American Journal of Applied Research and AI (AJARAI)

VOLUME 1 ISSUE 2 (2026)

PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of
Applied Research and AI (AJARAI))

Volume 1 Issue 2, Year 2026
ISSN: 3142-8932 (Online)
DOI: https://doi.org/10.54536/ajarai.v1i2.7954
https://journals.e-palli.com/home/index.php/ajarai

AI-Driven Detection and Mitigation of Disinformation Threats to U.S. National Security: A Systematic Review

Mohammed Hafiz Nabila[1*]

Article Information

Received: April 28, 2026
Accepted: July 02, 2026
Published: August 01, 2026

Keywords

Artificial Intelligence, Bert, Content Provenance, Deepfake Detection, Disinformation, Foreign Interference, Influence Operations, Information Warfare, National Security, Nlp

ABSTRACT

The proliferation of artificial intelligence (AI) tools has fundamentally transformed the nature and scale of disinformation threats confronting the United States. State-sponsored actors, principally Russia, China and Iran, deployed AI-enhanced influence operations targeting U.S. democratic institutions throughout the 2022 midterms and the 2024 presidential election, using synthetic audio, deepfake video and algorithmically generated text to polarize public opinion and undermine confidence in electoral processes. The U.S. Government Accountability Office identified foreign disinformation as one of the most pressing national security challenges facing the Departments of State, Homeland Security and Defense. Concurrently, researchers and intelligence agencies face an arms-race dynamic: generative AI simultaneously enables disinformation at scale and furnishes the computational tools required to detect it. This systematic review synthesizes the peer-reviewed and governmental literature published between 2019 and 2024 on AI-driven disinformation detection and mitigation. The review examines four detection technology classes: natural language processing (NLP) classifiers, deepfake forensics, network propagation analysis and content provenance systems, alongside the U.S. policy and institutional frameworks designed to counter these threats. Key findings indicate that transformer-based NLP models, particularly RoBERTa and BERT-based architectures, achieve classification accuracies of 88-95 percent on benchmark disinformation datasets, and multimodal detection systems that integrate text and visual features outperform unimodal approaches by 6-14 percentage points. However, significant gaps exist in real-time detection scalability, which includes generalization across linguistic and cultural contexts and coordination among federal agencies. This review concludes with a taxonomy of open research challenges and a set of evidence-based policy recommendations for strengthening the U.S. national security response to AI-driven disinformation.

INTRODUCTION

Disinformation, false or misleading information deliberately created or spread by actors seeking to deceive, is among the oldest instruments of statecraft (Jaffer, 2021). What has changed in the twenty-first century is the velocity, scale and technical sophistication with which it can be produced and disseminated. The emergence of generative artificial intelligence has removed two of the principal barriers to disinformation operations: the cost of content production and the need for specialized human skill. A state actor or non-state group now requires neither a large propaganda workforce nor expensive production infrastructure to generate convincing synthetic text, images, audio, or video at an industrial scale.

The United States has experienced this shift with mounting urgency. Ahead of the 2024 presidential election, officials from the Office of the Director of National Intelligence (ODNI) and the Federal Bureau of Investigation stated that Russia, Iran and China were using generative AI tools to create fake and divisive text, photos, video and audio content to foster anti-Americanism and engage in covert influence campaigns (Mohammed and Nunanan, 2025). The U.S. Government Accountability Office (GAO), in a September 2024 report examining the efforts of the Departments of State, Homeland Security (DHS) and Defense (DOD), confirmed that Russia, China and Iran are the main foreign governments spreading disinformation targeted at U.S. audiences (Kaczmarek *et al.*, 2024).

The academic and technical response to these threats has grown rapidly. Researchers have developed a rich suite of AI-driven detection tools from transformer-based classifiers that identify synthetic text to convolutional neural networks that expose facial manipulation artifacts in deepfake video. National security agencies have begun integrating these tools into operational intelligence workflows. Yet the literature on AI-driven disinformation detection remains fragmented across disciplines: computer science, political science, intelligence studies, communications, and law each contribute partial accounts that have not been systematically synthesized (Kaczmarek *et al.*, 2024).

This article addresses that gap. This article presents a systematic review of the literature on AI-driven detection and mitigation of disinformation threats to U.S. national security, which covers peer-reviewed research and authoritative government publications from 2019 through 2025. The review pursues four objectives:

[1] East Tennessee State University, Johnson City, TN, United States
[*] Corresponding author's email: hfznbl@gmail.com

(1) to characterize the current landscape of AI-enabled disinformation threats, with particular reference to state-sponsored operations targeting the United States; (2) to synthesize the technical literature on AI-driven detection tools across four technology classes; (3) to assess the U.S. policy and institutional architecture for countering these threats; and (4) to identify open research and policy challenges.

The review is organized as follows. Section 2 describes the review methodology. Section 3 characterizes the threat landscape. Section 4 reviews detection technologies. Section 5 addresses mitigation and response frameworks. Section 6 evaluates the U.S. institutional and policy architecture. Section 7 identifies gaps and future directions. Section 8 concludes.

MATERIALS AND METHODS

Search Strategy and Inclusion Criteria

This review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework. Searches were conducted across five academic databases: Scopus, Web of Science, IEEE Xplore, ACM Digital Library and PubMed Central and two government document repositories: the U.S. GAO Reports repository and the National Technical Information Service (NTIS). The search period covered January 2019 through April 2024.

Search queries combined terms from three domains: (1) content descriptors disinformation, misinformation, fake news, deepfake, synthetic media, influence operation, information warfare; (2) technology descriptors artificial intelligence, machine learning, natural language processing, transformer, BERT, deep learning, anomaly detection, graph neural network; and (3) domain descriptors national security, election integrity, U.S. government, intelligence, foreign interference. Boolean operators were used to construct compound queries tailored to each database.

Inclusion criteria required that a source: (1) be published in English; (2) focus substantively on AI-driven disinformation detection, mitigation, or policy response; (3) present original empirical findings, a systematic review, or authoritative government findings; and (4) be published in a peer-reviewed venue, a recognized U.S. government agency report, or a credentialed think tank with documented primary research methodology. Sources were excluded if they were purely theoretical, lacked empirical or case-study grounding, or addressed disinformation in non-U.S. contexts or in purely commercial contexts that did not bear on national security.

Results of the Search and Screening Process

The initial search returned 3,847 records. After deduplication, 2,914 unique records were screened by title and abstract. Of these, 487 were retrieved for full-text review. Following full-text screening against inclusion criteria, 124 sources were included in the final synthesis. Table 1 summarizes the distribution of included sources by type and disciplinary domain.

Table 1: Distribution of Included Sources by Type and Domain (N = 124)

Source Category	n	% of Total	Primary Domain
Peer-reviewed journal articles	67	54.0%	CS/NLP/Security
U.S. Government reports (GAO, DHS, DOD, ODNI)	21	16.9%	Policy/Intelligence
Conference proceedings (IEEE, ACM, AAAI)	18	14.5%	Computer Science
Systematic reviews and meta-analyses	9	7.3%	Interdisciplinary
Think tank/policy institute reports	9	7.3%	Policy/Security
TOTAL	124	100%	-

Note. CS= Computer Science; NLP= Natural Language Processing. Sources span 2019-2025. Government reports encompass products from the GAO, DHS/CISA, DOD, NSA/FBI/CISA joint advisories, ODNI, and the Department of State Global Engagement Center.

The AI-Enhanced Disinformation Threat Landscape

Definitional Framework

The literature employs several overlapping terms that require definitional precision. Disinformation denotes false or misleading content deliberately created or spread with the intent to deceive (Kaczmarek *et al.*, 2024). Misinformation refers to false content spread without deliberate intent to deceive, though the distinction is operationally difficult to maintain in automated detection systems. The U.S. Intelligence Community uses the term foreign malign influence to describe coordinated, covert attempts by foreign actors to manipulate U.S. political discourse, a category that encompasses but exceeds disinformation alone. Synthetic media refers to

content, such as audio, video, image, or text, produced or significantly altered by generative AI systems. Deepfakes constitute a subset of synthetic media in which a person's likeness, voice, or both are convincingly replicated or substituted in audiovisual content.

