International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 -6435 (Online)), Volume 15, Issue 4, April 2026)

The Dark Side of AI-Generated Content: An Exploration of the Misuse of AI in Creating and Disseminating Fake News and Propaganda

Dr. Rais Abdul Hamid Khan[1], Maheen Alim Kazi[2], Pranav Sachin Ahire[3], Aditya Bhausaheb Jadhav[4], Gayatri Ravindra Pendhari[5], Rutuja Bharat Kapadnis[6]

[1]*Professor, School of Computer Science and Engineering, Sandip University, Nashik*
[2]*B-Tech Scholar, School of Computer Science and Engineering, Sandip University, Nashik*
[3]*B-Tech Scholar, School of Computer Science and Engineering, Sandip University, Nashik*
[4]*B-Tech Scholar, School of Computer Science and Engineering, Sandip University, Nashik*
[5]*B-Tech Scholar, School of Computer Science and Engineering, Sandip University, Nashik*
[6]*B-Tech Scholar, School of Computer Science and Engineering, Sandip University, Nashik*

***Abstract*—** The fast-tracked development of generative Artificial Intelligence (AI) has revolutionized the digital world of communication by making it possible to automatically generate highly realistic text, image, audio, and video content. Although such technological developments are of immense use in the fields of education, healthcare, and entertainment, they also have a very serious "dark side" when used for nefarious ends. Currently, AI is increasingly being used for the malicious purpose of generating fake news reports and propaganda on an industrial scale, using advanced tools such as transformer-based Large Language Models (LLMs) and Generative Adversarial Networks (GANs) to produce believable falsehoods with minimal technical effort.

The research paper will delve into the technical underpinnings of AI-generated content (AIGC) and examine its misuse in the production of deepfakes—synthetic media that can convincingly imitate human likeness and voice to the extent that reality and reality-fiction become indistinguishable. Specifically focusing on the Indian scenario from 2023 to 2026, the paper will examine real-world case studies of AIGC's misuse in the form of political manipulation, celebrity identity theft (such as the 2023 Rashmika Mandanna case), financial voice-cloning scams, and the spread of war-related misinformation.

The results show that these AI-based misrepresentations are part of a problem called “truth decay,” where society becomes disillusioned with real media, thus undermining the democratic order, national security, and social cohesion. Moreover, the paper assesses the current state of the challenges in detecting these misrepresentations, where generation methods tend to improve at a rate that outpaces detection methods.

In this regard, the paper recommends a multi-faceted protection strategy that combines technical methods (such as digital watermarking and AI classifiers), new legal frameworks (such as the Digital Personal Data Protection Act in India), and full-scale public digital literacy campaigns. In conclusion, this study makes it clear that the integrity of the information environment must be collectively protected through immediate global collaboration and the development of sound ethical AI guidelines.

"The research also emphasizes the importance of a paradigm shift in the barrier to entry for misinformation; the democratization of AI tools ensures that malicious actors are no longer required to have specialized technical knowledge or financial resources to mount high-impact campaigns. In the Indian scenario, the research points out that the country’s massive multilingual environment presents a fertile ground for AI-powered propaganda to take root, where AI-generated content can be localized instantly to circumvent the need for moderation filters and exploit the social dynamics of the region.

In addition to the political and social implications, this paper examines the underlying psychological ‘truth decay’ where the perpetual presence of deepfakes leads to a condition of chronic epistemic distrust and social paranoia. This is where the ‘liar’s dividend’ takes hold, a secondary consequence where authentic evidence is often discredited by public figures as being AI-generated to avoid accountability."

In order to address these ever-unfolding threats, the research shifts from the theoretical worries to a **five-layered defense strategy**.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 -6435 (Online)), Volume 15, Issue 4, April 2026)

This strategy is built on redundancy and combines an AI-powered detection layer, a cryptographic verification layer (which includes digital watermarking), a platform-level moderation layer, a strong national governance layer, and a crucial human layer that concentrates on digital literacy. Through the examination of these interrelated approaches, this paper asserts that only an interdisciplinary strategy can address the problem, as a technical or legal fix alone is not enough to revive the integrity of the digital information environment.

