Theory and Practice of Science and Technology
2025, VOL. 6, NO. 6, 118-120
DOI: 10.47297/taposatWSP2633-456923.20250606

Research on Governance System for Generative AI Misinformation

Yiqi Lu*

School of Management, Northeastern University at Qinhuangdao, Qinhuangdao, Hebei, 066004, China

ABSTRACT
AI-generated misinformation (AIGM), characterized by high generation efficiency, strong content fidelity, and rapid dissemination capabilities, poses a systemic threat to social trust, democratic processes, and public safety. To address this challenge, active defense strategies, such as content watermarking and model provenance authentication, are integrated. This research breaks through technical bottlenecks in Bayesian Neural Networks (BNNs) for uncertainty quantification and federated learning with dynamic updates, establishing a dynamic ethical alignment theory. However, it faces challenges including rapid generative technology evolution, multimodal detection complexity, and privacy protection. Continuous development of universal foundational detection models and secure hardware-enhanced watermarking technologies is required.

KEYWORDS
Artificial intelligence-Generated misinformation; Deepfakes; Watermarking techniques; Active defense

1 Introduction

The rapid advancement of artificial intelligence (AI) technologies, especially large language models (LLMs) like the GPT series, Claude, and Gemini, as well as diffusion models such as DALL-E, Stable Diffusion, and Midjourney, has brought transformative productivity gains but also spawned a new type of information threat—Artificial Intelligence-Generated Misinformation (AIGM). This form of misinformation exhibits characteristics such as high generation efficiency, strong content fidelity, deep customization, and rapid propagation speed, far surpassing traditional human-fabricated false information. Radford et al. (2019) presciently noted in their seminal paper "Language Models are Unsupervised Multitask Learners" that as model scale increases, the fluency and coherence of generated text would significantly improve, but this might be "misused to generate misleading content." This foresight has become a stark reality [1].

The harmfulness of AIGM is substantial. It can fabricate news reports (e.g., generating false descriptions of political events), impersonate authoritative figures' statements (e.g., generating pseudo-scientific articles in the tone of renowned experts), create highly inflammatory social media content (e.g., deepfake videos of politicians' speeches), and produce highly personalized fraudulent information (e.g., targeted spear-phishing emails). Vosoughi et al. (2018) demonstrated in their Science paper "The spread of true and false news online" through large-scale data analysis that false news, especially novel falsehoods, spreads faster, wider, and deeper than true news [2]. AIGM, with its high degree of "novelty" and "fidelity," further amplifies this effect, posing systemic risks to social trust, democratic processes, public health (e. g., vaccine rumors), and financial market stability [3]. Consequently, developing efficient and robust methods for detecting AI-generated misinformation has become a core task in information security and AI governance. This review aims to systematically outline the main technical approaches, core methodologies, evaluation metrics, existing challenges, and future trends in AIGM detection, providing a reference for in-depth research and practical application in this field.

Table 1 Core Feature Comparison between AI-Generated Misinformation (AIGM) and Traditional Misinformation

Feature Dimension	Traditional Misinformation	AI-Generated Misinformation (AIGM)
Generation Speed	Relatively slow, reliant on manual fabrication	Extremely fast, batch generation on demand(seconds/thousands of words)
Generation Scale	Limited	Massive, theoretically infinite generation
Content Fidelity	Variable, prone to revealing flaws	Extremely high, fluent language, logical coherence, rich details
Customization Capability	Limited	Extremely strong, capable of precise generation targeting specific objects/scenarios
Cross-modal Fusion	Difficult, often requires multi-person collaboration	Seamless, text, image, audio, video can be generated interactively
Evolution & Evasion	Weak	Strong, models can rapidly iterate to evade existing detection
Traceability Difficulty	Relatively traceable (human traces)	Extremely high, machine-generation traces easily masked or imitated

2 Characteristics and generation mechanisms of AI-generated misinformation

Deeply understanding the essential characteristics of AIGM and its underlying technical generation mechanisms is fundamental to building effective detection methods. AIGM does not arise in a vacuum; its core features are rooted directly in the architectures and training paradigms of current mainstream generative AI models.

