Industrialized Deception: The Collateral Effects of LLM-Generated Misinformation on Digital Ecosystems*

Alexander Loth
alexander.loth@stud.fra-uas.de
Frankfurt University of Applied Sciences
Frankfurt am Main, Germany

Martin Kappes
kappes@fra-uas.de
Frankfurt University of Applied Sciences
Frankfurt am Main, Germany

Marc-Oliver Pahl
marc-oliver.pahl@imt-atlantique.fr
IMT Atlantique, UMR IRISA, Chaire Cyber CNI
Rennes, France

arXiv:2601.21963v3 [cs.CY] 29 Oct 2026

Abstract

Generative AI and misinformation research have evolved since our 2024 survey. This paper presents an updated perspective, transitioning from literature review to practical countermeasures. We report on changes in the threat landscape, including improved AI-generated content through Large Language Models (LLMs) and multimodal systems – addressing the adverse impacts of AI on information quality.

This work focuses on our practical contributions: *JudgeGPT*, a platform to evaluate human perception of AI-generated news, and *RogueGPT*, a controlled stimulus generation engine for research. Together, these tools form an experimental pipeline for studying how humans perceive and detect AI-generated misinformation.

Our findings show that detection capabilities have improved, but competition between generation and detection continues. We discuss mitigation strategies such as LLM-based detection, inoculation approaches, and the dual-use nature of generative AI.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; • **Information systems** → *Social networks*; • **Security and privacy** → *Social aspects of security and privacy*.

Keywords

Generative AI, LLM-Generated Misinformation, Fake News Detection, Collateral Effects, Digital Ecosystems, Deepfakes, Human Perception, Dual-Use AI

1 Introduction

Generative AI has changed how information—and misinformation—spreads online. The ability to generate convincing text at scale has enabled what can be termed *industrialized deception*: the automated production of misleading content affecting digital ecosystems. These digital ecosystems—comprising interconnected networks of platforms, users, algorithms, and content—have become the primary infrastructure through which information flows in modern society [17]. The health of these ecosystems depends on the trustworthiness of information circulating within them; when misinformation proliferates, it erodes trust not only in specific content but in the information infrastructure itself. In our 2024 survey [37], we examined the interplay between Generative AI and Fake News, covering enabling technologies, content creation, detection methods, and deepfake threats. Since then, the field has evolved in ways that warrant renewed examination.

Large Language Models (LLMs) have improved considerably, intensifying the competition between generation and detection of synthetic content [6, 49]. Researchers have explored the dual-use nature of LLMs—their capacity to both generate and detect misinformation [21, 47]—with societal implications extending to privacy, manipulation, and trust erosion [3].

This paper makes three key contributions. First, we provide an updated perspective on how the Generative AI and Fake News research domain has evolved since 2024, highlighting new challenges including the emergence of multimodal misinformation—where text, images, audio, and video are combined to create more convincing deceptive content [2, 28]—and the shift toward agentic AI systems capable of autonomous content generation and dissemination, which motivates moving beyond content-level detection toward behavioral-level analysis of coordinated inauthentic behavior [1, 60]. Second, we present our methodological contributions: the open-source tools *JudgeGPT*[1] [38] and *RogueGPT*[2], which together form an experimental pipeline for studying human perception of AI-generated news. JudgeGPT serves as an empirical data collection platform where participants evaluate news fragment authenticity, while RogueGPT provides controlled stimulus generation for research purposes. Our longitudinal expert perception survey reveals that large-scale text generation poses systemic risks of "epistemic fragmentation" and "synthetic consensus"—risks now formalized in Ferrara's "Generative AI Paradox" framework [19]—while experts express skepticism toward purely technical detection tools, preferring provenance standards and regulatory frameworks aligned with emerging "epistemic security" objectives [30, 40]. Third, we discuss emerging mitigation strategies and the role of AI itself in combating misinformation, including inoculation theory [33] and LLM-based detection approaches [21].

Our work contributes to efforts addressing AI's adverse impacts and collateral effects by examining both threats and countermeasures [23, 59]. The dual nature of Generative AI—as both a tool for creating deceptive content and for detecting it—warrants continued research as these technologies become more widely deployed.

This paper is structured as follows: Section 2 introduces key concepts and recent developments. Section 3 reviews relevant research since 2024. Section 4 presents our methodological contributions.

*This research is funded by the European Union project CyberSecDome under the grant agreement 101120779. Published in WWW Companion '26, pp. 295–302. DOI: 10.1145/3774905.3795471.



This work is licensed under a Creative Commons Attribution 4.0 International License.

[1] https://github.com/aloth/JudgeGPT
[2] https://github.com/aloth/RogueGPT

Section 5 synthesizes findings and discusses mitigation strategies. Section 6 presents conclusions and future directions.

2 The Domain

The term "Generative Artificial Intelligence (AI)" refers to AI technologies designed to produce new content. This content includes text, images, audio, and other media forms, often resembling human-generated output. Models like GPT-4 can produce text that evaluators struggle to distinguish from human writing [4].

Machine learning models such as Generative Adversarial Networks (GANs), transformers, and variational autoencoders enable this capability. Trained on large datasets, these models learn to generate new instances that reflect patterns in their training data [8, 26].

2.1 Evolution Since 2024

Since our 2024 survey [37], several developments have changed the Generative AI and Fake News research area:

LLMs and Multimodal Systems. Models like GPT-4o, Claude 3.5, and Gemini 1.5 have improved at generating coherent content across multiple modalities [18]. Reasoning models (e.g., OpenAI o1) and small language models for edge deployment have widened access to generation capabilities [32]. The emergence of large vision-language models (LVLMs) has changed both generation and detection of multimodal content [2].

Multimodal Misinformation Challenges. The combination of text, images, audio, and video in misinformation poses detection challenges that exceed single-modality approaches. Out-of-context misinformation—where authentic content is paired with misleading narratives—has emerged as a common form requiring cross-modal semantic analysis [20]. Recent advances include agentic frameworks that use web-grounded reasoning for verification [53] and tool-augmented detection agents [13].

Dual-Use Nature of LLMs. Recent research has explored how the same LLMs that can generate misinformation can also be used for detection [21, 49]. Sallami et al. demonstrate that LLMs exhibit both creative capabilities for generating convincing fake content and analytical capabilities for identifying it [49].