Keywords— Cyber Regulation, Deepfakes, ,Digital Literacy, Ethical AI, Fake News, Generative Adversarial Networks (GANs), Generative Artificial Intelligence (AI), Information Integrity, Large Language Models (LLMs), Liar's Dividend, Information Integrity, Propaganda, Social Media Algorithms, Synthetic Media, Truth Decay.

I. INTRODUCTION

There has been a paradigm shift in Artificial Intelligence (AI), from rule-based automation to sophisticated generative models that can independently generate human-like text, photorealistic images, and cloned audio. This technological shift, mainly fueled by the development of Large Language Models (LLMs) and Generative Adversarial Networks (GANs), has significantly transformed the communication paradigm around the world. Although these technologies form the foundation of today's digital ecosystem, ranging from automated journalism and recommendation systems to sophisticated medical diagnostics, they have also ushered in a powerful deception toolkit.

The heart of the issue is the unprecedented scalability and realism that generative AI offers. In the past, misinformation campaigns were resource-intensive, meaning that they required human labor to create and disseminate propaganda. However, with the advent of AI models, it is now possible to create "believable falsehoods" on a high frequency and with little technical knowledge or expense. This has led to the dawn of the age of industrial-scale misinformation, where synthetic media, also known as deepfakes, is used to manipulate public perception and evoke emotions through the creation of real-world entities saying or doing things that never happened.

India, with its massive digital ecosystem of over a billion internet-savvy users and growing smartphone penetration, offers a particularly susceptible context for the proliferation of AI-generated misinformation. The convergence of generative AI with social media algorithms has given rise to a "viral loop" where emotionally charged, AI-generated content is selectively amplified for engagement purposes, often circumventing the need for fact-checking. This is especially problematic in a multilingual country like India, where AI can generate localized propaganda in different regional dialects to tap into community-specific dynamics.

The ensuing crisis is one of "truth decay," an epistemological issue in which the constant presence of realistic fakes undermines the public's capacity to distinguish fact from fiction. This situation is not only deceptive; it also creates a secondary effect that is referred to as the "liar's dividend," in which genuine and legitimate evidence is often discredited by public figures as being AI-generated in order to escape accountability. High-profile cases, such as the 2023 viral deepfake video featuring actress Rashmika Mandanna and the employment of AI-generated messages during the 2024 electoral campaigns, have already shown the destructive power of such technologies on issues of personal privacy and democratic stability.

As a result, there is an imperative need to investigate the "dark side" of this technological development. This research paper proposes to fill the existing gap between the technical knowledge and the social implications by investigating the architectural processes of the misuse of AI, studying the practical cases in the Indian context from 2023 to 2026, and assessing the multi-layered defense systems needed to safeguard the integrity of the information environment. By doing so, this research paper will offer a guideline to achieve a balance between the innovation brought about by AI and the needed mechanisms to preserve the authenticity of the information environment in the digital era.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 -6435 (Online)), Volume 15, Issue 4, April 2026)

The democratization of AI technology has made it easier to produce high-fidelity misinformation, which has transitioned from being a state-sponsored activity to anyone with basic internet access. Unlike traditional propaganda, which needed a lot of creative resources, current Large Language Models (LLMs) and Generative Adversarial Networks (GANs) make it possible to produce "believable falsehoods" with little technical knowledge and investment. This has made it possible to produce content that can be customized for a particular demographic, making the deception both highly personalized and very hard to track because of the anonymity that comes with these technologies.

The social implications of this transformation are not only political but also psychological in nature. As synthetic media becomes indistinguishable from reality, it creates a situation of chronic "information overload" and "social paranoia." This situation creates a "loss of epistemic trust," where people begin to question the very possibility of objective truth. As deepfakes become capable of convincingly duplicating the image and voice of public figures, the confusion that ensues often leads to social fragmentation and polarization. Moreover, it creates the "liar's dividend," a very dangerous secondary consequence where actual incriminating evidence is discredited by the malicious as "AI-generated."

In the Indian scenario, the challenges are further exacerbated by the country's large multilingual population and the increasing use of smartphones among the country's one billion internet users. The 2024 election cycle marked an important milestone in this regard, as it was able to show the potential of AI-generated videos to spread political messages in several regional languages, thus effectively overcoming the linguistic constraints that have traditionally been a barrier to the spread of misinformation. These campaigns usually involve the use of emotionally charged content to target social fault lines, thus increasing the chances of the information spreading through engagement-driven social media algorithms.

