have emerged as competitive alternatives to GANs. Video diffusion models, including OpenAI's Sora (2024), are capable of generating highly realistic video sequences from text descriptions — with significant implications for deepfake detection, which historically relied on GAN-specific artefacts.

2.2.3 Voice Synthesis and Audio Deepfakes

Video deepfakes are increasingly accompanied by synthetic voice cloning using variational autoencoders or transformer-based architectures. Systems such as ElevenLabs and Microsoft's VALL-E produce convincing voice clones from as little as three seconds of reference audio. The combination of face-swap video and voice cloning produces multimodal deepfakes substantially more convincing than either component alone.

2.3 Typology of Deepfake Content

Deepfake Type	Description	Primary Harm	Detection Difficulty
Face Swap	Target's face replaced with source's face	Impersonation, NCII	Medium-High

Face Reenactment	Target's expression driven by source's motion	Political manipulation	High
Lip Sync	Target's lip movements altered to match audio	Quote fabrication, Fraud	Medium
Voice Cloning	Target's voice replicated synthetically	Social engineering, Fraud	Medium-High
Text-to-Video	Realistic video from text description	Mass disinformation	Very High

Table 1: Typology of AI-generated deepfake content and associated harm vectors

3. Deepfake Detection: Methods, Advances, and Limitations

3.1 The Detection Challenge

Deepfake detection is fundamentally an adversarial problem. Every advance in detection capability can be used to improve generation quality — if a detector identifies a specific artefact, the generator can be retrained to minimise it. Detection remains structurally disadvantaged: generation only needs to fool humans once, while detection must correctly identify every instance.

3.2 Computational Detection Approaches

Detection methods targeting biological signals include eye blink detection (Li et al., 2018, achieving >99% accuracy on first-generation deepfakes), facial blood flow via remote photoplethysmography (Ciftci et al., 2020), and gaze/pupil behaviour analysis. Spatial inconsistency detection exploits GAN-generated facial boundary artefacts, texture inconsistencies, and spectral domain anomalies. Temporal inconsistency detection identifies facial landmark jitter, inconsistent head pose dynamics, and audio-visual synchrony mismatches.

The state-of-the-art relies on purpose-trained deep neural networks. Key approaches include XceptionNet and MesoNet (early CNNs), the FaceForensics++ benchmark (Rossler et al., 2019), transformer-based detectors using vision transformers (ViTs), and multi-task learning models trained simultaneously on detection and auxiliary tasks.

3.3 Generalisation Problem & Human Detection

The most significant limitation of current detection is poor generalisation across deepfake generators. Models trained on one generator perform near chance levels on others — FaceForensics++ demonstrated in-distribution accuracy above 99% but cross-generator accuracy frequently below 60%. Human observers perform no better: research consistently finds detection accuracy at or near chance (48–57%), with confidence not correlating with accuracy.

Study (Year)	Deepfake Type	Human Accuracy
Korshunov & Marcel, 2018	GAN face swap	~57% (near chance)

Nightingale & Farid, 2022	StyleGAN faces	48.2% (below chance)
Groh et al., 2022	FaceForensics++	50.2% (chance level)
Köbis et al., 2021	Mixed GAN types	54% (slight above chance)

Table 2: Human deepfake detection accuracy (chance = 50%)

3.4 Platform Detection Infrastructure

Major platforms have invested in proprietary deepfake detection. The Facebook AI Deepfake Detection Challenge (DFDC, 2019–2020) produced a 119,000-video dataset; the winning model achieved 82.56% accuracy — still insufficient as a sole detection mechanism. YouTube and TikTok have developed internal detection tools, but independent evaluation is not possible due to lack of transparency.

4. Digital Trust: Theoretical Framework

4.1 Conceptualising Digital Trust

Trust is broadly defined as the willingness to be vulnerable to another's actions based on expectations of appropriate behaviour (Mayer et al., 1995). In the digital context, trust operates at multiple levels: information trust (accuracy of online content), source trust (authentic attribution), platform trust (reliability of content curation), interpersonal trust (authentic communications), and institutional trust (authentic communication from major institutions).

4.2 Trust Erosion and the Liar's Dividend

Deepfake technology threatens digital trust through direct mechanisms (specific incidents of deception) and indirect mechanisms (generalised uncertainty from deepfake awareness). Legal scholar Bobby Chesney and cybersecurity expert Danielle Citron coined the "liar's dividend": deepfake technology enables dishonest actors to claim authentic evidence is fabricated. Politicians confronted with genuine footage, defendants in legal proceedings, and institutions exposed by real recordings can plausibly invoke the deepfake possibility as grounds for dismissal.

