International Journal of Communication 19(2025), 3577–3601 1932–8036/20250005

Defining the Role(s) of AI in Disinformation Research—A Systematic Review

MARIA F. GRUB
EDDA HUMPRECHT
Friedrich-Schiller University Jena, Germany

This study examines the role of artificial intelligence (AI) in the generation, dissemination, detection, and correction of online disinformation in political contexts. We conducted a systematic review of N = 123 scientific contributions from social and computer science to explore how AI contributes to the spread of disinformation, for example, through deepfakes and social bots, and its potential role in combating these challenges via detection algorithms and fact-checking systems. The review identifies significant gaps in empirical research, especially about the effectiveness of detection models in real-world applications and the limited exploration of multilingual and multimodal disinformation. Furthermore, the field has yet to fully integrate theoretical frameworks from communication science with computational perspectives, which is crucial for a comprehensive understanding of AI's role in disinformation. The study underscores the need for interdisciplinary approaches that bridge computer and social sciences to better address the societal and political implications of AI-driven disinformation.

Keywords: disinformation, artificial intelligence, AI-driven disinformation, detection, dissemination

The spread of disinformation—"false, inaccurate, or misleading information that is shared with the intent to deceive the recipient" (Bontridder & Poullet, 2021)—threatens democracy, potentially polarizing society and fueling extremist views (Whyte, 2020). As more people rely on online rather than traditional media to stay informed (Landon-Murray, Mujkic, & Nussbaum, 2019), social media platforms play a crucial role in the dissemination and generation of disinformation. Most social media platforms and big tech companies (e.g., Google, Meta, OpenAI) rely on artificial intelligence (AI), for instance, for content moderation or recommender systems (Nah, Zheng, Cau, Siau, & Chen, 2023). AI shapes user experience; engagement-driven algorithms influence content trustworthiness and enable disinformation (Strömbäck et al., 2020). Because of limited control and detection mechanisms for AI-generated content, social media significantly contributes to the rapid dissemination of disinformation (Aïmeur, Amri, & Brassard, 2023), placing platforms under growing pressure to address this issue (Reisach, 2021).

Maria F. Grub: maria.grub@uni-jena.de
Edda Humprecht: edda.humprecht@uni-jena.de
Date submitted: 2025-02-20

Defining AI and assessing its empirical impact remain difficult because of the absence of a widely accepted conceptualization. Beyond technical workings, we must understand AI's relationship with humans, its communication patterns, and its scientific, ethical, societal, and political implications (Gil de Zúñiga, Goyanes, & Durotoye, 2024). This presents both theoretical and empirical challenges that no single research tradition can address alone (Schäfer, 2023). Consequently, AI needs to be investigated across various research fields, particularly at the intersection of computer and social sciences. In this article, we understand AI as "the ability of a system to perform tasks characteristic of human intelligence, such as learning and decision-making" (Kertysova, 2018, p. 57). This includes any self-learning system based on machine learning (ML), large language models (LLMs), or deep neural networks (DNNs) that contributes to the generation, dissemination, or detection of online disinformation (e.g., generative AI, deepfake technology, social bots). While we provide a working definition of artificial intelligence as a class of technologies capable of performing tasks typically requiring human intelligence, we recognize that in interdisciplinary research—such as concepts like "platforms" or "algorithms"—a unified definition may not be necessary or even beneficial. Rather than treating AI as a singular object, it is more productive to consider how specific technologies labelled "AI" are used in varying societal and political contexts. Our review, therefore, focuses less on technological essence and more on the roles and applications of AI as studied across disciplines.

AI and "fake news" have a traceable effect on the democratic and political understanding of individuals (Aïmeur et al., 2023), as we have witnessed during major political events such as the rapid spread of misleading health information during the Covid-19 pandemic (Bontridder & Poullet, 2021), social bot campaigns during the 2016 U.S.-presidential elections (Bessi & Ferrara, 2016), or more recently, the algorithmic promotion of far-right candidates on TikTok in the 2024 Romanian presidential elections (Kirby & Thorpe, 2024). Furthermore, technologies like deepfakes and generative AI have been used in various election contexts worldwide—for example, in India (Tomar, Nihal, Singh, Marwaha, & Tiwani, 2023), Nigeria (Ekpang, Iyorza, & Ekpang, 2023), and Brazil (Benevenuto & Melo, 2024)—to manipulate voters in favor of certain parties and candidates or encourage abstention from voting altogether (Łabuz & Nehring, 2024). With new AI technology emerging at a fast pace, the possibilities for disinformation to reach and deceive the population are only expected to increase (Strömbäck et al., 2020). Research on AI-driven disinformation has primarily focused on democracies located in the Global North (Kertysova, 2018; Whyte, 2020); however, as the examples above show, it is crucial to consider how the threat of AI-driven disinformation is perceived in other countries, especially in the Global South, as media systems and their resilience to mis- and disinformation vary across regions (Humprecht, Esser, & Van Aelst, 2020). Previous studies have, for example, examined resilience to disinformation, highlighting differences in susceptibility between countries and the significant role of trust (Boulianne & Humprecht, 2023).