Bias and Fairness Concerns. The research community has increasingly focused on biases embedded in AI detection systems, including gender bias in Fake News detection [46] and the need for fairness frameworks [48].

Agentic AI and the Operationalization of Influence. A notable development since 2024 is the emergence of agentic AI as a vehicle for industrialized deception. The threat model has shifted from human actors leveraging GenAI tools to autonomous agents capable of independent reasoning, planning, and execution [22]. Tseng et al. (2026) provide evidence of multi-agent pipelines systematizing Foreign Information Manipulation and Interference (FIMI), with specialized agentic components mapping behaviors to the DISARM framework [60]. This represents a shift from a "content abundance" problem to a "coordination abundance" problem—the constraint on disinformation campaigns is no longer human labor but compute [1]. Autonomous agents can perceive information environments, reason about psychological triggers, generate tailored multimodal content, and refine strategies based on real-time engagement metrics without human intervention.

Beyond Detection to Prevention. Research focus has shifted from detection toward prevention strategies, including inoculation approaches [33] and "prebunking" techniques [47].

As Generative AI improves, it both enables synthetic content creation and provides tools for detection. This dual-use nature motivates work on content authenticity verification. Cryptographic provenance standards such as C2PA offer an alternative to detection by establishing verifiable chains of content origin; our Origin Lens framework implements privacy-preserving on-device verification using a defense-in-depth approach [34, 36].

2.2 Structural Overview

Figure 1 illustrates the domain structure. At the core, Generative AI branches into two principal areas: creation and detection of Fake News. The Creation aspect encompasses Text Generation, Image Synthesis, Audio Generation, and Video Generation—representing the diverse capabilities to produce content indistinguishable from human-created material. The Detection branch addresses Content Verification, Social Media Analysis, and Crowd Sourcing for identifying synthetic content.

Adjacent to these themes are mitigation strategies (public awareness, regulatory policies) and ethical considerations (privacy concerns, bias and fairness) [3, 58]. Connecting these nodes are enabling technologies: Autoencoders, GANs, Transformers, GPTs, and VAEs.

2.3 Digital Ecosystems and Information Integrity

Digital ecosystems comprise interconnected networks of platforms, users, algorithms, and content that collectively shape how information flows through society. These ecosystems include social media platforms, search engines, news aggregators, messaging applications, and recommendation systems—each influencing what content users encounter and share [18]. The health of digital ecosystems depends on multiple factors: the trustworthiness of information sources, the transparency of algorithmic curation, and the resilience of users to manipulation [17].

From Fake News to Synthetic Reality. Ferrara (2026) argues that the prevailing focus on "deepfakes" or "misinformation" misses a broader socio-technical shift: the creation of *Synthetic Realities* [19]. This framework formalizes the threat as a layered stack: (1) *Synthetic Content*—the raw text, image, audio, or video artifacts; (2) *Synthetic Identity*—the fabrication of coherent personas that persist over time; (3) *Synthetic Interaction*—the simulation of social presence, engagement, and relationship-building; and (4) *Synthetic Institutions*—the manufacture of consensus through coordinated networks of fake outlets and organizations. This final layer is particularly relevant to "Industrialized Deception," as it implies the automation of credibility itself, not just content.

The Generative AI Paradox. Ferrara's "Generative AI Paradox" posits that as synthetic media becomes ubiquitous and indistinguishable from authentic content, societies will rationally discount *all* digital evidence. The cost of verification becomes prohibitively high compared to the cost of generation, leading to a market failure in the information ecosystem [19]. Trust is not merely eroded; it is rendered economically irrational. This aligns with the concept of "Epistemic Security" highlighted in recent policy discussions,

where the goal of defense shifts from "correcting false information" (which assumes a functioning marketplace of ideas) to "securing the conditions for knowledge creation" (which acknowledges that the marketplace itself is flooded) [30].

Generative AI poses systemic risks to these ecosystems through several interconnected mechanisms. First, LLMs enable the production of misleading content at high *scale and speed*, overwhelming traditional fact-checking and moderation systems [43]. Second, synthetic content tailored to specific audiences creates *epistemic fragmentation*—information bubbles with incompatible worldviews that fragment shared understanding, a key concern identified in expert surveys [30, 40]. Third, coordinated deployment of AI-generated content manufactures *synthetic consensus*, exploiting the Synthetic Institutions layer of Ferrara's stack to manipulate perceptions of public opinion [19]. Finally, as users become aware of AI-generated content, skepticism extends to authentic content, embodying the Generative AI Paradox where rational actors discount all digital evidence—a phenomenon we term *trust erosion*.

Platform algorithms amplify these effects by optimizing for engagement metrics that often favor sensational or emotionally charged content—characteristics that AI can readily generate [5, 27]. Understanding these ecosystem dynamics is useful for developing mitigation strategies that address not just individual pieces of misinformation but the structural conditions enabling their spread.

2.4 Functioning of Generative AI for Fake News Generation

Generative AI models synthesize new data by learning from existing datasets. They function through deep learning architectures such as GANs, VAEs, and Transformers. GANs pit two neural networks against each other to produce new, synthetic instances of data. GANs generate realistic images and videos to accompany synthetic Fake News stories.

Transformers utilize attention mechanisms to generate coherent sequences of text[61]. Transformers, like GPT models, are trained on vast corpora of text. Transformers are able to produce all kind of text, including Fake News[18, 50].

2.5 Technical Background

This section provides an overview of the key technologies and concepts foundational to understanding the intersection of Generative AI and Fake News. It introduces the essential definitions and methodologies employed in the survey.

2.5.1 Generative Artificial Intelligence. Generative AI refers to a subset of AI technologies designed to create content that mimics real-world data. These models learn to generate new data samples that are indistinguishable from authentic datasets.

2.5.2 Generative Adversarial Networks (GANs). GANs consist of two neural networks, the generator and the discriminator, which are trained simultaneously through adversarial processes. The generator creates data samples aimed at fooling the discriminator, while the discriminator evaluates them against real data, improving both models iteratively[25].

2.5.3 Variational Autoencoders (VAEs). VAEs are generative models that use a probabilistic approach to produce data. They learn to encode input data into a latent space and reconstruct it back, ensuring that generated samples adhere to the probability distribution of the input data[31].