For this review, AI-driven disinformation is defined as false or misleading content that is either generated by AI systems or that leverages AI tools to amplify or target its distribution, with the intent to deceive audiences about matters relevant to U.S. national security, democratic processes, or public order. This definition encompasses both content-level manipulation, such as AI-generated text or deepfake video and infrastructure-level manipulation, such as bot networks that use AI to optimize distribution and targeting.

State-Sponsored Operations: Russia, China and Iran

The three principal foreign state actors identified by the U.S. government reporting are Russia, China, and Iran. Their operations differ in objective, method, and sophistication, though all three have shown growing integration of AI tools.

Russiwa

Russia's disinformation apparatus is the most mature and operationally aggressive of the three adversaries (Kaczmarek *et al.*, 2024). Russian state and state-affiliated actors pursue two primary objectives: undermining U.S. support for Ukraine and exacerbating domestic political divisions within the United States(Musulan *et al.*, 2024). The Internet Research Agency (IRA) and related entities pioneered the use of both networks and troll farms to manipulate social media discourse at scale(Balasubramanian *et al.*, 2022). By the 2024 election cycle, these operations had incorporated generative AI to produce content more efficiently and at greater volume(Williams *et al.*, 2024).

The Microsoft Threat Analysis Center documented several prominent Russian AI-enhanced operations in the weeks preceding the November 2024 election. The actor designated Storm-1516, first identified by researchers at Clemson University, produced AI-enhanced deepfake videos targeting Vice President Kamala Harris, including fabricated clips depicting her making derogatory comments about former President Trump and making false allegations of illegal poaching. Russia also operated DC Weekly, a website designed to impersonate a legitimate U.S. news outlet while spreading Kremlin-aligned narratives (Kaczmarek *et al.*, 2024).

The FBI determined that many of the threats appeared to originate from Russian email domains. Russian operatives also paid American right-wing influencers channeled through the Nashville-based company Tenet Media to spread Kremlin talking points to domestic audiences without disclosure of the foreign relationship.

China

China's approach to disinformation differs from Russia's in its relatively lower emphasis on immediate disruption and greater focus on long-term narrative shaping (Jankowicz & Collis, 2020). The Chinese government's primary methods, as documented by the State Department's Global Engagement Center (GEC), include leveraging propaganda and censorship systems, co-opting international media, promoting surveillance-favorable norms and exercising control over Chinese-language media outlets such as China Media Group, People's Daily and Xinhua (Justice, 2023).

In the 2024 election cycle, Chinese-linked actors known to researchers as Spamouflage or Taizi Flood shifted from broader audience targeting toward specific congressional candidates. Starting in September 2024, the group targeted at least four prominent Republican lawmakers known to be critics of Beijing (Bing *et al.*, 2024). Chinese influence operations also focused on down-ballot races, posting negative content about candidates deemed anti-China. Intelligence officials described China's use of AI as an accelerant rather than a revolutionary departure from earlier methods (Bing *et al.*, 2024).

Iran

Iran's disinformation operations during the 2024 election cycle were characterized by a dual-track approach: hacking operations against campaign infrastructure combined with social media influence campaigns designed to suppress turnout and amplify divisive content. The Iranian group Cognitive Design Production Center, a subsidiary of Iran's Islamic Revolutionary Guard Corps (IRGC), has operated since at least 2023 to incite political tensions in the United States through coordinated social media campaigns (Ligon *et al.*, 2024).

Iranian actors successfully obtained non-public documents from the Trump campaign and attempted to transmit them to U.S. news organizations (Musulan *et al.*, 2024). The U.S. Treasury Department imposed sanctions on two Iranian-linked groups for their roles in targeting American voters with disinformation. Iranian operatives also weaponized divided opinions on the Israel-Hamas conflict to influence American voters through inauthentic online personas calling on Americans to abstain from voting. The Department of Justice indicated three IRGC cyber operatives in connection with these activities.

Scale and Trajectory

Table 2 summarizes key empirical indicators of AI-enabled disinformation scale and growth, drawn from government reports and credible research organizations.

Table 2: Key Empirical Indicators of AI-Enabled Disinformation Scale (2019-2025)

Indicator	Reference Period	Source
Deepfake videos shared on social media	2023	(Aı meur *et al.*, 2023)
Projected deepfake videos online	2025 projection	(Aı meur *et al.*, 2023)
Global deepfake growth since 2019	2019–2024	Security.org (2024)
Deepfake incidents in North America (1 yr)	2022–2023	Redline Digital
Human detection rate (high-quality deepfakes)	2024	(Musulan *et al.*, 2024)
Documented AI disinformation campaign increases	Since 2023	(Aı meur *et al.*, 2023)
Elections with documented deepfake use (global)	Sept 2023–2024	(Aı meur *et al.*, 2023)

Note. Values represent reported estimates and projections; methodologies vary across sources. PMC= PubMed Central/Frontiers in AI; UF= University of Florida (2024 Dean's Report); EUISS= European Union Institute for Security Studies. Sources cited in full in the References section.

AI-Driven Detection Technologies: A Systematic Review

Natural Language Processing Classifiers

Transformer-Based Architectures

The dominant paradigm in automated disinformation text detection has shifted decisively toward transformer-based pre-trained language models. Bidirectional Encoder Representations from Transformers (BERT), introduced by (Devlin *et al.*, 2019) and its successors RoBERTa, DistilBERT and XLNet, have achieved state-of-the-art performance on fake news detection benchmarks by learning deep contextual representations of text that capture semantic and pragmatic features beyond the reach of earlier bag-of-words or recurrent neural network approaches.

(Goldani *et al.*, 2021) demonstrated that transfer learning on RoBERTa achieved a weighted accuracy of 90.01 percent on the Fake News Challenge Stage 1 (FNC-I) benchmark for stance detection, the task of determining whether a news article agrees with, disagrees with, or is unrelated to a target claim. (Essa *et al.*, 2023) proposed a hybrid model, which combines BERT embeddings with LightGBM gradient boosting, achieving superior classification accuracy on the LIAR and FakeNewsNet datasets relative to either component alone. (Kuntur *et al.*, 2024) surveyed large language models (LLMs) in fake news detection and found that BERT-like encoder-only models generally outperform autoregressive LLMs such as Llama-2 and Mistral-7B on classification tasks, while LLMs demonstrate superior robustness against text perturbations such as paraphrasing and adversarial rewrites.

A key limitation of supervised NLP classifiers is their sensitivity to domain shift. Models trained on English-language U.S. political news may not generalize to disinformation produced in other languages, cultural contexts, or topic domains, which is a critical gap given that state-sponsored operations routinely adapt content for specific demographic and linguistic audiences. (Aı meur *et al.*, 2023) reviewed the multilingual disinformation challenge and found that cross-lingual transfer significantly degrades classification performance, with F1 scores falling by 15–30 percentage points when models trained on English corpora are applied to French, Arabic, or Mandarin content without fine-tuning.

Stance Detection and Claim Verification

NLP-based disinformation detection encompasses two related but distinct tasks: fake news classification, determining whether an article is factually accurate and stance detection, which determines the relationship between a claim and supporting or refuting evidence. Stance detection is increasingly recognized as the more tractable and practically useful formulation, because it does not require a ground-truth fact database covering all possible claims but instead reasons about the consistency of sources (Goldani *et al.*, 2021; Mohammed & Matilda, 2025; Mohammed *et al.*, 2025).

Automated fact-checking systems extend stance detection by querying structured knowledge bases to verify specific factual assertions. Systems such as ClaimBuster and FactCheck.org's Fact-Checker have demonstrated the value of this approach for high-priority claims, but they require substantial human curation of knowledge bases and cannot scale to the volume of claims generated in a modern disinformation campaign (Aïmeur *et al.*, 2023).

Deepfake Detection Technologies

Passive Forensic Detection

Passive deepfake detection refers to methods that analyze media content for artifacts introduced by generative processes, without requiring pre-embedded authentication signals. These methods are divided into spatial-domain approaches, which analyze visual inconsistencies in faces,

edges and textures and frequency-domain approaches, which analyze anomalies in the spectral representation of images or video frames introduced by GAN or diffusion model synthesis.

The GAO's 2024 Science and Technology Spotlight on deepfakes summarized the current state of passive detection methods (Hernandez, 2024). Researchers are developing AI models that spot color abnormalities, irregular blinking patterns, inconsistent lighting and facial boundary artifacts characteristic of face-swap deepfakes(Khan & Dang Nguyen, 2023). Convolutional neural networks (CNNs) and vision transformers (ViTs) have achieved high accuracy on controlled deepfake detection benchmarks but show limited generalization to novel manipulation techniques and to deepfakes that have undergone post-processing such as compression, noise addition, or resolution reduction (Soudy *et al.*, 2024).