While AI is a traceable contributor to the dissemination and generation of disinformation, it is also praised as a potential solution for identifying and counteracting online disinformation, for example, by flagging or automatically fact-checking information (Pathak, Srihari, & Natu, 2021). However, strategies to combat AI-driven disinformation are often context-dependent, and their success hinges on addressing different regions' unique needs and characteristics (Humprecht et al., 2020). This is particularly important

if we regard concepts like algorithm literacy and how we can effectively prepare citizens to navigate a growingly algorithm-governed online environment (Gagrčin, Naab, & Grub, 2024).

Despite the growing need to effectively detect, understand, and regulate AI-driven disinformation, there is currently no systematization of disinformation research, and potential blind spots exist in addressing AI's role regarding disinformation in political communication (e.g., Lee & Shin, 2022). The present study aims to expand the current knowledge of AI's role in the dissemination and detection of online disinformation by answering the following research question: *How do existing studies conceptualize and investigate the role of AI in the dissemination and detection of online disinformation?*

We conducted a systematic review of 123 scientific contributions in political communication research, focusing on articles that intersect computer science and social science to address AI and online disinformation in political contexts. The following sections contextualize our research interests, present the literature review results, and outline implications for future research.

Research Interests

Research on AI and disinformation lies at the intersection of computer and social sciences (Aïmeur et al., 2023). It is questionable, though, whether current approaches to studying disinformation are sufficient to fully grasp the impact of artificial intelligence. Computer science relies primarily on data articles, which refer to articles that describe datasets or use a dataset to train (and test) statistical models (e.g., disinformation detection). They often do not present or discuss theory but focus on the development and advancement of computational detection models. Communication science examines the impact of AI, applying traditional methods with both quantitative (e.g., content analysis) and qualitative (e.g., interviews) approaches (Kessler, Mahl, Schäfer, & Volk, 2025). However, AI and its impact hardly fit into current communication paradigms (Schäfer, 2023); hence, we need to bridge both research traditions by advancing current theoretical thinking and incorporating new methods into the field. This includes exploring different types of data (e.g., online data from social media, survey data); uncovering new and effective ways to analyze it; identifying key players in the field of AI-driven disinformation; understanding user perceptions, attitudes, and behaviors toward disinformation; and examining how individuals process and respond to AI-generated content (Ahmed, 2023). This systematic review aims to uncover how the two fields of study are—or, more precisely, can be—connected to assess the effects and impact of AI by exploring the research questions outlined in the following section.

The Dual Role of AI

The way AI influences online disinformation is multifaceted. Current research recognizes two distinct roles of AI: as a source of the disinformation problem and as a potential solution (Makhortykh, Sydorova, Baghumyan, Vziatysheva, & Kuznetsova, 2024). On the one hand, AI technologies such as deepfakes or social bots enable the rapid spread of disinformation, complicating efforts to detect and counteract it. On the other hand, AI offers tools like fact-checking systems and detection algorithms to combat disinformation (Weikmann & Lecheler, 2023). This dual role raises essential questions about the implications of AI's use and how it can be harnessed effectively. Additionally, it is crucial to understand

which modalities of AI-disinformation are being analyzed in current research—namely, textual, visual, audio, or multimodal—as this will help clarify the specific mechanisms through which AI contributes to and mitigates the spread of false information.

RQ1: What is the role of AI in the field of disinformation?
RQ2: What modalities of AI-disinformation are being analyzed?

Because of its rapid development, AI technology can create fake content that is almost indistinguishable from real information (Bontridder & Poullet, 2021). So-called "deepfakes"—multimodal material (audio and/or visual) that has been digitally manipulated—are spread online and are almost impossible to detect by untrained individuals (Kertysova, 2018). For example, Lee and Shin (2022) analyzed the credibility and engagement intentions of fake news and found that both significantly increased with higher source vividness, as with deepfakes.