4.3 Trust Calibration and the Authenticity Heuristic

Prior to the deepfake era, visual media benefited from a strong implicit authenticity heuristic — the assumption that visually plausible content is genuine. Deepfake technology represents a qualitative escalation by enabling photorealistic video fabrication without physical staging. The disruption of this heuristic creates a trust calibration problem, potentially leading to maladaptive responses including over-generalised distrust (dismissing authentic content as fake) and continued over-trust (failing to apply scepticism due to cognitive load).

5. Social Impact of Deepfakes on Digital Trust

5.1 Political and Electoral Implications

Political deepfakes represent the most studied and alarming application. High-profile cases include a deepfake of Ukrainian President Zelensky appearing to order troops to surrender (March 2022) and deepfake election candidate videos deployed in Slovakia and Taiwan (2023–2024). Research found that even subsequently identified deepfakes produced persistent reductions in candidate favourability ratings. A 2024 survey across six democracies found 41% of respondents concerned deepfakes would influence their upcoming election.

5.2 Journalism, Financial Fraud, and Interpersonal Trust

A 2023 Reuters Institute Digital News Report found only 32% of respondents across 46 countries trusted most online news "most of the time" — concerns about AI-generated content were among the most cited reasons. On the financial front, a 2024 Hong Kong fraud saw a finance employee transfer HK$200 million (approx. USD 25 million) after a video conference with deepfake impersonations of company executives. Meta reported removing tens of thousands of deepfake scam advertisements in 2023 alone.

5.3 Non-Consensual Intimate Imagery and Psychological Effects

Sensity AI (2023) estimated non-consensual intimate deepfake content constituted approximately 96% of all deepfakes identified online, primarily targeting women. Research by Franks and Chesney (2023) found deepfake NCII awareness reduced women's willingness to post photographs online and participate in online professional communities, regardless of personal targeting. Psychological consequences include epistemic anxiety (persistent uncertainty about content authenticity), cynical disengagement (protective withdrawal from online information), and confirmation bias amplification (deepfake uncertainty reinforcing pre-existing worldviews).

6. Demographic Moderators of Deepfake Impact

Contrary to intuition, younger users are not more resistant to deepfakes — digital familiarity does not translate to superior detection. Formal media literacy training, however, is associated with improved detection and calibrated trust responses. Political identity is a significant moderator: users are more likely to believe deepfake content consistent with existing views and dismiss authentic contrary content — representing a fundamental threat to shared political reality. Platform trust baseline also moderates impact: perceived trustworthiness of the platform functions as a moderating variable for the impact of synthetic content.

7. Regulatory, Platform, and Technical Responses

7.1 Legislative Landscape

The United States lacks comprehensive federal deepfake legislation, though the DEEPFAKES Accountability Act (proposed 2021) would require disclosure labels and criminalise certain non-consensual applications. The EU's Artificial Intelligence Act (2024) categorises deepfake dissemination as high-risk in political contexts; the Digital Services Act requires risk mitigation measures. China's Provisions on the Administration of Deep

Synthesis Internet Information Services (March 2022) represent the world's most comprehensive framework, requiring user consent, watermarking, and generation logs.

7.2 Platform Responses and Provenance Technology

Meta prohibits deepfakes "likely to deceive" in political contexts; YouTube prohibits misleading depictions of real people; TikTok requires AI-generated content labelling. The Coalition for Content Provenance and Authenticity (C2PA), including Adobe, Microsoft, and Google, has developed the Content Credentials standard — a cryptographic specification embedding verified metadata about content origin. This approach shifts the authentication burden from detecting fakes to verifying authenticity, though requiring coordinated adoption across the entire content ecosystem.

8. Media Literacy as a Trust-Preservation Strategy

Inoculation theory — derived from the analogy of biological immunisation — proposes that exposing individuals to weakened forms of misinformation techniques produces cognitive resistance to subsequent manipulative content. Research by Roozenbeek and van der Linden (Cambridge) demonstrated that brief inoculation interventions significantly improve users' ability to identify manipulative synthetic media. Critical visual literacy curricula, including understanding of deepfake artefacts, source verification habits, and awareness of emotional manipulation strategies, have demonstrated improved detection in experimental settings. However, limitations include the prohibitive cognitive overhead of applying critical analysis to all video in high-volume media environments, and the continuous evolution of generation technology rendering specific detection cues outdated.

