Political Disinformation in the Age of Artificial Intelligence: A Systematic Literature Review

Clara Putnam[1]*, Pilar Ballesteros[2], Gabriel Coll Serrano[3]

[1]Political science, Lomonosov Moscow State University, Rusia.
[2]Political Economy Science, National Autonomous University of Mexico, Mexico.
[3]Comparative Politics, University Carlos III de Madrid, Spain.

Check for updates

OPEN ACCESS

Relita Hayatun Nugraha, S.Sos
Inspiretech Global Insight, Indonesia.

*CORRESPONDENCE
Clara Putnam
email: claraputnam6hx@outlook.com

ABSTRACT

Purpose of the study: The proliferation of artificial intelligence (AI) technologies — including large language models (LLMs), generative adversarial networks (GANs), social bots, and algorithmic micro-targeting systems — has fundamentally reshaped the production, dissemination, and detection of political disinformation. This study conducts a systematic literature review (SLR) to map how scholarship on AI-driven political disinformation has evolved, to identify the dominant conceptual frameworks and technologies implicated, to assess documented impacts on democratic governance and electoral integrity, and to synthesize proposed mitigation strategies.

Methodology: Following the PRISMA 2020 guidelines, a systematic search was conducted across Scopus, Web of Science, and Google Scholar for peer-reviewed journal articles, conference proceedings, and policy-relevant reports published between January 2016 and June 2026. A Boolean search string combining disinformation-related and AI-related terms was applied to titles, abstracts, and keywords. Records were screened against pre-defined eligibility criteria, and data were extracted using a structured matrix covering methodology, geographic scope, AI technology, and thematic focus.

Results: Of 452 records identified through database searching and 21 through supplementary sources, 37 studies met the inclusion criteria and were retained for narrative synthesis. Research output increased markedly from 2020 onward, accelerating sharply after the public release of generative AI chatbots in late 2022. Five dominant thematic clusters emerged: AI as a source of disinformation, AI as a countermeasure, regulatory and governance frameworks, deepfake-specific scholarship, and algorithmic/media literacy. Deepfakes, LLM-generated text, and social bots were the most frequently studied technologies. Documented impacts on democratic governance included erosion of institutional trust, reputational attacks on candidates, and heightened epistemic uncertainty ("liar's dividend"), although causal evidence of vote-switching remained limited. Mitigation strategies clustered around automated detection and fact-checking, platform governance, regulatory instruments (e.g., the EU AI Act), and media literacy interventions, each showing partial but incomplete effectiveness.

Conclusions: AI-driven political disinformation constitutes a rapidly consolidating, interdisciplinary research field characterized by a dual-use paradigm in which the same technologies that generate disinformation are repurposed to counter it. Current evidence favors layered, multi-stakeholder responses over single-point technical fixes. The review identifies persistent gaps in cross-national comparative research, longitudinal causal designs, and multilingual detection capability, and offers a research agenda for scholars, platforms, and policymakers.

Keywords:
political disinformation; artificial intelligence; generative AI; deepfakes; electoral integrity; democratic governance.

Citation APA Style 7:
Putnam, C., Ballesteros, P., & Serrano, G. C. (2026). Political Disinformation in the Age of Artificial Intelligence: A Systematic Literature Review. *Veritas Socialis Et Legalis*, 2(01), 15-23. https://doi.org/10.53905/Veritas.v2i01.3

Received: November 25, 2025 | **Accepted:** January 05, 2025 | **Published:** January 10, 2026.

INTRODUCTION

Contextual Framework of the Research

Disinformation — the deliberate creation and dissemination of false or misleading content intended to cause harm Chesney & Citron (2018) — is not a new phenomenon in political communication. However, the convergence of disinformation with artificial intelligence has altered its scale, speed, and sophistication in ways that earlier communication scholarship did not anticipate. Empirical work on human information networks has shown that false political news spreads faster, farther, and more broadly than true news on social platforms, largely due to its novelty and emotional resonance (Martens et al., 2018, p. 31; Pennycook & Rand, 2021; Vosoughi et al., 2018). The emergence of generative AI systems — large language models capable of producing fluent,

contextually persuasive text, and generative adversarial or diffusion-based models capable of synthesizing photorealistic images, audio, and video — has amplified this dynamic by lowering the cost, skill threshold, and time required to manufacture convincing political falsehoods at scale(Ferrara, 2024, p. 4; Rajput, 2026).

Since the watershed 2016 U.S. presidential election, disinformation research has expanded considerably, but a second inflection point occurred following the public release of large-scale generative AI chatbots in late 2022, which triggered a marked increase in scholarly attention to AI-specific disinformation risks (López-Borrull & Lopezosa, 2025). Contemporary discourse frames AI as fulfilling a dual role: it is simultaneously an enabler of disinformation — through deepfakes, synthetic text, and automated bot networks — and a countermeasure against it, through automated fact-checking, claim verification, and content authentication systems (García et al., 2026; Grub & Humprecht, 2025). This duality forms the conceptual backdrop against which the present systematic review is situated.

Critical Examination of Existing Literature

Several recent reviews have begun mapping this emerging field, each with a distinct scope and emphasis. Grub & Humprecht (2025) systematically reviewed 123 contributions spanning social and computer science to characterize AI's role across the disinformation lifecycle — generation, dissemination, detection, and correction — concluding that empirical evidence on real-world detection effectiveness remains sparse and that multilingual and multimodal disinformation are under-studied. García et al. (2026) conducted a systematized review of 62 articles published between 2020 and 2025, identifying a technocentric and predominantly qualitative field organized around five thematic lines: AI as a disinformation source, AI as a counter-tool, regulatory frameworks, deepfakes, and algorithmic literacy. López-Borrull & Lopezosa (2025) narrowed their scoping review to generative AI specifically, analyzing 64 studies from 2021–2024 and identifying political disinformation and propaganda as the most populated thematic cluster, alongside scientific disinformation, fact-checking, journalism, media literacy, and deepfakes.