2.5.4 Transformer Models. Transformers are a type of neural network architecture designed for processing sequential data, particularly text. They rely on self-attention mechanisms to weigh the significance of different parts of the input data[61]. A recent breakthrough in this area is the development of 1-bit Large Language Models (LLMs), such as those introduced by Ma et al. (2024)[41], which achieve comparable performance to full-precision models with significantly reduced computational costs.

BERT. Bidirectional Encoder Representations from Transformers (BERT) is a model designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers[16].

GPT. Generative Pre-trained Transformer (GPT) models generate coherent text based on a given prompt. These models can perform various language tasks without task-specific training [45].

As described in Figure 2, the GPT model architecture is designed to capture and generate human-like text by processing input through a series of transformer blocks. Each block enhances the model's understanding of language context and structure, allowing for the generation of coherent and contextually relevant text. This mechanism allows the model to simulate various forms of written content, including Fake News, by leveraging learned patterns from extensive data sets.

3 Overview of Recent Developments

This section provides a condensed review of developments since our 2024 survey [37], organized around key themes.

3.1 The Agentic Shift in Misinformation

The 2025–2026 literature reveals a shift in the threat landscape: the human actor is increasingly removed from the loop, replaced by autonomous agents. Tseng et al. (2026) demonstrate multi-agent pipelines that operationalize the DISARM framework for investigating Foreign Information Manipulation and Interference (FIMI) [60]. While their work focuses on defensive applications, the architecture illustrates the dual-use potential: specialized agents can collaboratively map manipulative behaviors to standardized Tactics, Techniques, and Procedures (TTPs).

This implies that future detection systems must operate at the *behavioral level*—identifying agent strategies—rather than the content level, as content will be hyper-optimized to evade static classifiers. The ACM Europe Technology Policy Committee (2025) highlights this as a systemic risk, noting that current regulatory frameworks may govern "models" but fail to address emergent behaviors of "agents" exhibiting persistent operation and adaptive learning [1].

3.2 Advances in LLM-Based Misinformation

Sandrini and Somogyi (2023) analyze GenAI's effects on news consumption, finding that early-stage GenAI leads consumers toward deceptive content, though benefits emerge after reaching development thresholds [50]. Kumar et al. (2025) examine generative-AI-driven misinformation and propose counter-measures [32].

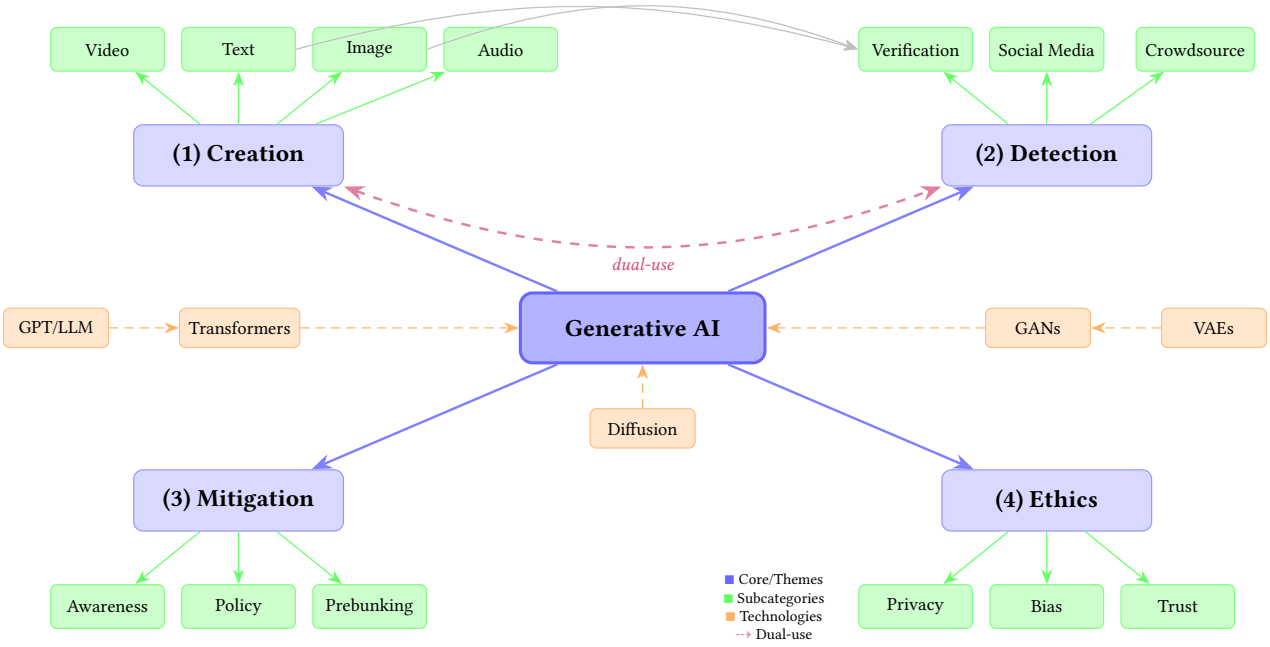


Figure 1: Structural overview of Generative AI's impact on Fake News. The dual-use nature (purple dashed arrow) illustrates how the same technologies enable both creation and detection of synthetic content.

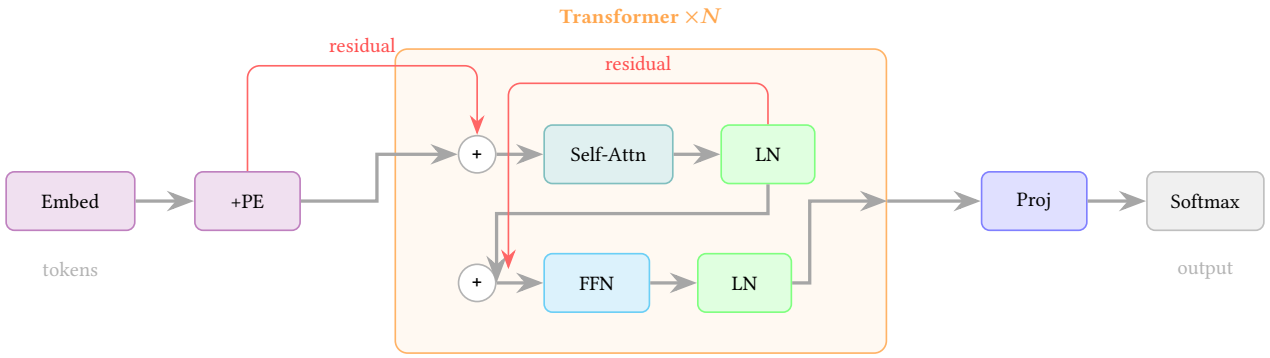


Figure 2: GPT architecture: tokens are embedded with positional encoding, processed through N transformer blocks (self-attention → layer norm → FFN → layer norm, with residual connections at each stage), then projected to output vocabulary.