9. Integrated Recommendations

9.1 For Technology Developers and AI Researchers

Prioritise detection generalisation: develop models identifying general characteristics of synthetic content rather than artefacts of specific generators.

Advance C2PA Content Credentials toward universal adoption across capture devices and distribution platforms.

Develop and mandate technically robust invisible watermarking for all AI-generated video content at the point of synthesis.

Share detection models, datasets, and benchmark standards with the broader research community.

9.2 For Social Media Platforms

Implement automated detection-based labelling of identified synthetic content with transparent confidence levels.

Integrate C2PA Content Credentials verification into content display interfaces.

Apply algorithmic friction — reduced amplification, interstitial prompts — to rapidly viralling video lacking provenance credentials.

Provide regular public transparency reports on synthetic media detection, action, and appeal volumes.

9.3 For Governments and Regulators

Enact comprehensive regulatory frameworks addressing generation, distribution, and use of synthetic media across electoral, judicial, commercial, and personal contexts.

Criminalise non-consensual intimate synthetic imagery with clear penalties and accessible civil remedies.

Fund national-scale inoculation and media literacy initiatives, particularly in school curricula.

Establish independent technical auditing of platform deepfake detection capabilities.

9.4 For Individuals and Civil Society

Practise lateral reading when encountering surprising or emotionally provocative video content.

Use reverse video search (Google Video Search, TinEye) and metadata inspection tools.

Recognise that content designed to provoke strong emotional responses is the most likely vehicle for deepfake manipulation.

Choose and promote platforms and news sources that adopt and display Content Credentials.

10. Conclusion

Deepfake technology presents a multidimensional challenge to the integrity of digital information environments and to the foundations of digital trust. Its rapid democratisation has created a situation in which fabricating convincing synthetic video is within reach of ordinary individuals, while reliable detection remains beyond consistent capability of either human observers or computational systems. The social consequences — political manipulation, journalistic erosion, financial fraud, non-consensual intimate imagery, and the generalised epistemic anxiety of navigating an authenticity-uncertain media environment — are already manifest and will intensify.

The trust impact of deepfake proliferation is structurally embedded in the awareness of synthetic media capability itself — in the liar's dividend, in the disruption of the visual authenticity heuristic, and in the cognitive burden of maintaining calibrated scepticism across a high-volume information environment. Addressing this challenge requires not merely better detection technology, but a comprehensive restructuring of the information infrastructure: robust content provenance systems, regulatory frameworks creating deterrence and accountability, platform governance integrating authenticity signals, and education systems cultivating critical visual literacy.

Key Finding Summary

Current deepfake detection achieves 82–99% accuracy in lab settings but drops to near-chance (50–60%) in cross-generator, real-world scenarios. Human detection averages 48–57% — at or below chance. Deepfake awareness alone is sufficient to measurably reduce general digital trust. Pre-bunking/inoculation interventions and content provenance infrastructure represent the most evidence-supported current mitigation strategies.

References

Artificial Intelligence and Detection Research

Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, 107(6), 1753–1820.

Ciftci, U. A., Demir, I., & Yin, L. (2020). FakeCatcher: Detection of synthetic portrait videos using biological signals. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7), 3742–3757.

Dolhansky, B., Bitton, J., et al. (2020). The DeepFake Detection Challenge (DFDC) dataset. arXiv preprint arXiv:2006.07397.

Goodfellow, I., et al. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. CVPR, 4401–4410.

Li, Y., Chang, M. C., & Lyu, S. (2018). In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking. IEEE International Workshop on Information Forensics and Security.

Rossler, A., et al. (2019). FaceForensics++: Learning to detect manipulated facial images. ICCV, 1–11.

Siarohin, A., et al. (2019). First order motion model for image animation. NeurIPS, 32.

Wang, S. Y., et al. (2020). CNN-generated images are surprisingly easy to spot. CVPR, 8695–8704.

Social Science, Trust, and Media Studies

Bruter, M., & Harrison, S. (2023). Deepfakes, disinformation and electoral integrity. Journal of Democracy, 34(2), 112–127.

Franks, M. A., & Chesney, R. (2023). The weaponisation of synthetic media. Georgetown Law Journal, 111(3), 451–512.

Groh, M., et al. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. PNAS, 119(1), e2110013119.

Köbis, N., et al. (2021). Fooled twice: People cannot detect deepfakes but think they can. iScience, 24(11), 103364.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of Management Review, 20(3), 709–734.

Newman, N., et al. (2023). Reuters Institute Digital News Report 2023. Reuters Institute, University of Oxford.

Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces. PNAS, 119(8), e2120481119.

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. Palgrave Communications, 5(1), 65.

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Impact on trust in news. Social Media + Society, 6(1), 1–13.

Policy Documents and Reports

Adobe, Arm, Intel, Microsoft et al. (2021). Content Credentials: An overview of the C2PA specification. Coalition for Content Provenance and Authenticity.

European Commission. (2024). Artificial Intelligence Act: Regulation (EU) 2024/1689. Official Journal of the European Union.

Sensity AI. (2023). The state of deepfakes: Landscape, threats, and impact. Sensity AI Research Report.

U.S. Department of Homeland Security. (2021). Increasing threats of deepfake identities. DHS Science and Technology Directorate.

Cyberspace Administration of China. (2022). Provisions on the administration of deep synthesis internet information services. China National Cybersecurity.

L. C. Kasireddy, L. Popuri, G. Karunanithi, A. Varghese, S. Ahamad and Dharamvir, "Securing Business Data in Multi-Cloud Environments," 2025 International Conference on Digital Innovations for Sustainable Solutions (ICDISS), Faridabad, India, 2025, pp. 1-6, doi: 10.1109/ICDISS68238.2025.11320589.

L. C. Kasireddy, S. Paruchuri, C. Janakamma, A. Sarawat, K. C. Ravi and R. Kumar Chandu, "Cloud-Oriented IoT: Distributed Power-Aware Security Scheme with Data Integrity and Performance Enhancement," 2025 World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS), Indore, India, 2025, pp. 1-6, doi: 10.1109/WorldSUAS66815.2025.11199185.

L. C. Kasireddy, A. Jeraldine Viji, P. K. Sholapurapu, D. Sowjanya Kolluru, D. U. Vishweshwar and P. Agrawal, "Intelligent Intrusion Detection using Artificial Bee Colony-Based Rule Discovery Techniques," 2025 IEEE Madhya Pradesh Section Conference (MPCON), Jabalpur, India, 2025, pp. 691-696, doi: 10.1109/MPCON66082.2025.11256592.

L. C. Kasireddy, S. Paruchuri, C. Janakamma, A. Sarawat, K. C. Ravi and R. Kumar Chandu, "Cloud-Oriented IoT: Distributed Power-Aware Security Scheme with Data Integrity and Performance Enhancement," 2025 World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS), Indore, India, 2025, pp. 1-6, doi: 10.1109/WorldSUAS66815.2025.11199185.

J. L., L. Chandrakanth Kasireddy, R. V. Palanivel, G. Sushma, K. Bhimaavarapu and P. V. Reddy, "Predictive Modeling in Economics: The Role of AI and Deep Learning," 2025 World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS), Indore, India, 2025, pp. 1-7, doi: 10.1109/WorldSUAS66815.2025.11199198.

N. Soni, L. C. Kasireddy, T. S., C. Sinhgadiya, S. Kumar and A. T. S., "A Recurrent Neural Network Framework for Effective DDoS Attack Detection in Cloud Computing," 2025 2nd International Conference on Multidisciplinary Research and Innovations in Engineering (MRIE), Gurugram, India, 2025, pp. 594-598, doi: 10.1109/MRIE66930.2025.11156616.

Jadhav, D., & Shinde, C. (2026). Sakhi: Stay safe stay fashionable. myresearchgo, 2(1), 1. https://doi.org/10.64448/myresearchgo.vol2.issue1.01.

Jadhav, A. (2026). AI-enhanced employee management system. myresearchgo, 2(1), 8. https://doi.org/10.64448/myresearchgo.vol2.issue1.02.

Rane, G., & Matteti, V. (2026). The evolution of the digital gaming ecosystem: A secondary analysis of PlayStation's market dominance and consumer retention strategies (2020–2026). Myresearchgo, 2(3), 1. https://doi.org/10.64448/myresearchgo.vol2.issue3.01.

Ansari, N., Sharma, A., & Yadav, S. (2026). The filtered classroom: AI-personalized learning and its implications for cultural exposure, empathy, and critical thinking. Myresearchgo, 2(3), 12. https://doi.org/10.64448/myresearchgo.vol2.issue3.02.

Junghare, P., Chheniya, J., Behare, M., Kashte, P., Belekar, S., Dhoble, V., & Kumari, S. (2026). Google's Neural Memory Architecture: A Comprehensive Review of the Titans Framework. Myresearchgo, 2(4), 75. https://doi.org/10.64448/myresearchgo.vol2.issue4.12.