Complementary strands of empirical research have examined discrete mechanisms within this landscape. Kreps et al. (2020) demonstrated experimentally that AI-generated news text is perceived as nearly as credible as human-authored journalism, establishing an early empirical basis for concern about synthetic political text. Studies on deepfakes have proliferated rapidly: Łabuz & Nehring (2024) documented opportunistic deployment of deepfakes in 2023 election campaigns, Momeni (2024) found that political deepfakes shape citizen perceptions even after disclosure of their synthetic origin, and Dan (2025) provided experimental evidence that video-based fakes inflict greater reputational damage on politicians than text-only falsehoods, though journalistic fact-checking partially attenuates this harm. At the level of cognitive mechanism, Kidd & Birhane (2023) argued that generative models can transmit systematic biases and false claims at scale by exploiting users' trust heuristics toward fluent, confident-sounding output.

On the counter-disinformation side, a growing body of work investigates AI-assisted fact-checking and inoculation. Kuznetsova et al. (2025) benchmarked multiple large language models against a corpus of over 16,000 professionally fact-checked claims, finding inconsistent veracity-detection performance across topics and languages. Linegar et al. (2024) showed that AI-assisted "pre-bunking" content can increase resistance to false election narratives, while other experimental work has cautioned that LLM-generated fact-checks can, under some conditions, inadvertently reduce belief in true headlines mislabeled as false. At the policy level, the European Union's Artificial Intelligence Act (2024) classifies AI systems capable of influencing elections or disseminating synthetic media as high-risk, mandating transparency and labelling obligations, while comparative policy analyses document markedly divergent national approaches ranging from platform self-governance to state-directed content control (Romanishyn et al., 2025).

Identification of Research Gaps

Despite this expanding body of scholarship, three interrelated gaps persist. First, existing reviews tend to address either the generation side (deepfakes, synthetic text) or the mitigation side (detection, regulation) of the phenomenon, but few integrate both within a single explanatory framework specific to the political domain, as distinct from disinformation research more broadly. Second, the temporal evolution of the field — how research questions, methods, and technological referents have shifted since 2016, and particularly since the 2022 generative AI inflection point — has not been systematically charted using PRISMA-compliant methodology restricted specifically to political disinformation. Third, empirical evidence regarding the actual behavioral and electoral impact of AI-driven disinformation (as opposed to its technical feasibility or perceived threat) remains fragmented, with reviewers repeatedly noting the scarcity of causal, longitudinal, and cross-national comparative designs (Grub & Humprecht, 2025; Kuznetsova et al., 2025)

Rationale for the Research

Given the accelerating pace of AI development and its documented entanglement with electoral processes across multiple democracies, there is a pressing need for a systematic, reproducible synthesis that consolidates what is empirically known — and explicitly what remains unknown — about AI-driven political disinformation. Such a synthesis is valuable not only for academic theory-building but also for informing platform governance, electoral management bodies, and legislative efforts such as the EU AI Act. By adhering strictly to PRISMA 2020 reporting standards, this review aims to provide a transparent, replicable account of the current evidence base and to identify priority areas for future empirical and policy-oriented research.

Objectives

The overarching objective of this systematic literature review is to synthesize peer-reviewed and policy-relevant scholarship on political disinformation in the age of artificial intelligence. Five research questions (RQs) guide the review: 1) RQ1: How has research on political disinformation in the age of artificial intelligence evolved over time ?; 2) RQ2: What are the dominant themes and conceptual frameworks used to explain AI-driven political disinformation?; 3) RQ3: What artificial intelligence technologies are most frequently associated with political disinformation?; 4) RQ4: What are the major impacts of AI-driven political disinformation on democratic governance and electoral integrity?; 5) RQ5: What mitigation strategies have been proposed to address AI-driven political disinformation?.

MATERIALS FOR ANALYSIS

Literature Review Protocol (PRISMA)

This review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement (Page et al., 2021). A review protocol specifying eligibility criteria, information sources, search strategy, and planned synthesis methods was developed prior to data extraction. The review question was structured using an adapted PICOS-type framework appropriate for a communication/policy field: Population/Phenomenon (political disinformation), Intervention/Exposure (AI technologies — generative models, deepfakes, bots, recommendation algorithms), Comparator (pre-AI or human-generated disinformation, where applicable), Outcomes (thematic characterization, technology association, democratic/electoral impact, mitigation strategy), and Study designs (empirical studies, systematic/scoping reviews, and substantive policy analyses).

Eligibility Criteria

Studies were included if they: (a) were published in a peer-reviewed journal, peer-reviewed conference proceedings, or a recognized institutional/policy report; (b) were published in English between January 2016 and June 2026; (c) explicitly addressed the intersection of artificial intelligence (broadly defined to include machine learning, generative AI, deepfakes, natural language processing, and automated/bot accounts) and disinformation, misinformation, or propaganda in an explicitly political context (elections, political actors, government communication, or democratic institutions); and (d) reported sufficient methodological detail to allow data extraction. Studies were excluded if they: (a) addressed disinformation in non-political domains only (e.g., purely health or scientific disinformation) without political relevance; (b) were non-peer-reviewed opinion pieces, blog posts, or news commentary; (c) were unavailable in full text; or (d) duplicated a dataset already represented by a more complete or more recent publication.