High-profile cases, such as the 2023 viral deepfake video of actress Rashmika Mandanna, have highlighted the need for a national conversation on digital safety and the limitations of existing legal frameworks.

However, the gap between the speed of content creation and the effectiveness of detection is still increasing. The existing laws, such as the IT Act, are struggling to cope with the issues of AI-generated identity fraud and the responsibility of the platforms on which it is hosted. This introduction emphasizes that the need to comprehend the "dark side" of AI is not only a technical requirement but also a need to maintain the integrity of democratic systems and social harmony in the digital era.

II. LITERATURE SURVEY

The academic study of misinformation has undergone a paradigm shift with the emergence of generative AI, from the human-mediated study of misinformation to the systemic implications of hyper-realistic, algorithmically generated content. The current literature on the subject has outlined several key aspects of this problem, from technical infrastructure to psychological susceptibility.

A. Taxonomy of Information Disorder

One of the key building blocks of the literature is the taxonomy of synthetic media. Scholars such as Wardle and Derakhshan have developed a comprehensive system of classification for "information disorder," which is based on the subtleties of intent and harm. This system of classification differentiates between misinformation (false information with no intent to harm), disinformation (intentionally false information), and malinformation (true information used for malicious purposes). In this regard, AI-generated content, particularly deepfakes, is recognized as a powerful vector for disinformation, which has the capability to simulate real human behavior with a high degree of accuracy, thus creating a continuum between satire and malicious content.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 -6435 (Online)), Volume 15, Issue 4, April 2026)

B. Socio-Political and National Security Risks

There has been extensive research on the effects of misinformation on the socio-political processes of democracy and the trust that the public has on institutions. Allcott and Gentzkow analyzed the economic incentives of misinformation networks and their detectable influence on election results. In a related study, Chesney and Citron investigated the "looming challenge" that deepfakes present to national security, suggesting that the capacity to produce high-stakes events (such as counterfeit political statements or battlefield footage) presents an unprecedented crisis of public confidence. This study highlights the fact that the danger is not only in the event of deception but in the "epistemic threat" that results when the public doubts all legitimate sources of information.

C. Technological Evolution and the Detection Gap

The literature emphasizes the "AI arms race" between content generation and detection. Research papers in IEEE publications emphasize that the exponential growth of Generative Adversarial Networks (GANs) is making it possible to produce synthetic content that does not have the "tells" of being manipulated. Although researchers such as Shu et al. have proposed machine learning solutions to detect fake news based on linguistic and network characteristics, and Li and Lyu have concentrated on detecting specific physiological cues (such as face-warping), the general view in the literature is that the detection gap is widening. The literature suggests that the more advanced the LLMs and GANs, the more unreliable technical detection is.

D. Psychological Vulnerabilities and "Truth Decay"

There is a growing body of research that examines the human side of AI misuse. Thaler and others have studied "motivated reasoning," a psychological phenomenon whereby people are more likely to believe synthetic media is true if it aligns with their pre-existing ideological views. This psychological vulnerability is the main cause of "truth decay," whereby society is so fed up with realistic fakes that trust in actual media is irreparably undermined.

Moreover, the literature points to the "liar's dividend," a perilous secondary consequence whereby the presence of deepfakes alone enables nefarious agents to discredit actual, incriminating evidence as being "AI-generated" in order to escape responsibility.

E. The Indian Context and Regulatory Challenges

In the Indian context, research indicates a greater susceptibility due to the widespread use of social media and the presence of a huge multilingual population. Research shows that more than 75% of Indians have been exposed to deepfake content, and this is a major concern. Researchers suggest that the current legal regime, including the IT Act, is not equipped to deal with the problem of identity fraud using AI. While the Digital Personal Data Protection Act of 2023 and the new IT Rules of 2021 are important developments, the literature suggests that a more proactive approach is needed to address the Indian context.