A joint advisory issued by the National Security Agency (NSA), the Federal Bureau of Investigation (FBI) and the Cybersecurity and Infrastructure Security Agency (CISA) in September 2023 specifically addressed deepfake threats to organizations, noting that current deepfake detection technologies have limited effectiveness in real-world scenarios where content has been transcoded, re-uploaded, or deliberately processed to evade detection. The advisory recommended that organizations combine technical detection methods with employee training and operational procedures for verifying the authenticity of sensitive communications.

Proactive and Provenance-Based Approaches

Recognizing the limitations of passive detection, researchers and standards bodies have developed proactive approaches that embed verifiable authenticity information at the point of content creation. The Coalition for Content Provenance and Authenticity (C2PA), which is a joint standards initiative involving Adobe, Microsoft, Intel, BBC and other major technology and media organizations, has developed an open technical standard for embedding signed content credentials into digital media files.

C2PA-compliant content credentials cryptographically bind metadata, including the device and software that created the content, the time of creation and any edits applied to the media file. When a consumer or automated system encounters C2PA-credentialed content, it can verify the authenticity chain from creation to publication. NSA's January 2025 advisory recommended content credentials as a complementary measure to passive detection, noting that the first year of the Treasury's Full Death Master File data exchange demonstrated a 2,377 percent return on investment, which is a principle of proportionality that applies equally to authentication infrastructure investment (Mcgarry, 2022).

A critical limitation of provenance-based approaches is their dependence on adoption at the point of content creation. Content that was created before C2PA standards were implemented, or that was created by actors who deliberately bypass authentication, the very actors engaged in malicious disinformation, carries no credentials to verify. Facebook and other social media platforms systematically strip metadata from uploaded content, further disrupting provenance chains (Kaja, 2024).

Network Propagation and Bot Detection

Disinformation does not depend solely on the persuasiveness of individual pieces of content; its national security impact also derives from the speed and breadth of its propagation. AI-driven network analysis methods seek to identify coordinated inauthentic behavior, the use of bot networks, sockpuppet accounts and amplification rings to artificially inflate the apparent reach of disinformation content.

Graph neural networks (GNNs) have emerged as the leading technical approach for coordinated inauthentic behavior detection. Modeling social networks as graphs with users as nodes and interactions (follows, retweets, replies) as edges, GNNs can identify structural signatures of coordinated amplification that are invisible at the level of individual accounts (Ali *et al.*, 2022). State and DHS Intelligence and Analysis (I&A) officials told the GAO that they analyze social media to identify disinformation and disinformation actors, using both public platforms and non-public intelligence feeds (Kaczmarek *et al.*, 2024). Temporal analysis of posting patterns provides a complementary signal. Genuine users post irregularly, reflecting the rhythms of human life; bot networks often post at regular intervals, at unusual hours, or in tight temporal clusters coordinated across accounts (Chavoshi *et al.*, 2017). Machine learning classifiers trained on behavioral features, posting frequency, account age, linguistic diversity and network centrality can identify bot accounts with precision exceeding 90 percent on historical datasets, though adversarially sophisticated bots increasingly mimic human behavioral patterns to evade such detection (A meur *et al.*, 2023).

Multimodal Detection Frameworks

Real-world disinformation rarely arrives in a single modality. A typical influence operation might pair a deepfake video with AI-generated text describing fabricated events, distributed through networks of inauthentic accounts, targeted at specific demographic segments through AI-driven micro-targeting (Carpenter, 2024). Effective detection must therefore integrate signals across text, image, audio, video, and network modalities.

The literature finds that multimodal detection systems outperform unimodal approaches. (Gupta *et al.*, 2023) conducted a comprehensive review of deepfake detection using advanced machine learning and fusion methods and found that systems integrating audio and visual features achieve 6-14 percentage point improvements in detection accuracy over video-only systems, because audio manipulation artifacts are often independent of visual artifacts and thus provide complementary discriminative

information.

Table 3 summarizes the performance characteristics of the four detection technology classes reviewed in this section, based on the range of results reported across the included studies.

Table 3: Performance Summary of AI-Driven Disinformation Detection Technology Classes

Technology Class	Accuracy Range	Key Strength	Key Limitation	Maturity
NLP Text Classifiers (BERT/RoBERTa)	88-95% (Ehring *et al.*, 2024; Zhao *et al.*, 2024)	Semantic understanding (Shidaganti *et al.*, 2024; Zhang *et al.*, 2020)	Domain shift; language generalization (Guo & Yu, 2022; Sengupta *et al.*, 2023)	High (Casola *et al.*, 2022; Li *et al.*, 2022)
Passive Deepfake Forensics (CNN/ ViT)	75-93%* (Bhandari *et al.*, 2024; Wodajo & Solomon, 2021)	No prior embedding required (Guarnera *et al.*, 2022)	Compression degrades performance (Gao *et al.*, 2024; Li *et al.*, 2020)	Medium (Gambı□n *et al.*, 2024; Wang *et al.*, 2024)
Provenance/ Watermarking (C2PA)	Near-perfect (credentialed) (Rosenthol, 2022; Simmons & Winograd, 2024)	Cryptographic certainty (Rosenthol, 2022; Simmons & Winograd, 2024)	Requires adoption at creation; stripped by platforms (Kaja, 2024; Longpre *et al.*, 2024)	Emerging (Longpre *et al.*, 2024; Rosenthol, 2022)
Network/Bot Detection (GNN)	88-92% (Peng *et al.*, 2024; Shareef *et al.*, 2024)	Detects coordination, not just content (Huang *et al.*, 2024; Qiao *et al.*, 2024)	Sophisticated bots mimic humans (Dehghan *et al.*, 2023; Ilias *et al.*, 2024)	Medium–High (Lo *et al.*, 2023; Xing *et al.*, 2021)
Multimodal Fusion Systems	90-97% (Roy *et al.*, 2023; Zhao *et al.*, 2021)	Complementary signal integration (Liang *et al.*, 2021; Singh *et al.*, 2019)	Computational cost; latency (Jiao *et al.*, 2024; Liang *et al.*, 2024)	Emerging–Medium (Chen *et al.*, 2024; Jiao *et al.*, 2024)

Note. Passive deepfake forensics accuracy degrades significantly (~15-25 pp) when test content has undergone re-compression or resolution reduction. Accuracy figures represent ranges across benchmark datasets as reported in the included studies; real-world performance is typically lower. pp = percentage points.*

Mitigation and Response Frameworks

Technical Countermeasures

Technical countermeasures encompass both the detection methods reviewed in Section 4 and active countermeasures that disrupt disinformation at the infrastructure level (Bradshaw & DeNardis, 2024). The latter includes account takedowns, content labeling, algorithmic demotion of suspected synthetic content, and the deployment of counter-narrative content to reduce the relative reach of disinformation (Bateman & Jackson, 2024).

Detection alone is insufficient to prevent harm (Simion, 2024). Disinformation can continue to spread even after deepfakes are identified and labeled, because corrections rarely achieve the same reach as the original false content (Kaczmarek *et al.*, 2024). This asymmetry, sometimes called the liar's dividend, means that technical countermeasures must be coupled with media literacy programs and institutional responses that address the trust and behavioral dimensions of disinformation susceptibility (Twomey *et al.*, 2023).

DARPA's Media Forensics program has been a significant U.S. government investment in technical deepfake detection research, funding university and private sector teams to develop methods that remain effective as generative models evolve (Christodorescu *et al.*, 2024). In September 2024, U.S. Senator Ben Cardin was targeted by an AI-driven deepfake operation in which adversaries used synthetic video to impersonate a Ukrainian official in an attempt to extract sensitive political information, which is a high-profile incident that reinforced the urgency of DARPA's initiatives and demonstrated that deepfake threats had migrated from public information environments into direct intelligence operations.

Policy and Legal Countermeasures

The U.S. legal framework for addressing disinformation is fragmented and contested. No single statute comprehensively addresses AI-generated disinformation. Relevant provisions are spread across the Foreign Agents Registration Act (FARA), the Countering America's Adversaries Through Sanctions Act (CAATSA), election law, wire fraud statutes and the Computer Fraud and Abuse Act (Lehman, 2023). The James M. Inhofe National Defense Authorization Act for Fiscal Year 2023 (NDAA FY2023) included provisions directing the DOD

to develop capabilities for identifying and countering foreign malign influence campaigns (Kaczmarek *et al.*, 2024).