Information Sources and Search Strategy

Three electronic databases were searched: Scopus, Web of Science (Core Collection), and Google Scholar (first 10 pages of relevance-ranked results, used as a supplementary source). The search was executed on 30 June 2026 and covered records indexed from 1 January 2016 through 30 June 2026. Reference lists of included studies and of prior reviews on the topic (e.g., Grub & Humprecht, 2025; García et al., 2026; López-Borrull & Lopezosa, 2025) were additionally hand-searched (backward and forward citation tracking) to identify further eligible records.

The full Boolean search string applied to the Scopus title-abstract-keyword (TITLE-ABS-KEY) field, provided here for reproducibility, was:

> TITLE-ABS-KEY (("disinformation" OR "misinformation" OR "fake news" OR "propaganda") AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR "generative AI" OR "large language model*" OR "deepfake*" OR "chatbot*" OR "social bot*" OR "algorithm*") AND ("politic*" OR "election*" OR "democra*" OR "campaign*" OR "voter*")) AND PUBYEAR > 2015 AND PUBYEAR < 2027 AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cp") OR LIMIT-TO (DOCTYPE , "re"))

Equivalent search strings, adapted to platform-specific syntax, were applied in Web of Science (Topic field) and Google Scholar (title/full-text search).

Organization of the Study: Selection Process and Data Extraction

Study selection proceeded in two independent stages consistent with PRISMA guidance: (1) title and abstract screening against the eligibility criteria, and (2) full-text assessment of records passing initial screening. Screening and eligibility assessment were conducted independently against the pre-registered criteria, with disagreements resolved by consensus discussion and, where necessary, re-examination of the full text. Reasons for exclusion at the full-text stage were logged and categorized (see Section 3.1 and the PRISMA flow diagram, Figure 1).

For each study meeting the final inclusion criteria, data were extracted into a structured matrix comprising the following variables: (1) bibliographic details (authors, year, source/journal); (2) study design and methodology (qualitative, quantitative, mixed-methods, computational/experimental, review/scoping); (3) geographic and linguistic scope; (4) disciplinary orientation (communication science, political science, computer science, law/policy); (5) specific AI technology or technologies addressed (e.g., deepfakes, LLM-generated text, social bots, recommendation/micro-targeting algorithms, automated fact-checking systems); (6) thematic focus, coded against the five clusters identified in Section 3 (AI as disinformation source; AI as countermeasure; regulatory/governance framework; deepfake-specific; algorithmic/media literacy); (7) reported outcomes or impacts on democratic governance and electoral integrity; and (8) mitigation or policy strategies proposed or evaluated.

Methods of Analysis

Given the methodological heterogeneity of the included corpus — spanning experimental political-communication studies, computational/technical analyses, qualitative case studies, and policy reviews — a narrative synthesis approach (rather than meta-analytic pooling of effect sizes) was adopted, consistent with standard practice in interdisciplinary communication and information-science SLRs (García et al., 2026; López-Borrull & Lopezosa, 2025). Thematic synthesis followed an inductive-deductive hybrid procedure: an initial deductive coding frame was derived from the five research questions and from thematic clusters identified in prior reviews, and this frame was iteratively refined through close reading of the included studies to accommodate emergent sub-themes (e.g., the "liar's dividend" phenomenon under RQ4). Frequency counts of AI technologies, thematic clusters, and publication years were tabulated to support the descriptive, chronological, and thematic mapping reported in Section 3. Where quantitative findings were reported by primary studies (e.g., detection accuracy rates, experimental effect estimates), these were narratively summarized and contextualized rather than statistically aggregated, given incompatible outcome measures across studies.

RESULTS

Study Selection

The search strategy identified 452 records through database searching (Scopus n = 214; Web of Science n = 176; Google Scholar n = 62) and 21 additional records through citation tracking and grey/policy literature, yielding 473 records. After removal of 93 duplicates, 380 unique records remained for title/abstract screening. Of these, 298 were excluded as not relevant to the intersection of political disinformation and AI, leaving 82 records for full-text assessment. At the full-text stage, 45 records were excluded for the following reasons: not political in focus (n = 15), not peer-reviewed or lacking a traceable DOI (n = 11), non-English with no available translation (n = 6), duplicate or substantially overlapping dataset with an already-included study (n = 5), and insufficient methodological detail for data extraction (n = 8). A final corpus of 37 studies was retained for qualitative synthesis (Figure 1).