F. Technical Mechanisms: The Adversarial Training Cycle

In addition to the general reference to GANs, the literature examines the particular "adversarial" connection between two neural networks: the generator and the discriminator. The generator is responsible for producing artificial content, while the discriminator tries to distinguish it from real-world data. Through a series of training cycles, the generator manages to produce content that is almost impossible to distinguish from reality. Moreover, studies on Large Language Models (LLMs), transformer models, demonstrate how they predict sequences of tokens according to contextual probability, allowing for the automatic creation of news stories and political narratives in seconds.

G. Propagation Networks and Algorithmic Amplification

A very important area of research is concerned with the dynamics of social networks and the use of AI-based social bots.

Social bots are designed to replicate human interaction patterns, giving the impression of a global consensus and accelerating the viral spread of misinformation. It has been observed in the literature that social media algorithms tend to favor information that has the potential to generate maximum engagement, which is naturally conducive to AI-based propaganda as opposed to correct and verified information.

H. Ethical Implications and Societal Bias

The ethical implications of the misuse of AI also include the targeting of marginalized groups. Research shows that the use of disinformation campaigns tends to leverage existing societal fault lines, and the AI technology itself tends to perpetuate the inherent biases that exist in the data used to train the technology. This results in greater levels of inequality and social division. Additionally, the literature also points to the "malinformation" challenge that results from the use of accurate information in conjunction with AI dissemination technology.

I. Geopolitical and Economic Aspects

Recent studies have found the application of AI technology in geopolitical affairs, where manipulated images of war and fake footage of battles are employed to mislead the global public opinion. On the economic front, the literature focuses on the effects of AI-generated news on stock markets, indicating that fake news has already resulted in temporary fluctuations in stock market values. The above-mentioned points indicate that the threat model of AI-generated propaganda consists of state actors, political interest groups, and cybercrime rings.

J. Theoretical Framework for Defense

Scholars argue that a "five-layered defense framework" is required to preserve information integrity. This framework proposes redundancy, progressing from a technical Detection Layer (AI classifiers) to the Human Layer (digital literacy).

The literature review indicates that although technical measures such as cryptographic watermarking are necessary, they should be supplemented by global governance and platform accountability to be successful.

III. METHODOLOGY

This research uses a combination of qualitative case study and analysis techniques to examine the mechanisms, effects, and mitigation strategies of AI-driven misinformation. Instead of using controlled experiments, this research combines observations of real-world evidence with academic analysis to create a comprehensive look at the world of synthetic media. The research is broken down into four main stages:

A. Academic Literature Review

The first stage of this research is a comprehensive review of existing academic literature from reputable sources such as IEEE Xplore to gain insight into the underlying mechanisms of AI-driven misinformation. This review will concentrate on the development of Generative Adversarial Networks (GANs), the predictive capabilities of Large Language Models (LLMs), and existing models of "information disorder" classification. By combining the academic knowledge that has been developed, this research sets a theoretical foundation for how synthetic media is created and why it represents an "epistemic threat" to society.

B. Case Study Analysis (2023-2026)

The heart of the study is the analysis of particular deepfake events and automated bot attacks that took place between 2023 and 2026. This stage uses data from news articles, social media incident analysis, and official reports from bodies such as the National e-Governance Division. Notable events such as the 2023 Rashmika Mandanna deepfake and the application of AI in the 2024 elections are analyzed to determine patterns of harm in terms of political, social, and legal spheres.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 -6435 (Online)), Volume 15, Issue 4, April 2026)

C. Technological Evaluation

This stage evaluates the technical capabilities of contemporary generative technology to determine how they are exploited by malicious actors. The evaluation will be on:

- LLMs: Analyzing how transformer models are used to predict sequences of tokens to automatically generate coherent and believable fake news articles.
- GANs: Evaluating the "adversarial training process" where a generator produces fake content and a discriminator tries to detect errors, resulting in highly believable synthetic media that can evade conventional detection methods.
- Social Bots: Investigating how AI-driven bots imitate human interaction patterns to produce a perception of a broad consensus on social media platforms.

D. Online Platform and Policy Assessment

The final stage of the project will evaluate the role of global communication networks in the spread of misinformation. This will include an assessment of the social media policies and algorithms that tend to promote misinformation over fact. Additionally, the project will assess the policy documents of governments, such as the Digital Personal Data Protection Act of 2023 in India, to determine the effectiveness of the current measures taken by governments.