The proposed Deepfakes Accountability Act would require disclosure labels on AI-generated media and establish civil and criminal penalties for malicious deepfake creation. As of the period covered by this review, the legislation had not advanced to enactment. Several states have enacted narrower deepfake legislation focused on electoral deepfakes and non-consensual intimate images.

Interagency and Cross-Sector Response

The Foreign Malign Influence Center (FMIC), established at the ODNI in 2022, serves as the coordination hub for intelligence community efforts to counter election disinformation. Its statutory functions include providing policymakers and Congress with comprehensive assessments, indications and warnings of foreign malign influence campaigns from Russia, China, Iran and other adversaries (Kaczmarek *et al.*, 2024). Ahead of the 2024 election, FMIC coordinated with the FBI's new election interference coordination hub, the first time the FBI operated a centralized coordination mechanism directly linking the Bureau with state and local election officials.

This work led to the identification and sanctioning of Tenet Media and its Russian funding relationships, as well as the imposition of sanctions on the Iranian Cognitive Design Production Center. The involvement of Treasury's financial intelligence capabilities represents an expansion of the counter-disinformation toolkit beyond information and technical domains into financial disruption.

U.S. Institutional and Policy Architecture

Agency Roles and Responsibilities

Table 4 maps the roles and primary authorities of the principal U.S. government agencies involved in countering AI-driven disinformation threats.

Table 4: U.S. Agency Roles in Countering Foreign Disinformation Threats

Agency	Primary Role	Key Authority / Program
Dept. of State/GEC	Identifies & publicizes foreign disinformation globally; monitors social media and open sources (Kuznetsov & Liang, 2023; Marangione, 2021)	Global Engagement Center
DHS/CISA	Educates the public on disinformation risks; partners with state and local election officials (Merivaki *et al.*, 2024)	Election Infrastructure ISAC (Muller & Thomas, 2020)
DOD/PA Office	Issues factual counter-statements; monitors and counters disinformation targeting U.S. forces (Kelley, 2024)	DOD Information Operations (Milewski, 2020)
ODNI/FMIC	Coordinates IC-wide assessments of foreign malign influence; statutory warning function (Kelley, 2024)	Foreign Malign Influence Center (est. 2022) (Kelley, 2024)
Treasury/OFAC	Financial sanctions against foreign disinformation networks; tracks foreign funding flows("Biden Administration Imposes Sanctions and Seeks to Cement Alliances to Counter China and Russia," 2021; "Government Agencies and Private Companies Undertake Actions to Limit the Impact of Foreign Influence and Interference in the 2020 U.S. Election," 2021)	CAATSA: Foreign influence sanctions authority (Batyuk, 2020; Ziegler, 2020)
DARPA	Funds foundational R&D in deepfake detection and media forensics (Nguyen *et al.*, 2022; Verdoliva, 2020)	Media Forensics (MediFor) Program (Nguyen *et al.*, 2022; Verdoliva, 2020)

Note. GEC = Global Engagement Center; CISA = Cybersecurity and Infrastructure Security Agency; I&A = Intelligence and Analysis; PA = Public Affairs; FMIC = Foreign Malign Influence Center; OFAC = Office of Foreign Assets Control; DARPA = Defense Advanced Research Projects Agency. Constructed from GAO (2024a) and EUISS (2025).

Persistent Coordination Gaps

Despite the multi-agency architecture described above, the GAO has identified persistent coordination failures. Most critically, the agencies do not use a common definition of disinformation(?, 2020; Fathaigh *et al.*, 2021). DHS/CISA applies its own definition deliberately created to mislead, harm, or manipulate, rather than the IC Lexicon definition used by most intelligence community agencies, creating potential gaps in threat assessment integration (Kaczmarek *et al.*, 2024). The absence of a unified legal authority for domestic counter-disinformation operations stemming from First Amendment constraints and concerns about government influence on speech leaves a gap between the robust foreign-focused authorities of the State Department and the more constrained authorities of domestic-facing agencies.

The collapse of the DHS Disinformation Governance Board in 2022 (Press, 2024; Ruiz, 2023). The most concerted recent attempt to create a coordinating mechanism specifically for disinformation demonstrated the political volatility of this institutional space. GAO's 2022 recommendation to establish a government-wide data analytics center for payment integrity offers a conceptual parallel: the analogous recommendation in the disinformation space, which is a coordinated federal AI-powered threat detection center (Maslej *et al.*, 2024), has not been acted upon.

Research and Policy Gaps

Technical Research Gaps

The systematic review identified six priority technical research gaps. First, real-time detection scalability remains an open problem. Most published detection systems are evaluated in offline, batch settings. Deploying these systems in the real-time information environment, where millions of pieces of content are posted per minute across multiple platforms, requires architectural innovation in both the detection models and the infrastructure supporting them.

Second, adversarial robustness is insufficiently addressed. The majority of reviewed detection papers do not evaluate performance against adversarially crafted disinformation content deliberately designed to evade detection systems (Cresci, 2020; Stiff & Johansson, 2021). As detection tools become publicly known, adversaries will develop corresponding evasion techniques, as has already occurred in the malware detection domain (Afianian *et al.*, 2019; İmamverdiyev & Baghirov, 2024). Benchmark evaluations must incorporate adversarial stress-testing.

Third, cross-lingual and cross-cultural generalization is inadequately studied. The bulk of the reviewed literature addresses English-language content (Helm *et al.*, 2024; Wang *et al.*, 2024). Yet state-sponsored disinformation operations routinely produce content in multiple languages for differentiated audience targeting (Glazunova *et al.*, 2022; Sinelnik & Hovy, 2024). Detection systems that perform well on English-language political news may fail on Arabic, Mandarin, or Farsi-language content produced by the same threat actors.

Fourth, multimodal integration architectures require further development. While the reviewed literature consistently demonstrates that multimodal detection outperforms unimodal detection (Jing *et al.*, 2022; Segura-Bédmar & Alonso-Bartolome, 2022). The computational cost and latency of existing multimodal systems remain barriers to operational deployment (Anggrainingsih *et al.*, 2024). Edge-optimized and distilled multimodal architectures represent an important frontier.

Fifth, ground-truth dataset scarcity limits progress in operational settings. Publicly available labeled disinformation datasets are small relative to the scale of real-world content and are often constructed under conditions that do not reflect the adversarial environment (Lakzaei *et al.*, 2024; Montoro-Montarroso *et al.*, 2023). Classified operational datasets held by intelligence agencies cannot be shared with academic researchers, creating a fundamental asymmetry between the research community and the threat actors (Academic-Practitioner Divide in Intelligence Studies, 2022; Schiffer & Chang, 2023).

Sixth, the explainability and auditability of AI detection decisions are underdeveloped. When an AI system flags a piece of content as disinformation, national security operators and legal authorities require a clear, auditable rationale for that determination. Black-box deep learning systems that cannot explain their outputs cannot be reliably integrated into enforcement or public communication workflows (Schmitt *et al.*, 2024; Szczepański *et al.*, 2021).

RESULTS AND DISCUSSION

The systematic review of 124 sources confirmed that AI-driven disinformation constitutes an operationally active and technically escalating threat to U.S. national security and that AI-driven detection technologies have achieved substantial laboratory performance but face consistent and documented gaps when deployed in real-world operational conditions. On the threat side, Russia, China and Iran have each integrated generative AI tools into active influence operations targeting U.S. democratic institutions, with Russia's operations confirmed by the GAO, the FBI and the Microsoft Threat Analysis Center as the most mature and most directly disruptive. Human detection rates for high-quality deepfakes have fallen to 24.5 percent, AI-enabled fake news websites increased tenfold in a single year following the public release of capable generative AI systems and deepfake volumes were projected to reach 8 million by 2025. On the detection side, transformer-based NLP classifiers achieve benchmark accuracy of 88 to 95 percent, passive deepfake forensic systems achieve 75 to 93 percent on controlled datasets, graph neural network bot detection achieves 88 to 92 percent on historical datasets and multimodal fusion systems that integrate text, audio and visual signals achieve the highest performance tier at 90 to 97 percent, outperforming unimodal approaches by 6 to 14 percentage points across reviewed studies.