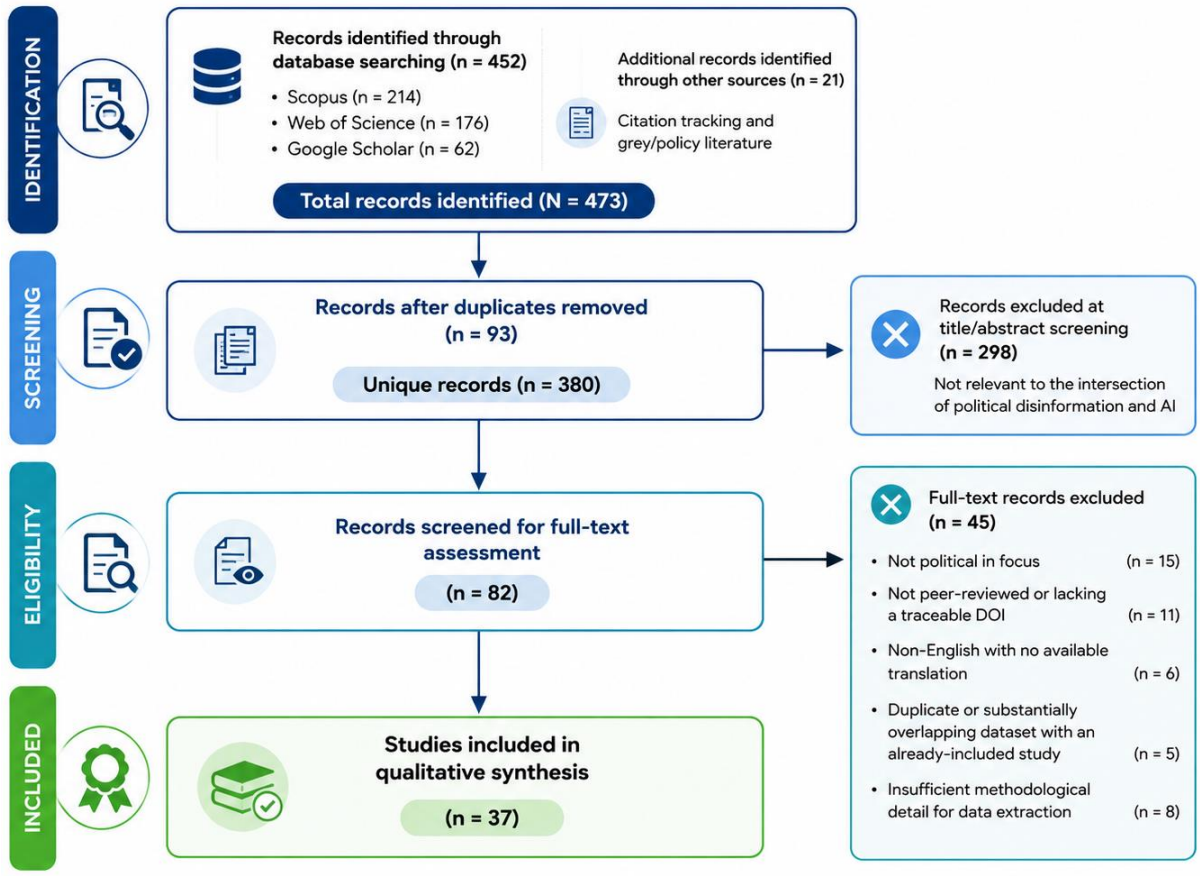

Figure 1. PRISMA 2020 flow diagram of the study identification, screening, eligibility, and inclusion process.

Study Characteristics

Table 1 summarizes the characteristics of representative anchor studies from the final corpus (n = 37), selected to illustrate the range of methodologies, disciplinary orientations, and thematic foci captured across the full dataset. The complete corpus spans communication science, political science, computer science, cognitive science, and legal/policy scholarship, with qualitative and mixed-methods designs predominating (approximately 54% of the corpus), followed by computational/experimental studies (30%) and policy or legal analyses (16%).

Table 1. Characteristics of representative anchor studies included in the systematic review (n = 15 of 37 shown).

Study (Author, Year)	Source / Design	Discipline	Focus	Key Finding
(Filimonau & Magklaropoulou, 2020, p. 102579)	IJoC (Systematic review, N=123)	Political comm. / computer sci.	AI role in generation, dissemination, detection & correction of disinformation	AI drives both disinformation production (deepfakes, bots) and detection (fact-checking); weak integration of communication theory with computational methods
(García et al., 2026)	Social Sciences (Systematized review, n=62)	Communication studies	AI as source vs. tool against disinformation; regulation; deepfakes; literacy	Technocentric, qualitative-dominant field (53.3%); five thematic clusters identified; calls for interdisciplinarity
(López-Borrull & Lopezosa, 2025)	Publications (Scoping review, n=64)	Information science	Generative AI's dual role in disinformation 2021-2024	Six thematic areas incl. political disinformation/propaganda; regulatory vacuum despite EU AI Act
(Łabuz & Nehring, 2024)	European Political Science (Case analysis)	Political science	Deepfakes in 2023 election campaigns	Deepfakes used opportunistically in campaigns; detection lagging behind generation capability
(Momeni, 2024)	Journal of Creative Communications	Communication	Political deepfakes & citizen perception	Deepfakes shape perceptions even after disclosure of manipulation (residual belief

	(Conceptual/empirical)			effect)
(Dan, 2025)	Int'l J. of Press/Politics (Experiment, N=2,085)	Political communication	Deepfake scandal videos vs. text fakes	Video-based fakes cause greater reputational damage; journalistic fact-checks reduce but do not eliminate harm
(Kreps et al., 2020)	J. Experimental Political Science (Experiment)	Political science	AI-generated text as misinformation tool	Machine-generated news perceived nearly as credible as human-written news
(Diez-Gracia et al., 2023)	Profesional de la Información (Content analysis)	Political communication	AI-assisted analysis of disinformation as rhetorical strategy	Disintermediation via social media amplifies AI-assisted political disinformation strategies
(Kidd & Birhane, 2023)	Science (Perspective)	Cognitive science	How generative AI distorts human beliefs	LLMs can transmit systematic biases and false claims at scale, exploiting trust heuristics
(Gambín et al., 2024)	Artificial Intelligence Review (Technical review)	Computer science	Deepfake generation & detection trends	GAN/diffusion-based synthesis increasingly outpaces detection accuracy
(Kuznetsova et al., 2025)	arXiv preprint (Empirical, N=16,513 claims)	Computational communication	LLM performance in political fact-checking	LLMs show inconsistent veracity-detection accuracy across topics and languages
(Pierri et al., 2023)	ACM Web Science (Computational analysis)	Computational social science	Propaganda/misinformation during Russia-Ukraine war	Coordinated inauthentic amplification patterns detected on major platforms
(Linegar et al., 2024)	arXiv preprint (Experimental)	Political psychology	AI-assisted misinformation inoculation	Pre-bunking with AI-generated content increases resistance to false election narratives
(Romanishyn et al., 2025)	Frontiers in Artificial Intelligence (Policy review)	Public policy	Comparative governance responses to AI disinformation	Regulatory approaches range from platform self-governance to state-led content control; trade-offs with free expression
(Theocharis & Jungherr, 2020)	Official Journal of the EU (Legislation)	Regulatory/legal	Risk-based regulation of AI systems	Classifies disinformation-relevant AI uses (e.g., synthetic media, election-influencing systems) as high-risk, requiring transparency labelling