E. Systematic Research Execution Steps

The research work adopts a systematic three-step execution approach to enable a thorough examination of the AI-based misinformation phenomenon. This includes:

1. Identification: Identifying the high-impact Indian deepfake incidents and propaganda activities taking place between 2023 and 2026.
2. Categorization: Categorizing these incidents based on their primary influence domains—namely, Political (election manipulation), Social (community polarization and harassment), and Legal (identity fraud and privacy infringement).
3. Response Evaluation: Assessing the efficacy of the technical, legal, and social implications generated by these particular incidents to determine the existing gaps in the information defense system.

F. Structural Analysis of the Deepfake Lifecycle

In order to address the current gap between the technical process of deepfake creation and its social context, the research methodology applies a structural analysis of the lifecycle of synthetic media. The research study investigates every stage of the lifecycle, starting from the initial Data Collection and Deep Learning Model (GANs) application for face and voice synthesis and ending with the final stage of Synthetic Media Output and Social Media Sharing. This allows the research study to identify specific "intervention points" where the application of detection tools or social media platform moderation is required.

G. Descriptive-Analytical Synthesis

The research study applies a descriptive-analytical synthesis instead of a purely experimental approach. This implies a combination of real-world observations of viral deepfake incidents with academic analysis from peer-reviewed journals. The aim is to develop a reasonable investigation of the "process" of misuse, analyzing the role of technical simplicity and high-speed global communication networks in undermining the public perception of truth.

H. Data Triangulation and Source Diversity

To ensure academic rigor and a balanced perspective, the study adopts a triangulation of diverse data sources. These are:

- Empirical Data: Real-time reports and news coverage (2023-2026) of deepfake incidents in India.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 -6435 (Online)), Volume 15, Issue 4, April 2026)

- Technical Literature: Specialized AI detection research papers and IEEE documentation to monitor the dynamic "detection gap".

- Governmental Overviews: Official reports from the National e-Governance Division to monitor the risks of national security and the current status of cyber-regulatory enforcement.

IV. DISCUSSION

The rise of AI-generated content (AIGC) has led to a complicated socio-technical setting where the pace of innovation tends to overwhelm the establishment of ethical and legal frameworks. This section examines the technical underpinnings of abuse, the infrastructure of its dissemination, and the far-reaching consequences of the Indian social and political landscape.

A. Technical Underpinnings of Synthetic Deception

Contemporary generative AI involves complex architectures that enable the automatic generation of plausible text, images, and audio with little human oversight. Generative Adversarial Networks (GANs) are at the heart of deepfake technology, which uses a competitive process where a generator produces fake content and a discriminator tries to detect its shortcomings. Through thousands of rounds of iteration, the generator learns to generate content that is almost indistinguishable from the real thing, thus effectively erasing the boundaries between what is real and what is fake. At the same time, Large Language Models (LLMs) utilize transformer architectures to predict token sequences based on contextual probabilities, which enables the large-scale production of well-structured news stories and political discourse that simulates human authorship. These technologies have moved from the realm of research to become available instruments for malicious actors, making possible the production of high-fidelity "believable falsehoods" at an industrial scale.

B. The Viral Pipeline: Dissemination and Amplification

The "dark side" of AI is magnified by the infrastructure of the modern global communication networks. AI-generated misinformation is not isolated; it is disseminated through automated bot networks that mimic human interaction to give the appearance of broad consensus. These bots are combined with social media engagement algorithms that favor emotionally engaging or sensationalized content to maximize user engagement, often to the detriment of truth. Moreover, targeted advertising enables malicious agents to target AI-generated propaganda at specific demographic segments, capitalizing on regional social dynamics and existing community prejudices. The intersection of synthetic content creation and algorithmic dissemination ensures that false information reaches millions of users before traditional fact-checking protocols can be brought to bear.

C. Indian Case Studies (2023-2026)

The Indian digital environment, with more than a billion internet-savvy users and a huge multilingual population, is a distinct setting in which to monitor the misuse of AI.

- Case Study 1: The Rashmika Mandanna Deepfake (2023): A viral video that superimposed the face of an actress with a digital replica marked the beginning of a new era in India, where the misuse of AI threatened the privacy of individuals. This event triggered a national discussion on the safety of women in the digital age and the need for tougher cyber laws to regulate the misuse of non-consensual synthetic media.