2.1 Core characteristic analysis

The most salient feature of AIGM is its high degree of surface plausibility. Transformer architecture-based large language models learn extremely complex linguistic statistical patterns and world knowledge representations through pre-training on massive amounts of unlabeled text [4]. This enables their generated content to be impeccable in grammar, common expressions, and basic factual associations. However, beneath this surface plausibility lurk factual errors (Hallucination), logical inconsistencies, and potential bias amplification. Ji et al. (2023) systematically analyzed multiple causes of hallucinations in LLMs in their "Survey of Hallucination in Natural Language Generation," including noisy training data, model overgeneralization, and prompt engineering induction [5]. These hallucinations are actively exploited in misinformation generation scenarios. Secondly, AIGM possesses strong adaptability and evolvability. Generative models (especially open-source ones) can be fine-tuned or optimized via prompt engineering for specific tasks (e. g., imitating a journalist's style, fabricating "evidence" conforming to a conspiracy theory) to produce highly customized and targeted deceptive content [6]. Finally, multi-modal fusion is an emerging trend. The combination of text generation models with image, audio, and video generation models (e. g., LLM + Diffusion Model) can create "fully-fledged" composite false information with text and audiovisuals, significantly increasing deception potential and detection difficulty (Deepfakes are a prime example).

2.2 Technical generation mechanisms

From a technical implementation perspective, AIGM relies primarily on the following paradigms:

(1) Autoregressive Generation

Examples include GPT series models. Based on previously generated token sequences, they predict and generate the next most probable token one-by-one (typically using Top-p or Top-k sampling). This mechanism excels at producing fluent long-form text but is prone to errors in long-range dependencies and complex logic and is susceptible to prompt-induced bias or false content.

(2) Diffusion Models Dominant in image and video generation (e.g., Stable Diffusion, DALL-E 3). They generate samples by iteratively denoising noisy data. These models can produce images and videos of extremely high visual fidelity, but also make forging celebrity likenesses or fabricating scenes effortless. Key challenges lie in controlling content consistency with the true intent (avoiding unintended generation) and preventing malicious use.

(3) Retrieval-Augmented Generation (RAG)

Combines retrieval from external knowledge bases with text generation. While intended to improve content accuracy, if retrieved sources contain misinformation or the model misinterprets results, it can authoritatively generate seemingly authoritative content based on false information. Attackers may also deliberately contaminate knowledge bases.

(4) Adversarial Attacks & Jailbreaking

Involves crafting adversarial examples (malicious prompts) to induce models to bypass their safety constraints and generate prohibited, biased, or false content they were designed not to produce (Carlini et al., 2023, "Extracting Training Data from Diffusion Models" shows diffusion models risk leaking private data via targeted unlearning attacks; "Jailbreak" attacks target LLM safety alignment mechanisms).

3 Active defense technologies: watermarking and provenance tracing

Actively embedding traceable identifiers in generated content or ensuring model provenance verifiability forms a crucial line of defense against AIGM dissemination.

3.1 Content watermarking

In the text watermarking domain, the Kirchenbauer watermark (KGW), proposed by Kirchenbauer et al. (2023), is a prominent current scheme. Its core mechanism involves implanting statistical features during text generation by controllably biasing token selection probabilities Other research approaches include modifying grammatical structures (e. g., regular use of specific sentence patterns), introducing subtle reversible spelling variants, or embedding information using the redundancy space in text representations. Key challenges in this field involve balancing four core elements: imperceptibility/Fidelity (ensuring the watermark doesn't affect semantics or reading experience), Robustness (resisting attacks like editing/paraphrasing), Capacity (amount of information the watermark can carry), and Security (preventing malicious removal or forgery).

In image/video watermarking, traditional digital watermarking methods are mature, while novel technical paths have emerged for AI-generated content:

(1)Invisible Watermarks: Often embed weak signals correlated with visual content in the frequency domain (e.g., DCT, DWT coefficients) or spatial domain, requiring resistance to common operations like compression, cropping, scaling. Recently, the latent space characteristics of diffusion models have inspired new robust watermark designs, such as modulating latent representations to embed imperceptible markers.

(2)Visible Watermarks: Exemplified by OpenAI adding identifiers to the corners of DALL-E 3 generated images. While direct and effective, this method faces limitations like easy cropping and compromised user experience.