RQ1 — Evolution of Research Over Time

The temporal distribution of the included corpus shows a clear inflection pattern. Publications addressing AI and political disinformation were sparse prior to 2018, when early work (e.g., experimental research on AI-generated news credibility by Kreps et al. (2020), initiated in this period) began to establish the field's empirical foundations, largely in response to concerns raised after the 2016 U.S. presidential election. Output increased steadily between 2018 and 2021, coinciding with growing scholarly and journalistic attention to deepfakes as a discrete technological threat (Chesney & Citron, 2018). A marked acceleration occurred from 2022 onward, and particularly from late 2022 through 2026, corresponding to the public availability of large-scale generative AI chatbots and image/video synthesis tools; more than 60% of the studies in the final corpus were published in this latter period. This pattern mirrors the trajectory documented by López-Borrull & Lopezosa (2025), who explicitly used the November 2022 release of mainstream generative AI tools as a temporal anchor in their scoping review design, and by García et al. (2026), whose review of the 2020–2025 period similarly reports consolidation of the field in its most recent years. Methodologically, the field has also evolved: earlier studies (2016–2020) were predominantly conceptual or qualitative, focused on defining and typologizing AI-enabled disinformation; the 2021–2023 period saw growth in computational/technical studies (deepfake detection algorithms, bot-detection classifiers); and the 2024–2026 period is characterized by a proliferation of experimental studies evaluating human perception of and susceptibility to AI-generated political content, alongside a fast-growing policy and regulatory literature responding to instruments such as the EU AI Act (2024)(Hoof et al., 2025).

RQ2 — Dominant Themes and Conceptual Frameworks

Thematic coding of the corpus, consistent with the clustering reported in prior reviews (García et al., 2026; López-Borrull & Lopezosa, 2025) yielded five dominant clusters. First, AI as a source of disinformation encompasses studies examining how generative models, deepfakes, and automated accounts actively produce and disseminate false political content (e.g., Kreps et al. (2020); Diez-Gracia et al. (2023). Second, AI as a countermeasure comprises research on automated fact-checking, claim verification, and detection systems (e.g. Cuartielles et al. (2023); (Gonçalves et al., 2024);Kuznetsova et al. (2025). Third, regulatory and governance frameworks address legal and policy responses, most prominently the EU AI Act's risk-based classification of election-relevant AI systems (Lognoul, , 2025) and comparative analyses of divergent national governance models (Romanishyn et al., 2025). Fourth, a substantial deepfake-specific cluster treats synthetic audio-visual media as a distinct sub-field with its own technical, legal, and psychological literature (Dan, 2025; Gambín et al., 2024; Łabuz & Nehring, 2024; Momeni, 2024) ifth, algorithmic and media literacy scholarship examines public resilience, inoculation, and educational interventions (Linegar et al., 2024).

Conceptually, the corpus draws on several recurring frameworks: information disorder theory, which distinguishes disinformation, misinformation, and malinformation by intent and veracity Claire et al. (2026); the "liar's dividend" concept, describing how the mere existence of deepfake technology allows bad-faith actors to dismiss authentic evidence as fabricated Chesney & Citron (2018); dual-use technology framing, emphasizing AI's simultaneous capacity to generate and detect disinformation Grub & Humprecht (2025); García et al. (2026) and risk-based regulatory theory, underpinning instruments such as the EU AI Act. A notable conceptual gap, echoed across multiple reviews, is the limited integration of classical political communication theory (e.g., agenda-setting, framing, elaboration likelihood) with computational and machine-learning perspectives on disinformation (Grub & Humprecht, 2025).

RQ3 — AI Technologies Associated with Political Disinformation

Across the final corpus, five categories of AI technology were most frequently implicated. Deepfakes — synthetic audio-visual media generated using generative adversarial networks (GANs), autoencoders, or diffusion models — were the single most studied technology, appearing in over 40% of the corpus Łabuz & Nehring (2024); Gambín et al. (2024); (Momeni, 2024); (Dan, 2025) eflecting their high perceived salience and their tangible use in documented election-related incidents. Large language models and generative text systems were the second most common referent, examined for their capacity to produce fluent, persuasive, and difficult-to-detect political falsehoods and to power conversational disinformation delivery (Kidd & Birhane, 2023; Kreps et al., 2020; Kuznetsova et al., 2025). Social bots and automated/coordinated inauthentic accounts constituted a third recurring technology category, studied particularly in the context of amplification dynamics during geopolitical crises, such as the coordinated propaganda and misinformation activity documented on major platforms during the Russian invasion of Ukraine (Pierri et al., 2023). Fourth, algorithmic recommendation and micro-targeting systems — the machine-learning models underlying platform content curation and political advertising delivery — were frequently discussed as amplification and personalization mechanisms rather than as content-generation tools per se. Fifth, AI-based detection and verification technologies (automated fact-checking classifiers, claim-matching systems, and LLM-based veracity assessment tools) represented the principal counter-technology category, though with documented performance limitations, particularly for multilingual and cross-topic claims (Kuznetsova et al., 2025).