3.3 Detection and Mitigation Strategies

Detection approaches have evolved from rule-based systems to LLM-based methods. Franco et al. (2025) propose using LLMs to prevent accidental sharing of misinformation [21], while Sallami and Aïmeur (2025) review prevention techniques beyond detection [47]. Research has also revealed gender biases in detection systems [46], prompting development of fairness frameworks [48]. Tommasel and Rodriguez (2025) address identifying misinformation spreaders in social networks [59].

3.4 Social Media and User Behavior

Godoy (2024) examines the moral intuitions of fake news spreaders [23], while Tintarev et al. (2024) study transparency in news recommendation systems [58]. Herder and Staring (2024) analyze "digital junkfood" consumption patterns on social media [27].

Critical to policy discussions is the "Transparency Penalty" identified by Nakano et al. (2026): disclosing AI authorship generally erodes perceived trustworthiness, competence, and warmth [44]. However, this effect is moderated by AI literacy—users with higher literacy are more tolerant of AI assistance and may even appreciate it. This finding complicates simple policy prescriptions for mandatory labeling; if labels universally reduce trust regardless of content quality, they may contribute to the "rational discounting" of evidence predicted by Ferrara's Generative AI Paradox [19].

3.5 Deepfakes and Multimodal Misinformation

Chun et al. (2024) find that older adults face challenges spotting deepfakes [10], while Verma (2024) shows that a single deceptive video could affect geopolitical relations [62].

The detection approach has shifted in 2026: single-modality detectors focusing on visual artifacts (warping, blending boundaries) or audio artifacts (robotic phrasing) are less effective against high-quality GenAI content. The frontier is *cross-modal consistency checking*. Hussain et al. (2026) demonstrate that the most reliable signal is temporal inconsistency between modalities—specifically, subtle desynchronization between lip movements and speech audio, or semantic mismatch between visual context and audio narrative [29]. Their Synchronization-Aware Feature Fusion (SAFF) and Cross-Modal Graph Attention Networks (CM-GAN) architectures achieve 98.76% accuracy on benchmarks like FaceForensics++ by explicitly modeling these cross-modal correlations.

Recent advances in Large Vision-Language Models (LVLMs) enable cross-modal semantic analysis [2], with multi-agent frameworks showing promise for complex verification tasks [53].

3.6 Ethical and Governance Considerations

Aïmeur (2025) addresses privacy concerns in generative AI [3], while Schrittwieser et al. (2024) examine trust and safety online [51]. The EU AI Act and similar frameworks are shaping deployment requirements for high-risk AI applications. Cena and Grasso (2025) explore users' mental models of conversational agents [9].

The 2025–2026 period marks the transition of C2PA from an emerging initiative to a global infrastructure standard. C2PA Specification v2.3 (December 2025) introduced support for live video streaming—addressing a major gap in real-time news verification—and manifests for unstructured text, extending provenance beyond media files to LLM outputs [11]. Adoption has scaled significantly, with Google integrating C2PA Assurance Level 2 into Pixel camera hardware and TikTok implementing mandatory labeling for realistic AI content. However, the Center for Democracy and Technology highlights a "validity gap": provenance proves origin (who signed it), not truth (is it factual?) [14]. Privacy concerns regarding embedded metadata (location, author identity) remain particularly acute for activists and whistleblowers.

4 Methodological Contributions: The Epistemic Security Experimental Pipeline

To operationalize the study of AI-mediated deception and quantify phenomena such as the "Generative AI Paradox" [19] and the "Agentic Shift," we introduce an experimental apparatus comprising two coupled components functioning as a closed-loop system.

4.1 RogueGPT: Controlled Stimulus Engine

RogueGPT[3] addresses the reproducibility challenge in misinformation research by replacing static datasets with a deterministic generation engine. Through a formal configuration schema, it enables the controlled injection of generative variables—specifically Model Architecture (M), Temperature (T), Style (S), and Format (F)—into the experimental design. This allows for the isolation of causal factors, represented formally as Stimulus $= f(M, T, S, F)$.

The engine ensures complete provenance by serializing the full generative context (system prompts, generation parameters) alongside the artifact, enabling retrospective analysis of deceptive strategies. Integration with OpenAI and Azure OpenAI APIs supports multi-model comparison across LLM versions, while manual entry capability allows incorporation of human-written control stimuli.

4.2 JudgeGPT: Psychometric Evaluation Platform

JudgeGPT[4] [38] serves as the measurement instrument for human epistemic resilience. Unlike binary classification tasks common in detection research, JudgeGPT employs continuous psychometric scales to capture the ambiguity of perception and the calibration of user confidence. Participants rate perceived origin (definitely human to definitely machine-generated), perceived veracity (definitely legitimate to definitely fake), and topic familiarity on graded scales.

By integrating response latency metrics and demographic profiling, the platform facilitates intersectional analysis of susceptibility. The architecture supports testing the efficacy of inoculation interventions [56], measuring whether prebunking warnings effectively engage analytical processing in real-time consumption environments.

4.3 Closed-Loop Verification System

The integration of these components via a unified document-oriented data topology allows for measurement of the "Perception-Accuracy Gap." Researchers first use RogueGPT to generate stimuli under controlled conditions, specifying model, style, format, and language parameters. Generated fragments with full provenance metadata are stored in a shared MongoDB collection. Participants access JudgeGPT, which retrieves fragments and presents them for evaluation. Responses are stored with links to fragment metadata, creating a dataset that enables precise attribution of perception effects to generation parameters.

This pipeline provides a standardized approach for measuring "Deceptive Potential," allowing quantification of threat escalation as generative models evolve.