(3)Model Fingerprinting: Involves embedding information specific to the model or user into generated content to trace leak sources or responsible parties.

Audio Watermarking typically embeds inaudible signals in specific frequency bands (e.g., echo watermarking, spread spectrum watermarking) or modifies phase information, requiring robustness against resampling, compression, and noise addition. By balancing audio perceptual quality and watermark robustness, this technology supports source tracing and anti-counterfeiting for AI-generated audio content.

3.2 Model provenance tracing & authentication

In the field of model provenance tracing and authentication:

(1)Model Fingerprinting/Signing: Aims to embed unique digital fingerprints or signatures during model training or deployment, causing model outputs to carry identifiable patterns related to architecture, training data distribution, or hyperparameters. Research by Nie et al. (2023) indicates that analyzing latent fingerprints in generated samples can infer the source model, providing a technical path for tracing developers or users of malicious models.

(2)Secure Publishing & Authentication Frameworks: This consists of two main directions: (1) Content Credentials: Promoted by the Coalition for Content Provenance and Authenticity (C2PA), this standard works to create verifiable metadata chains for digital content (images, videos, audio, documents), recording creator (human or AI), editing tools, modification history, etc. For example, Adobe's Content Credentials system already enables users to view and verify content credentials with supported tools [3].(2) Model Cards & Usage Agreements: Require AI model providers to publish Model Cards detailing capabilities, limitations, training data, potential biases, and intended uses. Stringent usage agreements can lower the risk of models being used for misinformation generation.

(3)Blockchain & Distributed Ledger Technology: Leverages immutability to record AI model creation, modification, deployment information, and hashes/provenance credentials of key generated content (including watermark info), providing a trusted platform for traceability and verification.

These active defense technologies, particularly standardized, interoperable watermarking and provenance schemes (e.g., C2PA), constitute critical infrastructure for a healthy AI-generated content ecosystem, offering core capabilities for platforms, law enforcement, and users to verify content authenticity and origin.

4 Conclusion

AI-generated misinformation (AIGM) is a complex and severe challenge of the digital age. Leveraging the powerful capabilities of generative AI, it manufactures and disseminates deceptive content with unprecedented efficiency, scale, and fidelity, seriously threatening the health of the information ecosystem, the foundation of social trust, and the stability of democratic systems. Addressing this challenge cannot rely on single-point solutions; it necessitates constructing a comprehensive defense system integrating technological innovation, platform governance, laws and regulations, public education, and ethical norms.

At the technological level, this review systematically outlines the forward-looking active defense strategies—content watermarking and provenance authentication (e.g., the C2PA standard). While each method has its advantages and limitations, the primary focus of future research lies in enhancing the robustness, generalization capabilities, and explainability of detection models. Exploring new directions such as universal foundational detection models, secure hardware-enhanced watermarking, privacy-preserving detection, and emphasizing training and evaluation in adversarial environments is crucial.

The battle between AI-generated misinformation and its detection is an ongoing, dynamic contest. This contest is not solely about technological superiority but involves profound social governance, ethical choices, and value judgments. Only through global collaboration, interdisciplinary research, responsible innovation, and widespread public awareness can we effectively harness the dual-edge nature of generative AI, prevent its abuse for spreading lies and division, and ensure this transformative technology ultimately serves human truth, understanding, and progress. Continuous investment in detection technology R&D, robust platform governance, improved legal frameworks, and enhanced public literacy are essential choices for building a digital future more resilient to the harms of AI-generated misinformation.

About the Author

* Corresponding Author: Yiqi Lu, hallielu113@gmail.com

References

[1] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI blog, vol. 1, no. 8, p. 9, 2019.

[2] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," science, vol. 359, no. 6380, pp. 1146-1151, 2018.

[3] D. M. Lazer et al., "The science of fake news," Science, vol. 359, no. 6380, pp. 1094-1096, 2018.

[4] A. Vaswani et al., "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[5] Z. Ji et al., "Survey of hallucination in natural language generation," ACM computing surveys, vol. 55, no. 12, pp. 1-38, 2023.

[6] R. Bommasani et al., "On the opportunities and risks of foundation models," arXiv preprint arXiv:2108.07258, 2021.