The results further confirm that each detection technology class carries specific and documented operational limitations that reduce its real-world effectiveness below benchmark levels. NLP classifiers suffer domain shift, with F1 scores falling by 15 to 30 percentage points when models trained on English-language content are applied to Arabic, Mandarin, or Farsi content produced by the same threat actors. Passive deepfake detection accuracy falls by 15 to 25 percentage points when video content has been re-compressed to the quality levels typical of social media platforms, which is the primary distribution channel for operational deepfakes. Bot detection faces adversarial evasion as sophisticated networks increasingly mimic human behavioral patterns and the 2024 Tenet Media operation demonstrated that the most advanced Russian influence operations route disinformation through genuine human accounts that behavioral classifiers cannot distinguish from uncompromised users. Provenance-based C2PA systems offer near-certain verification for credentialed content but cannot address content produced by adversaries who deliberately bypass authentication. The U.S. institutional architecture produced measurable results in 2024, including the identification and sanctioning of Tenet Media and the Iranian Cognitive Design Production Center, but the GAO identified persistent coordination failures including the absence of a common definitional framework across agencies and no unified authority spanning the full trajectory of a disinformation campaign from foreign origin to domestic amplification.

Discussion

The results of this review reveal three structural tensions that define the current state of the field and that must be resolved if AI-driven detection is to provide adequate national security protection. The first tension is between detection accuracy and operational scalability. Multimodal fusion systems achieve the highest accuracy of any reviewed detection class but require computational resources and latency tolerances that are incompatible with real-time deployment across millions of pieces of content per day. The methods that can operate at social media scale in real time are less accurate and more vulnerable to adversarial evasion. This gap is not merely a technical inconvenience. It means that the best-performing detection systems are unavailable in the operational environment where disinformation does its most significant damage, which is the real-time information environment during high-stakes events such as elections, military crises and public health emergencies. Closing this gap requires architectural innovation in model distillation and edge-optimized multimodal design that the reviewed literature has identified as a priority but not yet resolved.

The second tension is between detection performance and adversarial adaptation. Every published detection method is, in principle, a specification of the artifacts that a sufficiently motivated adversary can learn to conceal. This review confirms that this arms-race dynamic is already observable across all four technology classes. Post-processing of deepfake video destroys the spatial and frequency artifacts that passive forensic classifiers depend upon. Adversarially crafted text is increasingly designed to evade transformer-based classifiers through paraphrasing and semantic substitution. Bot networks operated by sophisticated state actors increasingly replicate human behavioral patterns to defeat temporal and structural detection. The bulk of the reviewed literature evaluates detection performance against non-adversarial content. This is a fundamental methodological limitation because it overstates operational effectiveness. Detection research must incorporate adversarial stress-testing as a standard evaluation requirement, not an optional supplement, if published accuracy figures are to reflect real-world national security utility.

The third tension is between the geographic and jurisdictional scope of the threat and the constrained scope of the U.S. government response. Foreign disinformation operations achieve their most significant domestic impact not by delivering content directly to U.S. audiences from foreign servers, but by seeding content into domestic amplification networks where it travels through genuine human channels that are legally and constitutionally protected from government intervention. The Tenet Media operation demonstrated this model with precision. Russian funds flowed to an American company that paid American influencers who distributed Kremlin narratives to American audiences through platforms that operate under First Amendment protections. The State Department's Global Engagement Center has robust authority to counter foreign disinformation abroad. DHS/CISA operates under significant constraints in the domestic information environment. No single agency possesses end-to-end authority spanning the full trajectory from foreign origin to domestic audience impact. The collapse of the DHS Disinformation Governance Board in 2022 confirmed that the political environment for expanding domestic counter-disinformation authority is highly constrained. Addressing this jurisdictional gap requires legislative action that creates coordination mechanisms without creating the government speech intervention that constitutional and democratic norms rightly prohibit. Achieving that balance is the most difficult institutional design challenge identified in this review and the one for which the existing literature offers the least guidance.

CONCLUSION

AI has turned disinformation into a cheap, scalable weapon. State and non-state actors now use it against American democratic institutions with increasing skill. The 2024 election showed this clearly. AI robocalls suppressed votes in New Hampshire. Deepfake videos targeted senior officials. Foreign-funded influencer networks spread Kremlin narratives through American voices. IRGC operatives hacked campaign systems

and tried to weaponize stolen data. Researchers have built strong detection tools in response. Transformer-based classifiers reach 88 to 95 percent accuracy on benchmark datasets. Multimodal systems that combine text, audio and video beat single-mode systems by 6 to 14 percentage points. C2PA-based provenance systems offer strong certainty for credentialed content. Network analysis tools catch coordinated inauthentic behavior with over 90 percent precision on past data. Still, humans struggle to spot deepfakes. Detection rates have fallen to 24.5 percent. Deepfake volume may reach 8 million by 2025. Fake news sites using AI grew tenfold in one year. This is an uneven fight. Generative AI is cheap and improves fast. Detection tools cost more, face attacks, and struggle outside their training data. Closing this gap needs two things. First, better technical research on speed, robustness, and explainability. Second, institutional reform: shared definitions, agency coordination, and stronger laws. Both must move fast.

REFERENCES

Afianian, A., Niksefat, S., Sadeghiyan, B., & Baptiste, D. A. (2019). Malware Dynamic Analysis Evasion Techniques. *ACM Computing Surveys, 52*(6), 1–28. https://doi.org/10.1145/3365001

Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining, 13*(1), 30–30. https://doi.org/10.1007/s13278-023-01028-5

Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: A review. *Social Network Analysis and Mining, 13*(1), 30. https://doi.org/10.1007/s13278-023-01028-5

Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A. E., Al-Dhaqm, A., Nasser, M., ... & Saif, A. (2022). Financial fraud detection based on machine learning: A systematic literature review. *Applied Sciences, 12*(19), 9637. https://doi.org/10.3390/app12199637

Ali, A., Razak, S. A., Othman, S. H., Eisa, T. A. E., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H., & Saif, A. (2022). Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review. *Applied Sciences, 12(*19), 9637–9637. https://doi.org/10.3390/app12199637

Anggrainingsih, R., Hassan, G. M., & Datta, A. (2024). Transformer-based models for combating rumours on microblogging platforms: a review [Review of *Transformer-based models for combating rumours on microblogging platforms: a review]*. *Artificial Intelligence Review, 57*(8). Springer Science+Business Media. https://doi.org/10.1007/s10462-024-10837-9

Balasubramanian, S. K., Bilgic, M., Culotta, A., Hemphill, L., Nikolich, A., & Shapiro, M. A. (2022). Leaders or Followers? A Temporal Analysis of Tweets from IRA Trolls. *Proceedings of the International AAAI Conference on Web and Social Media, 16,* 2–11. https://doi.org/10.1609/icwsm.v16i1.19267

Bateman, J., & Jackson, D. (2024). *Countering disinformation effectively: An evidence-based policy guide*. Carnegie Endowment for International Peace.

Bateman, J., & Jackson, D. (2024). *Countering disinformation effectively: An evidence-based policy guide*. Carnegie Endowment for International Peace.

Batyuk, V. (2020). Balance of Power between the U.S. and Russia. *World Economy and International Relations, 64*(8), 63–69. https://doi.org/10.20542/0131-2227-2020-64-8-63-69

Bhandari, M., Shrestha, S., Karki, U., Adhikari, S., & Gaihre, R. (2024). Predicting manipulated regions in deepfake videos using convolutional vision transformers. *Computing and Artificial Intelligence., 2*(2), 1409–1409. https://doi.org/10.59400/cai.v2i2.1409

Biden Administration Imposes Sanctions and Seeks to Cement Alliances to Counter China and Russia. (2021). *American Journal of International Law, 115*(3), 536–545. https://doi.org/10.1017/ajil.2021.28

Bing, C., Paul, K., & Bing, C. (2024). US voters targeted by Chinese influence online, researchers say. *Reuters.*

Bing, C., Paul, K., & Bing, C. (2024). US voters targeted by Chinese influence online, researchers say. Reuters. September.