4.4 Empirical Findings

Our studies using this apparatus are reported in companion publications [38, 40]. Data collection reveals that participants struggle to distinguish GPT-4 generated content from human-written text, with accuracy rates approaching chance levels for certain news styles [35, 38]. A perception-accuracy gap exists: increased suspicion does not improve detection accuracy, and asymmetric cognitive fatigue degrades fake detection by 10.2 percentage points under sustained exposure [39]. Demographic predictors show weak effects: political orientation is only weakly associated with detection performance [38], and age shows no significant correlation with

[3] https://github.com/aloth/RogueGPT

[4] https://github.com/aloth/JudgeGPT

accuracy, challenging the "digital native" assumption [35]. Familiarity with fake news is associated with higher suspicion but only negligibly with detection accuracy [39].

These findings align with broader research indicating that LLM-generated content is increasingly difficult to detect [24, 42].

5 Synthesis and Mitigation Strategies

Simon et al. (2023) discuss ethical considerations for AI in journalism, emphasizing the balance between benefits and risks [55]. Weisz et al. (2023) propose design principles for generative AI that prioritize safety [63].

5.1 Mitigation Approaches

Countering Generative AI's adverse effects requires strategies spanning technology, education, and policy:

Technological Approaches: Detection algorithms using the same LLMs employed for generation have demonstrated measurable effectiveness [21, 49]. However, the competition between generation and detection has become adversarial. Tahmasebi et al. (2026) demonstrate that many state-of-the-art detectors rely heavily on sentiment correlations—assuming, for instance, that fake news is negative or inflammatory—making them vulnerable to "sentiment attacks" that rewrite false claims to sound neutral or positive [57]. Their AdSent framework shows that such attacks can degrade detection performance (F1-score) by over 20%, confirming that adversaries are now optimizing latent features to traverse detector decision boundaries. This necessitates sentiment-agnostic training strategies that force models to learn veracity features independent of emotional tone.

Recent work on multimodal LLM-based detection systems, such as TRUST-VL [64] and agentic multi-persona frameworks [7], demonstrates improved accuracy through combining multiple reasoning perspectives. Tool-augmented agents using Monte Carlo Tree Search have achieved high performance in complex multimodal verification [13]. Future detectors should be adversarially aware and sentiment-agnostic, moving beyond stylistic analysis toward features that capture veracity rather than surface patterns.

Inoculation and Prebunking: Lewandowsky and Van Der Linden's (2021) inoculation theory has gained traction, emphasizing preemptive education to build resilience [33]. Spearing et al. (2025) provide empirical support in the GenAI context: "pre-emptive source discreditation"—warning users about the manipulative tactics of a source before exposure—is more effective than reactive debunking [56]. This is relevant for GenAI content, where the volume makes reactive fact-checking impractical. Our JudgeGPT platform offers opportunities to measure not just detection accuracy but also the efficacy of such inoculation interventions.

Provenance and Authenticity Infrastructure: Content authenticity initiatives provide cryptographic verification of content origin as an alternative to detection-based approaches. The C2PA standard has matured with v2.3 supporting live streaming and text manifests [11], while Google DeepMind's open-source SynthID Text provides a complementary watermarking layer using tournament-based token probability adjustment that resists modification yet remains invisible to humans [15]. This "defense-in-depth" approach—if metadata is stripped, the watermark may remain—aligns with our Origin Lens framework, which performs privacy-first on-device C2PA verification, combining cryptographic provenance with heuristic metadata analysis, watermark detection, and graded confidence indicators [36]. However, the "validity gap" remains: provenance proves origin, not truth [14]. Challenges persist around manifest stripping, analog-hole attacks, and privacy implications for whistleblowers.

Platform Design: Herder and Staring explore interface designs that help users manage social media consumption and avoid misinformation [27]. Transparency mechanisms and friction-inducing interventions can slow the reflexive sharing that accelerates misinformation spread.

Collaborative Measures: Shu et al. (2020) explore collaborative approaches involving governments, private sector, and civil society [54]. The development of shared benchmarks and evaluation frameworks enables progress in detection capabilities.

5.2 Stakeholder Landscape

Key stakeholders include academics developing detection algorithms [23], technology developers deploying AI detection systems, platforms enforcing misuse prevention, and policymakers creating regulations [3].

5.3 Open Research Directions

Several unresolved issues demand continued attention. *Adversarial robustness* remains a challenge: the arms race has moved to the feature level, with adversaries optimizing latent features to evade detectors, as demonstrated by sentiment attacks that degrade F1-scores by over 20% [12, 57]. *Multimodal challenges* require detection approaches that can analyze cross-modal semantic consistency and identify out-of-context manipulations [2, 64]. Global misinformation campaigns necessitate *cross-lingual detection* capabilities [20, 52], while ensuring *bias and fairness* in detection systems requires explicit attention to avoid creating new forms of harm [48]. The operationalization of Foreign Information Manipulation and Interference (FIMI) through multi-agent pipelines demands *behavioral-level detection* that analyzes Tactics, Techniques, and Procedures (TTPs) within frameworks like DISARM rather than isolated artifacts [7, 18, 60]. Finally, *digital ecosystem resilience* requires infrastructure-level approaches to information integrity, including provenance standards and platform design interventions [17, 36].

Future research must navigate technological innovations alongside broader societal, ethical, and psychological dimensions of this challenge.

6 Conclusion

This paper has examined how Generative AI has changed the disinformation landscape since our 2024 survey. We documented developments in LLM capabilities, multimodal misinformation, and the dual-use nature of these technologies for both generation and detection.

Our methodological contributions—*JudgeGPT* [38] and *RogueGPT*—provide an experimental pipeline for studying human perception of AI-generated news. Our companion studies using this pipeline

reveal that participants struggle to distinguish LLM-generated content from human-written text, with accuracy approaching chance levels for certain news styles [35, 38]. Our longitudinal expert survey found that specialists view large-scale text generation as posing systemic risks of "epistemic fragmentation" and "synthetic consensus"—findings now formalized in Ferrara's "Generative AI Paradox," which argues that the cost of verification has become prohibitively high compared to the cost of generation, rendering trust economically irrational [19]. Experts express skepticism toward purely technical detection tools, preferring provenance standards aligned with emerging "epistemic security" objectives [30, 40].