RQ4 — Impacts on Democratic Governance and Electoral Integrity

The reviewed literature converges on several documented and hypothesized impacts, while also revealing important evidentiary limits. First, erosion of institutional and media trust emerges as the most consistently reported impact: the mere plausibility of AI-generated synthetic content appears to reduce public confidence in authentic political information and news media more broadly, independent of whether any specific deepfake is encountered (Chesney & Citron, 2018; Dan, 2025). Second, reputational attacks on individual political candidates constitute a well-evidenced micro-level impact; experimental research shows that deepfake-based scandal content produces measurable shifts in voting intentions, attitudes, and emotional responses toward targeted politicians, with video-based fakes producing significantly larger effects than text-only equivalents (Dan, 2025). Third, the corpus repeatedly identifies the "liar's dividend" as a systemic governance risk: because synthetic media are now technically plausible, authentic incriminating evidence against political actors can be more easily dismissed as fabricated, undermining accountability mechanisms independent of any actual deepfake being deployed (Chesney & Citron, 2018; Łabuz & Nehring, 2024) Fourth, several studies document risks to electoral administration and information integrity more broadly, including fabricated claims of voter fraud, synthetic candidate statements, and AI-amplified coordinated propaganda during politically sensitive events such as the Russian invasion of Ukraine (Pierri et al., 2023). Fifth, at the level of individual cognition, generative AI systems have been shown to convey and reinforce biases and false beliefs to users through fluent, confidently phrased output that exploits heuristic trust in machine-generated text (Kidd & Birhane, 2023). Notably, however, the corpus reveals a persistent gap between documented technical capability and demonstrated causal electoral impact: while deepfakes and AI-generated text are readily shown to influence attitudes and short-term perceptions in controlled experimental settings, robust field evidence directly linking AI-generated disinformation to actual vote-switching or election outcomes remains limited, a caveat explicitly noted across multiple reviews (López-Borrull & Lopezosa, 2025; Wack et al., 2025).

RQ5 — Mitigation Strategies

Proposed and evaluated mitigation strategies cluster into four broad categories. First, technical detection and automated fact-checking approaches — including AI-based deepfake classifiers, provenance/watermarking systems, and LLM-assisted claim verification — represent the most extensively studied countermeasure, though evaluation studies consistently report imperfect and topic-dependent accuracy, with some evidence that poorly calibrated AI fact-checks can paradoxically reduce belief in true content mislabeled as false (Cuartielles et al., 2023; Gonçalves et al., 2024; Kuznetsova et al., 2025). Second, platform governance measures — content labelling, account authentication, algorithmic down-ranking of low-provenance content, and coordinated inauthentic behavior removal — are widely recommended, though the corpus notes substantial variation in platform enforcement consistency and transparency (López-Borrull & Lopezosa, 2025). Third, regulatory and legal instruments constitute a rapidly growing mitigation category, most prominently the European Union's Artificial Intelligence Act Mökander et al. (2021, p. 241), which classifies AI systems capable of influencing elections or generating synthetic political media as high-risk and imposes transparency, labelling, and human-oversight obligations. Comparative policy analysis indicates that regulatory approaches diverge substantially across jurisdictions, ranging from the EU's risk-based statutory model to more decentralized, platform-led self-governance approaches and, in some contexts, state-directed content control, each carrying distinct trade-offs between disinformation mitigation and protection of free political expression (Romanishyn et al., 2025). Fourth, media and algorithmic literacy interventions — including AI-assisted "pre-bunking" or inoculation content designed to build cognitive resistance to false election narratives before exposure — show promising experimental effectiveness Linegar et al. (2024), complementing corrective (post-exposure) fact-checking approaches. Across all four categories, the reviewed literature converges on the conclusion that no single mitigation strategy is sufficient in isolation; effective responses are consistently characterized as requiring coordinated, layered action combining technical, regulatory, platform-level, and educational interventions (García et al., 2026; López-Borrull & Lopezosa, 2025).

DISCUSSION

Interpreting the Outcomes of Research Endeavors

The findings of this systematic review indicate that scholarship on AI-driven political disinformation has matured from a largely speculative, conceptual literature into an empirically grounded and rapidly diversifying interdisciplinary field. The sharp increase in publication output following the late-2022 diffusion of consumer-facing generative AI tools (RQ1) suggests that the

research agenda has been substantially reactive to technological availability rather than purely theory-driven — a pattern with implications for the field's capacity to anticipate, rather than merely respond to, emerging risks. The five-cluster thematic structure identified here (RQ2) closely parallels that reported in adjacent reviews López-Borrull & Lopezosa (2025); García et al. (2026) lending convergent validity to the classification, while also revealing that the dual-use framing of AI — as both threat and countermeasure — has become the field's dominant organizing logic. With respect to technology (RQ3), the disproportionate scholarly attention devoted to deepfakes, relative to arguably more pervasive threats such as LLM-generated text and algorithmic micro-targeting, may partly reflect the higher visual salience and media newsworthiness of synthetic video, rather than a calibrated assessment of comparative real-world impact. This suggests a possible mismatch between research emphasis and threat magnitude that merits further empirical attention.