- Case Study 2: Multilingual Electoral Manipulation (2024): In the 2024 election cycle, AI was employed to create political messages in different regional languages. This enabled propaganda to overcome the usual linguistic constraints, making it hard for voters to distinguish between actual messages from candidates and manipulated videos, thus undermining the democratic process.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 -6435 (Online)), Volume 15, Issue 4, April 2026)

- Case Study 3: Voice Cloning and Financial Fraud: Adversaries employed AI to create voice clones of prominent individuals and relatives for the purpose of executing impersonation scams. Financial frauds like these show that the misuse of AI is not limited to political rhetoric but also poses a direct economic risk to individuals and organizations.

D. The Phenomenon of "Truth Decay" and the Liar's Dividend

The aggregate impact of these technologies is a psychological condition referred to as "truth decay," in which the public is subjected to a level of realistic fakes that ultimately leads to a degradation of public trust in all authentic forms of media. Studies have shown that the public is prone to motivated reasoning, in which they are more apt to accept AI-generated propaganda that supports their pre-existing ideological viewpoints. This type of culture also breeds a secondary consequence that is referred to as the "liar's dividend," in which the presence of deepfakes alone allows for the plausible denial of genuine incriminating evidence in the form of it being AI-generated.

E. Regulatory and Detection Gaps

Although there has been a growing use of AI detection classifiers, the "detection gap" still persists due to the faster development of generation tools. In the Indian context, although the IT Rules 2021 and the Digital Personal Data Protection Act (2023) are positive developments, they tend to be challenged by the pace of technological development and the international character of digital propaganda. Moreover, social media platforms make it even more difficult to detect by deleting the necessary metadata that could have been used to determine the authenticity of a file. Thus, the discussion concludes that a technical fix is not enough but that a harmonized approach is necessary.

The legal issues are further exacerbated by the anonymity and low costs of generative technology, which enable the perpetration of malicious activity with a high degree of impunity,. Although the legislative initiatives of India, such as the IT Rules 2021 and the Digital Personal Data Protection Act, are laudable, the absence of a standardized global framework for AI regulation makes it difficult to track the origin of international misinformation,. Moreover, the extant laws are often ill-equipped to deal with the nuances of AI-related impersonation and identity fraud,,.

From a technological standpoint, we are seeing an "arms race" between GAN models, which are improving at an exponential rate, always staying one step ahead of the computational models that aim to detect them,. This gap is further exacerbated by human cognitive biases, such as "motivated reasoning," where people tend to trust and accept synthetic media that confirms their pre-existing beliefs, even if technical warnings are present,. Social media sites are also being criticized for not having an effective content verification process in place, often relying on algorithms that value engagement over the accuracy of the content being shared.

Lastly, the policy dilemma is characterized by a crucial conflict between shielding society from dangerous propaganda and upholding the basic right to freedom of speech. The ethics of AI are still largely uncharted, and the policies on moderation are also inconsistent across various geographical locations. Without the application of a multi-layered defense system that includes digital watermarking, mandatory labeling of AI-generated content, and digital literacy initiatives at the national level, AI-driven misinformation is bound to spread and threaten the information ecosystem.

V. CONCLUSION

Generative Artificial Intelligence is a revolutionary paradigm in the digital age, which holds both unparalleled opportunities and challenges for the stability of society. This research has clearly shown that the "dark side" of AI, namely its misuse for the creation and spreading of fake news, propaganda, and deepfakes, has moved from a theoretical threat to a serious threat to democracy, economic security, and privacy.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 -6435 (Online)), Volume 15, Issue 4, April 2026)

The unprecedented level of automation, scalability, and hyper-realism enabled by Large Language Models (LLMs) and Generative Adversarial Networks (GANs) has made it possible for malicious actors to create "believable falsehoods" easily, thus causing a systemic loss of trust in information, which is known as "truth decay."