Several key insights emerge from this analysis. First, the threat has evolved beyond "fake news" to "Synthetic Reality"—a layered stack comprising synthetic content, identity, interaction, and institutions—requiring defenses that address each layer [19]. The dual-use nature of LLMs offers opportunities for "fighting fire with fire" approaches [21, 49], while multimodal misinformation requires detection approaches that analyze cross-modal semantic consistency [2, 64]. Prevention and prebunking strategies may prove more effective than reactive detection, as inoculation theory gains empirical support [33, 47]. Bias and fairness in detection systems require explicit attention to avoid creating new forms of harm [48]. Digital ecosystem resilience requires infrastructure-level interventions including content provenance standards and platform design changes aligned with epistemic security objectives [30, 36]. Finally, the rise of agentic AI systems introduces new vectors for scaled misinformation campaigns that demand behavioral-level detection and proactive governance [1, 60].

Our review suggests that purely technical countermeasures—such as watermarking or detection classifiers—face significant challenges due to the rapid adaptability of generative models. The competition between generation and detection capabilities continues, with current mitigation strategies struggling to keep pace.

Future research should explore proactive approaches including adversarial testing, provenance infrastructure, and governance frameworks. Our JudgeGPT-RogueGPT pipeline offers one foundation for investigating human perception of AI-generated content.

Addressing the adverse impacts of generative AI on information quality will require efforts combining technical safeguards, media literacy initiatives, platform accountability, and policy frameworks.

As our study continues, we invite experts to participate in our ongoing survey: https://github.com/aloth/verification-crisis.

References

[1] ACM Europe Technology Policy Committee. 2025. *Systemic Risks Associated with Agentic AI: A Policy Brief.* Technical Report. Association for Computing Machinery. https://www.acm.org/media-center/2025/october/agentic-ai-regulation

[2] Wei Ai, Yilong Tan, Yuntao Shou, Tao Meng, Haowen Chen, Zhixiong He, and Keqin Li. 2026. The Paradigm Shift: A Comprehensive Survey on Large Vision Language Models for Multimodal Fake News Detection. *Computer Science Review* 60 (2026), 100893. doi:10.1016/j.cosrev.2026.100893

[3] Esma Aïmeur. 2025. Privacy in AI. In *Human-Centered AI: An Illustrated Scientific Quest.* Springer Nature Switzerland, Cham, 537–572. doi:10.1007/978-3-031-61375-3_9

[4] Celeste Biever. 2023. ChatGPT broke the Turing test—the race is on for new ways to assess AI. *Nature* 619, 7971 (2023), 686–689. doi:10.1038/d41586-023-02361-7

[5] Samantha Bradshaw, Hannah Bailey, and Philip N. Howard. 2021. *Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation.* Technical Report. Computational Propaganda Research Project, Oxford Internet Institute, Oxford, UK. https://demtech.oii.ox.ac.uk/research/posts/industrialized-disinformation/

[6] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712 [cs] http://arxiv.org/abs/2303.12712

[7] Roopa Bukke, Soumya Pandey, Suraj Kumar, Soumi Chattopadhyay, and Chandranath Adak. 2025. Agentic Multi-Persona Framework for Evidence-Aware Fake News Detection. arXiv:2512.21039

[8] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun. 2023. A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. doi:10.48550/arXiv.2303.04226 arXiv:2303.04226 [cs]

[9] Federica Cena and Francesca Grasso. 2025. Exploring users' mental models of conversational agents: a systematic review. *Behaviour & Information Technology* 45, 10 (2025), 2189–2212. doi:10.1080/0144929X.2025.2573436

[10] Rachel Wan Ying Chun, Shilan Huang, Dion Hoe-Lian Goh, Chei Sian Lee, and Yin-Leng Theng. 2024. Can Seniors Spot Deepfakes? A Diary Study of Deepfake Identification Strategies. *Proceedings of the Association for Information Science and Technology* 61, 1 (2024), 877–879. doi:10.1002/pra2.1126

[11] Coalition for Content Provenance and Authenticity. 2025. C2PA Technical Specification Version 2.3. https://spec.c2pa.org/specifications/specifications/2.3/specs/C2PA_Specification.html

[12] Federico Cocchi, Lorenzo Baraldi, Samuele Poppi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2023. Unveiling the Impact of Image Transformations on Deepfake Detection: An Experimental Analysis. In *Image Analysis and Processing – ICIAP 2023* (Cham, 2023) *(Lecture Notes in Computer Science)*, Gian Luca Foresti, Andrea Fusiello, and Edwin Hancock (Eds.). Springer Nature Switzerland, 345–356. doi:10.1007/978-3-031-43153-1_29

[13] Xing Cui, Yueying Zou, Zekun Li, Peipei Li, Xinyuan Xu, Xuannan Liu, and Huaibo Huang. 2026. T^{2}Agent: A Tool-augmented Multimodal Misinformation Detection Agent with Monte Carlo Tree Search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. AAAI Press, 175–183. doi:10.1609/aaai.v40i1.36977

[14] Shruti Das. 2025. *The Promise and Risk of Digital Content Provenance.* Technical Report. Center for Democracy and Technology. https://cdt.org/insights/the-promise-and-risk-of-digital-content-provenance/

[15] Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, et al. 2024. Scalable watermarking for identifying large language model outputs. *Nature* 634, 8035 (2024), 818–823. doi:10.1038/s41586-024-08025-4

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Association for Computational Linguistics, Minneapolis, MN, USA, 4171–4186. doi:10.18653/v1/N19-1423

[17] Alina Fastowski and Gjergji Kasneci. 2025. Understanding Knowledge Drift in LLMs Through Misinformation. In *Discovering Drift Phenomena in Evolving Landscapes (DELTA 2024) (Lecture Notes in Computer Science).* Springer Nature Switzerland, Cham, Switzerland, 74–85. doi:10.1007/978-3-031-82346-6_5

[18] Emilio Ferrara. 2024. GenAI against humanity: nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science* 7, 1 (2024), 549–569. doi:10.1007/s42001-024-00250-1

[19] Emilio Ferrara. 2026. The Generative AI Paradox: GenAI and the Erosion of Trust, the Corrosion of Information Verification, and the Demise of Truth. *Future Internet* 18, 2 (2026), 73. doi:10.3390/fi18020073

[20] Rafael Martins Frade, Rrubaa Panchendrarajan, and Arkaitz Zubiaga. 2026. MultiCaption: Detecting disinformation using multilingual visual claims. arXiv:2601.11220

[21] Mirko Franco, Valentin Grimm, and Eelco Herder. 2025. Preventing Accidental Sharing of Misinformation Using Large Language Models. In *Proceedings of the 2025 International Conference on Information Technology for Social Good (GoodIT '25).* ACM, New York, NY, USA, 244–252. doi:10.1145/3748699.3749798