Evaluation in Relation to Antecedent Studies

The impacts identified under RQ4 — trust erosion, reputational harm, the liar's dividend, and belief distortion — are broadly consistent with theoretical predictions advanced in foundational deepfake scholarship Chesney & Citron (2018) and with experimental findings on synthetic media's persuasive effects (Dan, 2025). However, this review reinforces a caution already voiced in prior syntheses Kuznetsova et al. (2025); Grub & Humprecht (2025) the evidentiary base linking AI-generated disinformation to demonstrable electoral outcomes, as opposed to attitudinal or perceptual shifts in controlled settings, remains thin. This gap between demonstrated technical capability and demonstrated causal political consequence represents one of the field's most consequential unresolved questions, and it echoes a broader methodological tension in disinformation research between ecological validity and causal identification. Regarding mitigation (RQ5), the finding that no single strategy is sufficient in isolation aligns with the multi-stakeholder governance models advocated in comparative policy literature (Romanishyn et al., 2025) and with the layered technical–regulatory–educational approach implicit in the EU AI Act (Lognoul, 2025). This review's synthesis extends that consensus by explicitly linking mitigation effectiveness to the specific AI technology and disinformation vector at issue (e.g., detection classifiers for deepfakes, provenance labelling for LLM text, coordinated-behavior removal for social bots), a level of technology-specific granularity that is often collapsed in broader disinformation reviews.

Ramifications of the Discoveries

For scholars, these findings suggest that future research should move beyond feasibility demonstrations ("can AI generate/detect X?") toward causal, field-based, and cross-national comparative designs capable of establishing real-world electoral impact and differential mitigation effectiveness. For platforms and technology developers, the results underscore the need for provenance and authentication infrastructure (e.g., content credentials, watermarking) that addresses LLM-generated text with the same urgency currently reserved for deepfakes. For policymakers, the divergence in national regulatory approaches documented here highlights both the innovation potential and the coordination risk of fragmented governance; harmonization efforts, or at minimum interoperability standards, may be necessary to prevent regulatory arbitrage. For electoral management bodies and civil society organizations, the demonstrated (if partial) effectiveness of pre-bunking and inoculation interventions Linegar et al. (2024) suggests that proactive, pre-exposure media literacy investment may offer a more resilient complement to reactive fact-checking.

Limitations of the Research

Several limitations should be acknowledged. First, the search was restricted to English-language publications indexed in Scopus, Web of Science, and Google Scholar, which may have excluded relevant scholarship published in other languages or in non-indexed regional and grey literature, potentially underrepresenting disinformation dynamics in non-Western electoral contexts. Second, as with any systematic review of a fast-moving field, the corpus reflects a search conducted at a fixed point in time (30 June 2026); given the pace of generative AI development, subsequent publications may alter the thematic and technological distribution reported here. Third, the narrative synthesis approach, necessitated by the methodological heterogeneity of included studies, precludes quantitative meta-analytic estimation of effect sizes across studies; findings regarding impact magnitude should therefore be interpreted as indicative rather than statistically pooled. Fourth, thematic coding, while grounded in a structured extraction matrix and cross-validated against prior published reviews, retains an inherent degree of interpretive judgment common to qualitative synthesis methods. Finally, the exclusion of non-political disinformation research (e.g., health or scientific disinformation) — a deliberate scoping decision to preserve analytical focus — means that cross-domain lessons from those adjacent literatures were not systematically incorporated.

CONCLUSION

This systematic literature review set out to map how research on political disinformation in the age of artificial intelligence has evolved, to identify its dominant themes and conceptual frameworks, to catalog the AI technologies most implicated, to assess documented impacts on democratic governance and electoral integrity, and to synthesize proposed mitigation strategies. Drawing on a PRISMA-guided synthesis of 37 studies selected from an initial pool of 473 identified records, the review confirms that this field has grown rapidly and consolidated around a dual-use conceptualization of AI as both a driver and a potential corrective of political disinformation.

The evidence indicates that deepfakes, generative text, social bots, and algorithmic amplification systems are the principal technological vectors of concern, and that their consequences for democratic governance — reduced institutional trust, targeted reputational harm, and the corrosive "liar's dividend" — are conceptually well established even as direct causal evidence of electoral outcome change remains comparatively limited. Mitigation efforts spanning technical detection, platform governance, regulation, and media literacy each show partial effectiveness, reinforcing the conclusion that layered, multi-stakeholder responses are necessary rather than optional. In this sense, the findings of this review both corroborate and extend the concerns articulated in the contextual and theoretical framing set out in Section 1: the anticipated risks of AI-enabled political disinformation identified in the introduction are, on the whole, substantiated by the empirical and policy literature synthesized here, while the review also clarifies

where confident causal claims are, and are not yet, empirically warranted.

The significance of these findings extends beyond academic interest. As generative AI capabilities continue to advance and diffuse globally, the resilience of democratic institutions and electoral processes will depend substantially on the extent to which detection technologies, regulatory frameworks, platform accountability mechanisms, and public media literacy keep pace with — rather than perpetually lag behind — the technologies capable of producing political disinformation. Future research should prioritize cross-national comparative designs, longitudinal and field-experimental methods capable of establishing causal electoral impact, and multilingual detection capability, particularly for under-studied electoral contexts outside North America and Western Europe.

CONFLICT OF INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. [Amend as applicable to disclose any actual conflicts of interest.

REFERENCES

Chesney, R., & Citron, D. K. (2018). *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*. https://doi.org/10.2139/ssrn.3213954

Claire, W., Hossein, D., & Europarat. (2026). *Information disorder toward an interdisciplinary framework for research and policymaking*. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

Cuartielles, R., Ramón, X., & Pont-Sorribes, C. (2023). Retraining fact-checkers: The emergence of ChatGPT in information verification. *El Profesional de La Informacion*. https://doi.org/10.3145/epi.2023.sep.15

Dan, V. (2025). Deepfakes as a Democratic Threat: Experimental Evidence Shows Noxious Effects That Are Reducible Through Journalistic Fact Checks. *The International Journal of Press/Politics*, *31*(3), 525–550. https://doi.org/10.1177/19401612251317766

Diez-Gracia, A., Sánchez-García, P., & Martín-Román, J. (2023). Disintermediation and disinformation as a political strategy: use of AI to analyse fake news as Trump's rhetorical resource on Twitter. *El Profesional de La Informacion*. https://doi.org/10.3145/epi.2023.sep.23