Bipasha, S. (2025). Literature Survey of Image Forgery Detection Using Machine Learning. In *Theseus (Ammattikorkeakoulujen)*. http://www.theseus.fi/handle/10024/883366

Bradshaw, S., & DeNardis, L. (2024). Technical infrastructure as a hidden terrain of disinformation. *Journal of Cyber Policy, 9*(3), 316–332. https://doi.org/10.1080/23738871.2024.2419010

Carpenter, P. (2024). *FAIK: A practical guide to living in a world of deepfakes, disinformation, and AI-generated deceptions*. John Wiley & Sons.

Carpenter, P. (2024). *FAIK: A practical guide to living in a world of deepfakes, disinformation, and AI-generated deceptions.* John Wiley & Sons.

Casola, S., Lauriola, I., & Lavelli, A. (2022). Pre-trained transformers: an empirical comparison. *Machine Learning with Applications, 9*, 100334–100334. https://doi.org/10.1016/j.mlwa.2022.100334

Chavoshi, N., Hamooni, H., & Mueen, A. (2017). *Temporal Patterns in Bot Activities*. 1601–1606. https://doi.org/10.1145/3041021.3051114

Chavoshi, N., Hamooni, H., & Mueen, A. (2017, April). *Temporal patterns in bot activities*. In Proceedings of the 26th international conference on world wide web companion (pp. 1601-1606).

Chen, J., Seng, K. P., Smith, J. S., & Ang, L.-M. (2024). Situation Awareness in AI-Based Technologies and Multimodal Systems: Architectures, Challenges and Applications. *IEEE Access, 12,* 88779–88818. https://doi.org/10.1109/access.2024.3416370

Christodorescu, M., Craven, R., Feizi, S., Gong, N. Z., Hoffmann, M. R., Jha, S., Jiang, Z., Kamarposhti, M. S., Mitchell, J. B., Newman, J., Probasco, E., Qi, Y., Shams, K., & Turek, M. (2024). Securing the Future of GenAI: Policy and Technology. *arXiv*

(Cornell University). https://doi.org/10.48550/arxiv.2407.12999

Cresci, S. (2020). A decade of social bot detection. *Communications of the ACM, 63*(10), 72–83. https://doi.org/10.1145/3409116

Dehghan, A., Siuta, K., Skorupka, A., Dubey, A., Betlen, A., Miller, D., Xu, W., Kamiński, B., & Prałat, P. (2023). Detecting bots in social-networks using node and structural embeddings. *Journal Of Big Data, 10*(1). https://doi.org/10.1186/s40537-023-00796-3

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (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 (pp. 4171–4186). ACL. https://doi.org/10.18653/v1/N19-1423

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). 4171–4186. https://doi.org/10.18653/v1/n19-1423

Ehring, D., Menekse, I., Luttmer, J., & Nagarajah, A. (2024). *Automatic identification of role-specific information in product development: a critical review on large language models [Review of Automatic identification of role-specific information in product development: a critical review on large language models]. Proceedings of the Design Society, 4,* 2009–2018. Cambridge University Press. https://doi.org/10.1017/pds.2024.203

Essa, E., Omar, K., & Alqahtani, A. (2023). Fake news detection based on a hybrid BERT and LightGBM models. *Complex & Intelligent Systems, 9*(6), 6581–6592. https://doi.org/10.1007/s40747-023-01098-0

Essa, E., Omar, K., & Alqahtani, A. (2023). Fake news detection based on a hybrid BERT and LightGBM model. *Complex & Intelligent Systems, 9,* 7021–7035. https://doi.org/10.1007/s40747-023-01098-0

Fathaigh, R. Ó., Helberger, N., & Appelman, N. (2021). The perils of legally defining disinformation. *Internet Policy Review, 10*(4). https://doi.org/10.14763/2021.4.1584

Gambín, Á. F., Yazidi, A., Vasilakos, A. V., Haugerud, H., & Djenouri, Y. (2024). Deepfakes: current and future trends. *Artificial Intelligence Review, 57*(3). https://doi.org/10.1007/s10462-023-10679-x

Gao, J., Xia, Z., Marcialis, G. L., Dang, C., Dai, J., & Feng, X. (2024). DeepFake detection based on high-frequency enhancement network for highly compressed content. *Expert Systems with Applications, 249,* 123732–123732. https://doi.org/10.1016/j.eswa.2024.123732

Garon, J. M. (2022). When AI goes to war: corporate accountability for virtual mass disinformation, algorithmic atrocities, and synthetic propaganda. N. Ky. L. Rev., 49, 181.

Glazunova, S., Bruns, A., Hurcombe, E., Montaña-Niño, S., Coulibaly, S., & Obeid, A. K. (2022). Soft power, sharp power? Exploring RT 's dual role in Russia's diplomatic toolkit. *Information Communication & Society, 26*(16), 3292–3317. https://doi.org/10.1080/1369118x.2022.2155485

Goldani, G. H., Hamooni, H., & Mueen, A. (2021). Fake news detection based on neural networks with word embeddings. *2021 IEEE 12th International Conference on Information and Communication Systems (ICICS),* 1–6.

Government Agencies and Private Companies Undertake Actions to Limit the Impact of Foreign Influence and Interference in the 2020 U.S. Election. (2021). American Journal of International Law, 115(2), 309–317. https://doi.org/10.1017/ajil.2021.10

Guarnera, L., Giudice, O., Niesner, M., & Battiato, S. (2022). On the Exploitation of Deepfake Model Recognition. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* 61–70. https://doi.org/10.1109/cvprw56347.2022.00016

Guo, X., & Yu, H. (2022). On the Domain Adaptation and Generalization of Pretrained Language Models: A Survey. *arXiv (Cornell University).* https://doi.org/10.48550/arxiv.2211.03154

Gupta, G., Raja, K., Gupta, M., Jan, T., Thompson-Whiteside, S., & Prasad, M. (2023). A Comprehensive Review of DeepFake Detection Using Advanced Machine Learning and Fusion Methods. *Electronics, 13*(1), 95–95. https://doi.org/10.3390/electronics13010095

Helm, P., Bella, G., Koch, G., & Giunchiglia, F. (2024). Diversity and language technology: how language modeling bias causes epistemic injustice. *Ethics and Information Technology, 26(*1). https://doi.org/10.1007/s10676-023-09742-6

Hernandez, F. J. (2024). *Deepfake Technology in Corporate Cybersecurity: Emerging Threats and Defense Mechanisms.* https://doi.org/10.36227/techrxiv.172833229.93826709/v1

Huang, H., Tian, H., Zheng, X., Zhang, X., Zeng, D., & Wang, F. (2024). CGNN: A Compatibility-Aware Graph Neural Network for Social Media Bot Detection. *IEEE Transactions on Computational Social Systems, 11(*5), 6528–6543. https://doi.org/10.1109/tcss.2024.3396413

Ilias, L., Kazelidis, I. M., & Askounis, D. (2024). Multimodal Detection of Bots on X (Twitter) using Transformers. *IEEE Transactions on Information Forensics and Security, 19,* 7320–7334. https://doi.org/10.1109/tifs.2024.3435138

İmamverdiyev, Y., & Baghirov, E. (2024). EVASION TECHNIQUES IN MALWARE DETECTION: CHALLENGES AND COUNTERMEASURES. *Problems of Information Technology, 15*(2), 9–15. https://doi.org/10.25045/jpit.v15.i2.02

Jaffer, N. (2021). FAKE NEWS AND DISINFORMATION IN MODERN STATECRAFT. *INSTITUTE OF REGIONAL STUDIES ISLAMABAD, 39*(1), 3–33.

JAFFER, N. (2021). FAKE NEWS AND DISINFORMATION IN MODERN STATECRAFT. *INSTITUTE OF REGIONAL STUDIES ISLAMABAD, 39*(1), 3-33.

Jankowicz, N., & Collis, H. (2020). Enduring Information Vigilance: Government after COVID-19. *The US Army War College Quarterly Parameters, 50*(3). https://doi.org/10.55540/0031-1723.2671

Jiao, T., Guo, C., Feng, X., Chen, Y., & Song, J. (2024). A Comprehensive Survey on Deep Learning Multi-Modal Fusion: Methods, Technologies and Applications. *Computers, Materials & Continua/ Computers, Materials & Continua (Print), 80*(1), 1–35. https://doi.org/10.32604/cmc.2024.053204

Jing, J., Wu, H., Sun, J., Fang, X., & Zhang, H. (2022). Multimodal fake news detection via progressive fusion networks. *Information Processing & Management, 60*(1), 103120–103120. https://doi.org/10.1016/j.ipm.2022.103120

Justice, D. of. (2023). Global Engagement Center Special Report: How The People's Republic Of China Seeks To Reshape The Global Information Environment. *Florida International University Digital Commons (Florida International University)*. https://digitalcommons.fiu.edu/srhreports/resources/resources/194

Kaczmarek, K., Karpiuk, M., & Melchior, C. (2024). Disinformation as a threat to state security. Przegląd *Nauk o Obronności, 9*(20), 45–54.