[22] Gartner. 2024. Gartner Identifies the Top 10 Strategic Technology Trends for 2025. Press release. https://www.gartner.com/en/newsroom/press-releases/2024-10-21-gartner-identifies-the-top-10-strategic-technology-trends-for-2025

[23] Daniela Godoy. 2024. On the moral intuitions of fake news spreaders. In *Companion Proceedings of the 16th ACM Web Science Conference (WebSci Companion '24).* ACM, New York, NY, USA, 65–68. doi:10.1145/3630744.3663605

[24] Dion Hoe-Lian Goh, Jonathan Pan, and Chei Sian Lee. 2024. Humans Versus Machines: A Deepfake Detection Faceoff. *Proceedings of the Association for Information Science and Technology* 61, 1 (2024), 917–919. doi:10.1002/pra2.1139

[25] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27 (NIPS 2014).* Curran Associates, Inc. arXiv:1406.2661

[26] Roberto Gozalo-Brizuela and Eduardo C. Garrido-Merchan. 2023. ChatGPT is not all you need. A State of the Art Review of large Generative AI models. doi:10.48550/arXiv.2301.04655 arXiv:2301.04655 [cs]

[27] Eelco Herder and Jouke Staring. 2024. Digital Junkfood on Social Media: To Each Their Own Poison. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media (HT '24)*. ACM, New York, NY, USA, 126–135. doi:10.1145/3648188.3678163

[28] Redwan Hussain, Mizanur Rahman, and Prithwiraj Bhattacharjee. 2026. Toward Generalized Detection of Synthetic Media: Limitations, Challenges, and the Path to Multimodal Solutions. In *Proceedings of the Seventh International Conference on Trends in Computational and Cognitive Engineering (TCCE 2025) (Lecture Notes in Networks and Systems)*. Springer, 80–91. doi:10.1007/978-3-032-33805-1_8

[29] Tarak Hussain, B. Tirapathi Reddy, Kondaveti Phanindra, Sailaja Terumalasetti, and Ghufran Ahmad Khan. 2026. Decoding Deception: State-of-the-art Approaches to Deep Fake Detection. *Frontiers in Big Data* 8 (Jan. 2026), 1670833. doi:10.3389/fdata.2025.1670833

[30] International Panel on the Information Environment (IPIE). 2025. *Expert Survey on the Global Information Environment 2025: Safeguarding Epistemic Security*. Technical Report. International Panel on the Information Environment. https://www.ipie.info/research/sfp2025-4

[31] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations (ICLR 2014)*. arXiv:1312.6114

[32] Sanjeev Kumar, Siva Sai, Vinay Chamola, Aanchal Gaur, Chitwan Agarwal, Kaizhu Huang, and Amir Hussain. 2025. Peeping into the Future: Understanding and Combating Generative AI-Based Fake News. *Cognitive Computation* 17, 3, Article 103 (2025). doi:10.1007/s12559-025-10457-7

[33] Stephan Lewandowsky and Sander van der Linden. 2021. Countering Misinformation and Fake News Through Inoculation and Prebunking. *European Review of Social Psychology* 32, 2 (2021), 348–384. doi:10.1080/10463283.2021.1876983

[34] Alexander Loth. 2021. *Decisively Digital: From Creating a Culture to Designing Strategy*. John Wiley & Sons, Inc., Hoboken, NJ, USA. https://www.wiley.com/en-us/Decisively+Digital%3A+From+Creating+a+Culture+to+Designing+Strategy-p-9781119737285

[35] Alexander Loth. 2026. The Indistinguishability Threshold: Measuring Cognitive Vulnerabilities to AI-Generated Disinformation. In *Companion Publication of the 18th ACM Web Science Conference (WebSci Companion '26), PhD Symposium* (Braunschweig, Germany). ACM, New York, NY, USA, 79–82. doi:10.1145/3795513.3807421

[36] Alexander Loth, Dominique Conceicao Rosario, Peter Ebinger, Martin Kappes, and Marc-Oliver Pahl. 2026. Origin Lens: Reclaiming Trust on the AI-Mediated Web Through On-Device Image Provenance Verification. In *Companion Publication of the 2026 18th ACM Web Science Conference (WebSci Companion '26)* (Braunschweig, Germany). ACM, New York, NY, USA, 1–5. doi:10.1145/3795513.3806658

[37] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2024. Blessing or Curse? A Survey on the Impact of Generative AI on Fake News. doi:10.48550/arXiv.2404.03021 arXiv:2404.03021 [cs.CL]

[38] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2026. Eroding the Truth-Default: A Causal Analysis of Human Susceptibility to Foundation Model Hallucinations and Disinformation in the Wild. In *Companion Proceedings of the ACM Web Conference 2026 (WWW Companion '26)* (Dubai, United Arab Emirates). ACM, New York, NY, USA, 1125–1132. doi:10.1145/3774905.3795832 arXiv:2601.22871 [cs.CY]

[39] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2026. Interrupting the Chain: Human Perception of AI-Generated Disinformation Through a Kill Chain Lens. In *INFORMATIK 2026: Digitale Resilienz (Lecture Notes in Informatics (LNI), Vol. P-384)*. Gesellschaft für Informatik, Bonn, Germany, 321–330. doi:10.18420/inf2026_22 arXiv:2608.21389 [cs.CY]

[40] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2026. The Verification Crisis: Expert Perceptions of GenAI Disinformation and the Case for Reproducible Provenance. In *Companion Proceedings of the ACM Web Conference 2026 (WWW Companion '26)* (Dubai, United Arab Emirates). ACM, New York, NY, USA, 980–988. doi:10.1145/3774905.3795484 arXiv:2602.02100 [cs.CY]

[41] Shuming Ma, Hongyu Wang, Lingxiao Ma, Lei Wang, Wenhui Wang, Shaohan Huang, Li Dong, Ruiping Wang, Jilong Xue, and Furu Wei. 2024. The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits. doi:10.48550/arXiv.2402.17764 arXiv:2402.17764 [cs]

[42] Antonis Maronikolakis, Hinrich Schütze, and Mark Stevenson. 2021. Identifying Automatically Generated Headlines using Transformers. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda (NLP4IF)*. Association for Computational Linguistics, 1–6. doi:10.18653/v1/2021.nlp4if-1.1