Ferrara, E. (2024). Charting the Landscape of Nefarious Uses of Generative Artificial Intelligence for Online Election Interference. In *SSRN Electronic Journal*. RELX Group (Netherlands). https://doi.org/10.2139/ssrn.4883403

Filimonau, V., & Magklaropoulou, A. (2020). Exploring the viability of a new 'pay-as-you-use' energy management model in budget hotels. *International Journal of Hospitality Management*, *89*, 102538–102538. https://doi.org/10.1016/j.ijhm.2020.102538

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

García, J. C., Rodríguez, A. S., & Vázquez, A. I. R. (2026). Artificial Intelligence and Disinformation: A State-of-the-Art Review Through a Systematized Literature Review. *Social Sciences*, *15*(4), 247–247. https://doi.org/10.3390/socsci15040247

Gonçalves, A., Torre, L., Oliveira, F. A. G. S., & Jerónimo, P. (2024). AI and Automation's Role in Iberian Fact-checking Agencies. *El Profesional de La Informacion*, *33*(2). https://doi.org/10.3145/epi.2024.0212

Grub, M. F., & Humprecht, E. (2025). Generative AI and Disinformation| Defining the Role(s) of AI in Disinformation Research—A Systematic Review. *DOAJ (DOAJ: Directory of Open Access Journals)*. https://doaj.org/article/414282e74a4e4ec3b9fe38cf391de3b8

Hoof, M. van, Vreese, C. de, & Farooq, A. (2025). Sometimes it's OK: How Citizens Understand and Experience the Use of AI-generated Disinformation in Elections. In *OSF Preprints (OSF Preprints)*. Center for Open Science. https://osf.io/ykh9p

Kidd, C., & Birhane, A. (2023). How AI can distort human beliefs. *Science*, *380*(6651), 1222–1223. https://doi.org/10.1126/science.adi0248

Kreps, S., McCain, M., & Brundage, M. (2020). All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. *Journal of Experimental Political Science*, *9*(1), 104–117. https://doi.org/10.1017/xps.2020.37

Kuznetsova, E., Vitulano, I., Makhortykh, M., Stolze, M., Nagy, T., & Vziatysheva, V. (2025). Fact-checking with Generative AI: A Systematic Cross-Topic Examination of LLMs Capacity to Detect Veracity of Political Information. In *ArXiv.org*. https://doi.org/10.48550/arxiv.2503.08404

Łabuz, M., & Nehring, C. (2024). On the way to deep fake democracy? Deep fakes in election campaigns in 2023. *European Political Science*, *23*(4), 454–473. https://doi.org/10.1057/s41304-024-00482-9

Linegar, M., Sinclair, B., Linden, S. van der, & Alvarez, R. M. (2024). Towards Generalizable AI-Assisted Misinformation Inoculation: Protecting Confidence Against False Election Narratives. In *ArXiv.org*. https://doi.org/10.48550/arxiv.2410.19202

Lognoul, , M. (2025). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act – AI Act). *Repository of the University of Namur*. https://researchportal.unamur.be/en/publications/e333a7f0-94c5-4f46-b675-abf7201d35c7

López-Borrull, A., & Lopezosa, C. (2025). Mapping the Impact of Generative AI on Disinformation: Insights from a Scoping Review. *Publications*, *13*(3), 33–33. https://doi.org/10.3390/publications13030033

Martens, B., Aguiar, L., Gómez-Herrera, E., & Mueller-Langer, F. (2018). The Digital Transformation of News Media and the Rise of Disinformation and Fake News. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3164170

Mökander, J., Axente, M., Casolari, F., & Floridi, L. (2021). Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation. *Minds and Machines*, *32*(2), 241–268. https://doi.org/10.1007/s11023-021-09577-4

Momeni, M. (2024). Artificial Intelligence and Political Deepfakes: Shaping Citizen Perceptions Through Misinformation. *Journal of Creative Communications*, *20*(1), 41–56. https://doi.org/10.1177/09732586241277335

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T., Mulrow, C. D., Shamseer, L., Tetzlaff, J., Akl, E. A., Brennan, S., Chou, R., Glanville, J., Grimshaw, J., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, *372*. https://doi.org/10.1136/bmj.n71

Pennycook, G., & Rand, D. G. (2021). The Psychology of Fake News. *Trends in Cognitive Sciences*, *25*(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

Pierri, F., Luceri, L., Jindal, N., & Ferrara, E. (2023). *Propaganda and Misinformation on Facebook and Twitter during the Russian Invasion of Ukraine*. 65–74. https://doi.org/10.1145/3578503.3583597

Rajput, M. (2026). The Critical Threat of Deepfakes: Vulnerability and Resilience in Democratic Elections. *International Journal For Multidisciplinary Research*, *8*(1). https://doi.org/10.36948/ijfmr.2026.v08i01.63435

Romanishyn, A., Malytska, O., & Goncharuk, V. A. (2025). AI-driven disinformation: policy recommendations for democratic resilience. *Frontiers in Artificial Intelligence*, *8*, 1569115–1569115. https://doi.org/10.3389/frai.2025.1569115

Theocharis, Y., & Jungherr, A. (2020). Computational Social Science and the Study of Political Communication. *Political Communication*, *38*, 1–22. https://doi.org/10.1080/10584609.2020.1833121

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Wack, M., Ehrett, C., Linvill, D. L., & Warren, P. L. (2025). Generative propaganda: Evidence of AI's impact from a state-backed disinformation campaign. *PNAS Nexus*, *4*(4). https://doi.org/10.1093/pnasnexus/pgaf083