Kaczmarek, K., Karpiuk, M., & Melchior, C. (2024). Disinformation as a threat to state security. Przegląd *Nauk o Obronności, 9*(20), 45-54.

Kaja, R. (2024). A Comprehensive Content Verification System for ensuring Digital Integrity in the Age of Deep Fakes. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2411.19750

Kelley, M. J. (2024). Understanding Russian Disinformation and How the Joint Force Can Address It. *The US Army War College Quarterly Parameters, 54(*2). https://doi.org/10.55540/0031-1723.3286

Khan, S. A., & Dang-Nguyen, D. (2023). Deepfake Detection: Analyzing Model Generalization Across Architectures, Datasets, and Pre-Training Paradigms. *IEEE Access, 12*, 1880–1908. https://doi.org/10.1109/access.2023.3348450

Kuntur, S., Wróblewska, A., Paprzycki, M., & Ganzha, M. (2024). Under the influence: A survey of large language models in fake news detection. *IEEE Transactions on Artificial Intelligence, 6*, 458–476. https://doi.org/10.1109/TAI.2024.3398989

Kuntur, S., Wróblewska, A., Paprzycki, M., & Ganzha, M. (2024). Under the influence: A survey of large language models in fake news detection. *IEEE Transactions on Artificial Intelligence, 6,* 458–476. https://doi.org/10.1109/TAI.2024.3398989

Kuznetsov, N., & Liang, F. (2023). Digital diplomacy of USA and China in the era of datalization. *Vestnik of Saint Petersburg University International Relations, 16*(2), 191–200. https://doi.org/10.21638/spbu06.2023.206

Lakzaei, B., Chehreghani, M. H., & Bagheri, A. (2024). Disinformation detection using graph neural networks: a survey. *Artificial Intelligence Review, 57*(3). https://doi.org/10.1007/s10462-024-10702-9

Lehman, C. F. (2023). *Modernize the Criminal Justice System: An Agenda for the Ne w Congress*. Manhattan Institute.

Lehman, C. F. (2023). *Modernize the Criminal Justice System: An Agenda for the Ne w Congress*. Manhattan Institute, Apr, 20.

Li, I., Pan, J., Goldwasser, J., Verma, N., Wong, W. P., Nuzumlalı, M. Y., Rosand, B., Li, Y., Zhang, M., Chang, D., Taylor, R. A., Krumholz, H. M., & Radev, D. (2022). Neural Natural Language Processing for unstructured data in electronic health records: A review [Review of *Neural Natural Language Processing for unstructured data in electronic health records: A review]. Computer Science Review, 46*, 100511–100511. Elsevier BV. https://doi.org/10.1016/j.cosrev.2022.100511

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020, June 1). Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr42600.2020.00327

Liang, C. X., Tian, P., Yin, C. H., Yua, Y., An-Hou, W., Ming, L., Wang, T., Bi, Z., & Liu, M. (2024). A Comprehensive Survey and Guide to Multimodal Large Language Models in Vision-Language Tasks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2411.06284

Liang, X., Qian, Y., Guo, Q., Cheng, H., & Liang, J. (2021). AF: An Association-Based Fusion Method for Multi-Modal Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(12), 9236–9254. https://doi.org/10.1109/tpami.2021.3125995

Ligon, G. S., Harms, M., Grace, E., & Ursch, B. (2024). *One Year Later: The Impact of the Oct. 7 Attack in the United States.*

Lo, W. W., Kulatilleke, G. K., Sarhan, M., Layeghy, S., & Portmann, M. (2023). XG-BoT: An explainable deep graph neural network for botnet detection and forensics. *Internet of Things, 22,* 100747–100747. https://doi.org/10.1016/j.iot.2023.100747

Longpre, S., Mahari, R., Obeng-Marnu, N., Brannon, W., South, T., Kabbara, J., & Pentland, S. (2024). *Data Authenticity, Consent, and Provenance for AI Are All Broken: What Will It Take to Fix Them?* https://doi.org/10.21428/e4baedd9.a650f77d

M., G., Jon. (2022). When AI Goes to War: Corporate Accountability for Virtual Mass Disinformation, Algorithmic Atrocities, and Synthetic Propaganda. *NSUWorks (Nova Southeastern University)*. https://nsuworks.nova.edu/law_facarticles/452

Marangione, M. S. (2021). Words as Weapons: The 21st Century Information *War. Security and Intelligence, 6*(1). https://doi.org/10.18278/gsis.6.1.7

Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., & Clark, J. A. (2024). Artificial Intelligence Index Report 2024. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2405.19522

Mcgarry, B. W. (2022). *FY2022 National Defense Authorization Act: Context and Selected Issues for Congress.*

Merivaki, T., Suttmann-Lea, M., McCreary, M.-C., & Daniel, T. (2024). The #TrustedInfo2022 dataset: States' trust-building social media campaigns during the 2022 election cycle. *State Politics & Policy Quarterly, 24(*4), 468–478. https://doi.org/10.1017/spq.2024.14

Milewski, D. (2020). The Analysis of Narratives and Disinformation in the Global Information Environment Amid Covid-19 Pandemic. *EUROPEAN RESEARCH STUDIES JOURNAL,* 3–17. https://doi.org/10.35808/ersj/1848

Mohammed Hafiz Nabila, & Nunana Klenam Djokoto. (2025). Narratives that build nations: Storytelling as a strategic tool against disinformation in the United States. *ARCouncil Journal of Education and Sociology, 4(*10). https://doi.org/10.5281/zenodo.17313368

Mohammed Hafiz Nabila, & Matilda Thompson. (2025). Countering foreign influence: The role of strategic communications in national security policy making in the United States. *Sarcouncil Journal of Arts, Humanities and Social Sciences, 4*(9). https://doi.org/10.5281/zenodo.17135544

Mohammed Hafiz Nabila, Rohany Abdul Shaibu, Glory Edinam Afeti, & Esinu Aku Adza. (2025). Disinformation as a driver of political polarization: A strategic framework for rebuilding civic trust in the U.S. *World Journal of Advanced Research and Reviews, 27*(1), 916–925. https://doi.org/10.30574/wjarr.2025.27.1.2564

Mohammed Hafiz Nabila, & Matilda Thompson. (2025). The impact of disinformation on national security policymaking: A review. *EPRA International Journal of Multidisciplinary Research (IJMR), 11*(6). https://doi.org/10.36713/epra22781

Montoro-Montarroso, A., Cantón-Correa, J., Rosso, P., Chulvi, B., Panizo-LLedot, Á., Huertas-Tato, J., Figueras, B. C., Rementería, M. J., & Gómez-Romero, J. (2023). Fighting disinformation with artificial intelligence: fundamentals, advances and challenges. *El Profesional de La Informacion*. https://doi.org/10.3145/epi.2023.may.22

Muller, S. R., & Thomas, C. E. (2020). Election Infrastructure Security: Grants and Reimbursement to the States for Usage of their National Guards in State Active Duty Status to Provide Cybersecurity for Federal Elections. *Annals of Computer Science and Information Systems, 24,* 73–78. https://doi.org/10.15439/2020km7

Musulan, A., Xia, V., Kosak-Hine, E., Gibbs, T., Sujaya, V., Rabbany, R., Godbout, J., & Pelrine, K. (2024). Online Influence Campaigns: Strategies and Vulnerabilities. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2501.10387

Musulan, A., Xia, V., Kosak-Hine, E., Gibbs, T., Sujaya, V., Rabbany, R., ... & Pelrine, K. (2024). Online Influence Campaigns: Strategies and Vulnerabilities. arXiv preprint arXiv:2501.10387.

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyên, T. T., Pham, Q., & Nguyen, C. (2022). Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding, 223*, 103525–103525. https://doi.org/10.1016/j.cviu.2022.103525

Padalko, H. (2025). *AI and Information Manipulation: Russia's Interference in the US Elections.* Centre for International Governance Innovation.