[43] Raj Gaurav Maurya, Vaibhav Shukla, Raj Abhijit Dandekar, Rajat Dandekar, and Sreedath Panat. 2025. Simulating Misinformation Propagation in Social Networks using Large Language Models. In *Proceedings of the CIKM Workshop on Large-Scale Agent-based Social Simulations (LASS)*. arXiv:2511.10384

[44] Hiroki Nakano, Jo Takezawa, Fabrice Matulic, Chi-Lan Yang, and Koji Yatani. 2026. Understanding Reader Perception Shifts upon Disclosure of AI Authorship. In *Proceedings of the 31st International Conference on Intelligent User Interfaces (IUI '26)*. ACM, New York, NY, USA, 2131–2146. doi:10.1145/3742413.3789076

[45] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. *Improving Language Understanding by Generative Pre-Training*. Technical Report. OpenAI. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[46] Dorsaf Sallami and Esma Aïmeur. 2025. Does Gender Matter? Examining and Mitigating Gender Bias in Fake News Detection. In *Foundations and Practice of Security (FPS 2024) (Lecture Notes in Computer Science)*. Springer Nature Switzerland, Cham, Switzerland, 249–266. doi:10.1007/978-3-031-87499-4_16

[47] Dorsaf Sallami and Esma Aïmeur. 2025. Exploring beyond detection: a review on fake news prevention and mitigation techniques. *Journal of Computational Social Science* 8, 1, Article 23 (2025). doi:10.1007/s42001-024-00351-x

[48] Dorsaf Sallami and Esma Aïmeur. 2025. Fairframe: a fairness framework for bias detection and mitigation in news. *AI and Ethics* 5, 3 (2025), 2467–2483. doi:10.1007/s43681-024-00568-6

[49] Dorsaf Sallami, Yuan-Chen Chang, and Esma Aïmeur. 2024. From Deception to Detection: The Dual Roles of Large Language Models in Fake News. arXiv:2409.17416

[50] Luca Sandrini and Robert Somogyi. 2023. Generative AI and deceptive news consumption. *Economics Letters* 232 (2023), 111317. doi:10.1016/j.econlet.2023.111317

[51] Sebastian Schrittwieser, Andreas Ekelhart, Esther Seidl, and Edgar Weippl. 2024. Safe or Scam? An Empirical Simulation Study on Trust Indicators in Online Shopping. In *Proceedings of the 21st International Conference on Security and Cryptography (SECRYPT 2024)*. SciTePress, 560–567. doi:10.5220/0012852600003767

[52] Mina Schütz, Jaqueline Böck, Medina Andresel, Armin Kirchknopf, Daria Liakhovets, Djordje Slijepčević, and Alexander Schindler. 2022. AIT_FHSTP at CheckThat! 2022: Cross-Lingual Fake News Detection with a Large Pre-Trained Transformer. In *Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum (CEUR Workshop Proceedings, Vol. 3180)*. CEUR-WS.org, 660–670. https://ceur-ws.org/Vol-3180/paper-53.pdf

[53] Mir Nafis Sharear Shopnil, Sharad Duwal, Abhishek Tyagi, and Adiba Mahbub Proma. 2025. MIRAGE: Agentic Framework for Multimodal Misinformation Detection with Web-Grounded Reasoning. arXiv:2510.17590v1

[54] Kai Shu, Amrita Bhattacharjee, Faisal Alatawi, Tahora H. Nazer, Kaize Ding, Mansooreh Karami, and Huan Liu. 2020. Combating disinformation in a social media age. *WIREs Data Mining and Knowledge Discovery* 10, 6, Article e1385 (2020). doi:10.1002/widm.1385

[55] Felix M. Simon, Sacha Altay, and Hugo Mercier. 2023. Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. *Harvard Kennedy School (HKS) Misinformation Review* 4, 5 (2023). doi:10.37016/mr-2020-127

[56] Emily R. Spearing, Constantina I. Gile, Amy L. Fogwill, Toby Prike, Briony Swire-Thompson, Stephan Lewandowsky, and Ullrich K. H. Ecker. 2025. Countering AI-generated misinformation with pre-emptive source discreditation and debunking. *Royal Society Open Science* 12, 6 (June 2025), 242148. doi:10.1098/rsos.242148

[57] Sahar Tahmasebi, Eric Müller-Budack, and Ralph Ewerth. 2026. Robust Fake News Detection using Large Language Models under Adversarial Sentiment Attacks. In *Proceedings of the ACM Web Conference 2026 (WWW '26)* (Dubai, United Arab Emirates). ACM, New York, NY, USA, 1717–1726. doi:10.1145/3774904.3792606

[58] Nava Tintarev, Bart P. Knijnenburg, and Martijn C. Willemsen. 2024. Measuring the benefit of increased transparency and control in news recommendation. *AI Magazine* 45, 2 (2024), 212–226. doi:10.1002/aaai.12171

[59] Antonela Tommasel and Juan Manuel Rodriguez. 2025. Countering the Spread: An Approach to Identify Misinformation Spreaders in Social Media. *IEEE Transactions on Computational Social Systems* 12, 5 (2025), 3403–3415. doi:10.1109/TCSS.2025.3550029

[60] Kevin Tseng, Juan Carlos Toledano, Bart De Clerck, Yuliia Dukach, and Phil Tinn. 2026. An Agentic Operationalization of DISARM for FIMI Investigation on Social Media. arXiv:2601.15109 Presented at the International Conference on Military Communication and Information Systems (ICMCIS), Bath, UK, May 2026.

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*. Curran Associates, Inc. arXiv:1706.03762

[62] Nitin Verma. 2024. "One Video Could Start a War": A Qualitative Interview Study of Public Perceptions of Deepfake Technology. *Proceedings of the Association for Information Science and Technology* 61, 1 (2024), 374–385. doi:10.1002/pra2.1035

[63] Justin D. Weisz, Michael Muller, Jessica He, and Stephanie Houde. 2023. Toward General Design Principles for Generative AI Applications. arXiv:2301.05578 [cs] http://arxiv.org/abs/2301.05578

[64] Zehong Yan, Peng Qi, Wynne Hsu, and Mong Li Lee. 2025. TRUST-VL: An Explainable News Assistant for General Multimodal Misinformation Detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 5588–5604. doi:10.18653/v1/2025.emnlp-main.284