Within the Indian scenario between 2023 and 2026, it has been found that the large multilingual population and deep penetration of social media in the country have made it extremely vulnerable to viral deepfakes and regional political propaganda. The case study analysis of real-life scenarios, including the 2023 Rashmika Mandanna case and the 2024 election manipulation, validates that the use of synthetic media is being harnessed to target not only individual reputation but also the collective democratic integrity. Moreover, the results have also revealed the existence of a "detection gap," where the rate of evolution of content generation techniques is consistently outpacing the identification technologies. This has led to the "liar's dividend," a psychological effect where the presence of deepfakes itself enables the liar to convincingly deny the existence of real evidence, further polarizing the collective understanding of the truth.

However, to effectively mitigate such emerging threats, this paper recommends that a technological or legal remedy alone is not sufficient. Rather, the paper recommends a multi-layered defense strategy that combines five key layers:

1. Detection Layer: Leveraging the use of sophisticated AI-based classifiers to detect synthetic media.
2. Verification Layer: Deploying cryptographic watermarking and content provenance tools.
3. Platform Layer: Adhering to strict moderation algorithms and accountability policies for social media platforms.
4. Governance Layer: Amending existing national and international legal frameworks, including the Digital Personal Data Protection Act (2023) in India, to provide for adequate punishment of AI-based fraud.
5. Human Layer: Initiating broad-scale digital literacy initiatives to enable users to critically assess digital content.

Finally, it is only through an urgent and interdisciplinary effort that the integrity of the information environment can be protected. Future research needs to concentrate on the development of real-time detection tools and the improvement of ethical standards for AI to ensure that the pursuit of innovation does not have to be at the expense of public trust. In this way, the negative consequences of misinformation generated by AI can be mitigated while still reaping the rewards of the creative and productive power of artificial intelligence

ACKNOWLEDGMENT

We sincerely thank the Department of Computer Engineering at Sandip University, Nashik, for fostering a supportive academic environment and providing the resources needed to carry out this research. We are especially grateful to our project guide and faculty members for their constant encouragement and valuable insights. Their expertise helped us better understand the technical aspects of Generative Adversarial Networks (GANs) and Large Language Models (LLMs), while also guiding us to critically examine the ethical challenges and risks associated with these technologies.

We also extend our appreciation to the researchers and government organizations whose work significantly contributed to this study. Reports from the National e-Governance Division and other publications on India's digital ecosystem played a key role in shaping our understanding of deepfakes, misinformation, and the growing issue of "truth decay" between 2023 and 2026. Case studies such as the Rashmika Mandanna deepfake incident and multilingual electoral manipulation provided valuable context for developing our proposed multi-layered defense strategy.

Lastly, we are deeply thankful to our families and friends for their continuous support, motivation, and belief in our work. This research reflects a shared commitment to preserving trust and authenticity in today's rapidly evolving digital landscape.

REFERENCES

[1] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," J. Econ. Perspect., vol. 31, no. 2, pp. 211–236, 2017.

[2] B. Chesney and D. Citron, "Deepfakes and the new disinformation war," Foreign Affairs, vol. 98, no. 1, pp. 147–155, 2019.

[3] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," ACM SIGKDD Explor., vol. 19, no. 1, pp. 22–36, 2017.

[4] P. L. Kharvi, "Understanding the Impact of AI-Generated Deepfakes on Public Opinion, Political Discourse, and Personal Security in Social Media," IEEE Security & Privacy, 2024.

[5] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," Proc. IEEE CVPR Workshops, 2019.

[6] M. Thaler, "The Fake News Effect: Motivated Reasoning Using Trust in News," American Economic Journal, 2024.

[7] C. Vaccari and A. Chadwick, "Deepfakes and disinformation," Soc. Media + Soc., vol. 6, no. 1, 2020.

[8] P. Singh, "AI-Generated Fraud in India," 2025.

[9] National e-Governance Division, "Reports on AI and Misinformation," 2023–2026.

[10] Times of India, "Deepfake Reports (2023–2026)," 2026.

[11] Home Security Heroes, "2023 State of Deepfakes: Realities, Threats, and Impact," 2023.

[12] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, arXiv:1312.6114.

[13] AI Detection Research Papers, "Technical Trends in Identification," 2025.

[14] C. Wardle and H. Derakhshan, "Information Disorder: Toward an interdisciplinary framework for research and policymaking," Council of Europe, 2017 (Cited for the Taxonomy used in).

[15] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.