Peng, H., Zhang, J., Huang, X., Hao, Z., Li, A., Yu, Z., & Yu, P. S. (2024). Unsupervised Social Bot Detection via Structural Information Theory. *arXiv (Cornell University*). https://doi.org/10.1145/3660522

Press, U. (2024). Parameters Summer 2024. T*he US Army War College Quarterly Parameters, 54*(2). https://doi.org/10.55540/0031-1723.3281

Qiao, B., Zhou, W., Li, K., Li, S., & Hu, S. (2024). Dispelling the Fake: Social Bot Detection Based on Edge Confidence Evaluation. I*EEE Transactions on Neural Networks and Learning Systems, 36*(4), 7302–7315. https://doi.org/10.1109/tnnls.2024.3396192

Rosenthol, L. (2022). C2PA: the world's first industry standard for content provenance (Conference Presentation). 26–26. https://doi.org/10.1117/12.2632021

Roy, S. K., Deria, A., Hong, D., Rasti, B., Plaza, A., & Chanussot, J. (2023). Multimodal Fusion Transformer for Remote Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing, 61,* 1–20. https://doi.org/10.1109/tgrs.2023.3286826

Ruiz, L. G. (2023). Disinformation, Misinformation and Limits on Freedom of Expression During the Covid-19 Pandemic: A Critical Inquiry. The *Age of Human Rights Journal, 21*. https://doi.org/10.17561/tahrj.v21.8149

Schiffer, P., & Chang, F. (2023). How Openness Could Strengthen Academia's Partnerships with the Intelligence Community. *Issues in Science and Technology, 39*(4), 25–27. https://doi.org/10.58875/fqun1916

Schmitt, V., Villa-Arenas, L.-F., Feldhus, N., Meyer, J., Spang, R. P., & Möller, S. (2024). *The Role of Explainability in Collaborative Human-AI Disinformation Detection*. 2157–2174. https://doi.org/10.1145/3630106.3659031

Segura-Bédmar, I., & Alonso-Bartolome, S. (2022). Multimodal Fake News Detection. *Information, 13*(6), 284–284. https://doi.org/10.3390/info13060284

Sengupta, S., Ghosh, S., Nakov, P., & Mitra, P. (2023). Can You Answer This? – Exploring Zero-Shot QA Generalization Capabilities in Large Language Models (Student Abstract). *Proceedings of the AAAI Conference on Artificial Intelligence, 37*(13), 16318–16319. https://doi.org/10.1609/aaai.v37i13.27019

Shareef, Sk. K., Chaitanya, R. K., Chennupalli, S., Chokkakula, D., Kiran, K. V. D., Pamula, U., & Vatambeti, R. (2024). Enhanced botnet detection in IoT networks using zebra optimization and dual-channel GAN classification. *Scientific Reports, 14*(1).

https://doi.org/10.1038/s41598-024-67865-2

Shidaganti, G., Shetty, R., Edara, T., Srinivas, P., & Tammineni, S. C. (2024). Exploratory analysis on the natural language processing models for task specific purposes. *Bulletin of Electrical Engineering and Informatics, 13(*2), 1245–1255. https://doi.org/10.11591/eei.v13i2.6360

Simmons, J. C., & Winograd, J. M. (2024). Interoperable Provenance Authentication of Broadcast Media using Open Standards-based Metadata, Watermarking and Cryptography. *arXiv (Cornell University).* https://doi.org/10.48550/arxiv.2405.12336

Sinelnik, A., & Hovy, D. (2024). *Narratives at Conflict: Computational Analysis of News Framing in Multilingual Disinformation Campaigns.* 225–237. https://doi.org/10.18653/v1/2024.acl-srw.21

Singh, A. P. (2024). Safeguarding Authenticity in the Digital Realm: A Holistic Approach Integrating Content Provenance, Secure Watermarking, and Transparent Labeling to Combat Deepfakes. *International Journal for Multidisciplinary Research, 6(*3), 1-14.

Singh, M., Singh, R., & Ross, A. (2019). A comprehensive overview of biometric fusion. *Information Fusion, 52*, 187–205. https://doi.org/10.1016/j.inffus.2018.12.003

Soudy, A. H., Sayed, O., Tag-Elser, H., Ragab, R., Mohsen, S., Mostafa, T., Abohany, A. A., & Slim, S. O. (2024). Deepfake detection using convolutional vision transformers and convolutional neural networks. *Neural Computing and Applications, 36*(31), 19759–19775. https://doi.org/10.1007/s00521-024-10181-7

Soudy, A. H., Sayed, O., Tag-Elser, H., Ragab, R., Mohsen, S., Mostafa, T., ... & Slim, S. O. (2024). Deepfake detection using convolutional vision transformers and convolutional neural networks. *Neural Computing and Applications, 36*(31), 19759-19775.

Stiff, H., & Johansson, F. (2021). Detecting computer-generated disinformation. I*nternational Journal of Data Science and Analytics, 13(*4), 363–383. https://doi.org/10.1007/s41060-021-00299-5

Swanström, N., & Logan, T. J. (2025). *G7 Strategy for Countering Russian Information Operations in the Indo-Pacific Region: A Framework for Enhanced Multilateral Coordination and Response.*

Szczepański, M., Pawlicki, M., Kozik, R., & Choraś, M. (2021). New explainability method for BERT-based model in fake news detection. *Scientific Reports, 11*(1). https://doi.org/10.1038/s41598-021-03100-6

Twomey, J. G., Ching, D., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2023). Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. *PLoS ONE, 18*(10). https://doi.org/10.1371/journal.pone.0291668

Verdoliva, L. (2020). Media Forensics and DeepFakes: An Overview. *IEEE Journal of Selected Topics in Signal Processing, 14*(5), 910–932. https://doi.org/10.1109/jstsp.2020.3002101

Wang, X., Zhang, W., & Rajtmajer, S. (2024). Monolingual and Multilingual Misinformation Detection for Low-Resource Languages: A Comprehensive Survey. *arXiv (Cornell University).* https://doi.org/10.48550/arxiv.2410.18390

Wang, Z., Cheng, Z., Xiong, J., Xu, X., Li, T., Veeravalli, B., & Yang, X. (2024). A Timely Survey on Vision Transformer for Deepfake Detection. *arXiv (Cornell University).* https://doi.org/10.48550/arxiv.2405.08463

Williams, A. R., Burke-Moore, L., Chan, R. S.-Y., Enock, F. E., Nanni, F., Sippy, T., Chung, Y.-L., Gabašová, E., Hackenburg, K., & Bright, J. (2024). Large language models can consistently generate high-quality content for election disinformation operations. *arXiv (Cornell University).* https://doi.org/10.48550/arxiv.2408.06731

Wodajo, D., & Solomon, A. (2021). Deepfake Video Detection Using Convolutional Vision Transformer. *arXiv (Cornell University*). https://doi.org/10.48550/arxiv.2102.11126

Xing, Y., Shu, H., Zhao, H., Li, D., & Guo, L. (2021). Survey on Botnet Detection Techniques: Classification, Methods, and Evaluation. *Mathematical Problems in Engineering,* 2021, 1–24. https://doi.org/10.1155/2021/6640499

Zhang, Z., Wu, Y., Zhao, H., Li, Z., Zhang, S., Zhou, X., & Zhou, X. (2020). Semantics-Aware BERT for Language Understanding. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(5), 9628–9635. https://doi.org/10.1609/aaai.v34i05.6510

Zhao, T., Tian, S., Daly, J. R., Geiger, M., Jia, M., & Zhang, J. (2024). Information retrieval and classification of real-time multi-source hurricane evacuation notices. *International Journal of Disaster Risk Reduction, 111,* 104759–104759. https://doi.org/10.1016/j.ijdrr.2024.104759

Zhao, Y., Cao, X., Lin, J., Yu, D., & Cao, X. (2021). Multimodal Affective States Recognition Based on Multiscale CNNs and Biologically Inspired Decision Fusion Model. *IEEE Transactions on Affective Computing, 14*(2), 1391–1403. https://doi.org/10.1109/taffc.2021.3093923

Ziegler, C. E. (2020). Sanctions in U.S. - Russia Relations. *Vestnik RUDN International Relations, 20*(3), 504–520. https://doi.org/10.22363/2313-0660-2020-20-3-504-520