Apart from being a tool for generating disinformation, AI also plays a vital role in its dissemination, for example, through micro-targeting specific interest groups (Bontridder & Poullet, 2021) or using social bots or troll farms to enhance the spread of online disinformation (Aïmeur et al., 2023). Social bots are fully or semi-automated non-human actors that autonomously mimic human communication on digital platforms (Bontridder & Poullet, 2021). They can also be programmed to perform harmful actions, such as fostering disinformation and conflict (Hajli, Saeed, Tayvidi, & Shirazi, 2022). Hajli et al. (2022) report that a person's online opinion is affected by interactions with bots, as social bots provide people with disinformation or escalate arguments on social media. Furthermore, AI and algorithms are affected by algorithmic bias, meaning that human biases are reflected in algorithmic outputs and can become part of automated processes, leading to reduced human control over AI output and the unrestricted spread of biased information (Kertysova, 2018). To grasp the scope of AI in political communication, we identify the dimensions of AI-generated online disinformation and the potential outcomes of AI-driven disinformation, such as its effect on political campaigns and elections, trust in democratic institutions, and news credibility.

RQ3: How is online disinformation generated and/ or disseminated?
RQ4: What are the effects and impacts of AI-generated disinformation on citizens and society?

As mentioned before, AI is not only seen as a cause of online disinformation but also as a potential solution that could supplement human-based approaches to combat disinformation, such as fact-checking (Aïmeur et al., 2023). New AI technologies are being developed to detect and regulate online disinformation, for example, by filtering and removing false information, blocking or suspending accounts that share disinformation, or (de)prioritizing certain content online (Bontridder & Poullet, 2021). Disinformation detection models use machine learning (ML) or deep learning (DL) algorithms to identify false information, manipulated content, or social bots responsible for the spread of disinformation. The models can be differentiated in their approach (knowledge-based, feature-based, and learning-based) and mode of classification (e.g., ensemble methods, kernelized models, linear models) (Kaliyar, Goswami, & Narang, 2021). Different approaches and classifiers are employed depending on the complexity of the data and the required accuracy. Choosing one model over another involves a trade-off between performance (predictive power) and explainability. For example, DL models like Convolutional Neural Networks (CNNs) or Generative

Adversarial Networks (GANs) can achieve outstanding performance; however, their decision-making process is considered a "black box," making them difficult to interpret and ethically questionable for flagging content as disinformation. Linear models are high in explainability but lack the performance ratings of other methods (Gongane, Munot, & Anuse, 2024). Deciding on a model involves considering whether to prioritize high performance or interpretability.

RQ5: What are the different functionalities of detection models?

However, as a study by Paschen, Kietzmann, and Kietzmann (2019) found, AI shows higher potential in creating disinformation than in distinguishing false from accurate news. We review current state-of-the-art detection models that aim to identify online disinformation and propose countermeasures accurately. For this evaluation, different reliability metrics for measuring the correctness of a model's predictions—accuracy, precision, recall, and F1-score—will be considered (Rasyid, Sibaroni, & Ihsan, 2023). A detailed overview of the reliability metrics is provided in the supplementary material.[1] The results of applicable detection studies (i.e., studies that report at least one of the metrics beforehand) will be compared to assess the effectiveness of different approaches.

RQ6: How effective are the detection and resulting countermeasures of AI-disinformation?

Review Method

We conducted a systematic review of the literature at the intersection of AI and disinformation research. We chose this method because it allows for a structured, category-based synopsis of current knowledge, helps identify research gaps, and bridges existing literature, which is especially valuable in such an interdisciplinary field.

Data Collection

The data collection comprised two distinct steps (see Figure 1). First, we conducted a database search of Web of Science (WoS), WoS Theses and Dissertations Database, Google Scholar, and Association for Computing Machinery (ACM) Digital Library for records between January 2005 and April 2024. These databases cover a wide array of research in both social and computer sciences, and their combination has previously yielded promising results (Valente, Holanda, Mariano, Furata, & Da Silva, 2022). Because of the vast output on Google Scholar, only publications from 2022 onward were included from this database. We identified peer-reviewed, English-language scientific articles and conference papers in the fields of social and behavioral science and computer science. Furthermore, as PhD supervision is close to peer review, dissertations were also included in the sample.

A search string combining search terms related to both fields was created using the block-building method (Guba, 2008) to screen titles, abstracts, and keywords. The search string includes disinformation and AI as key interests. Because not all research areas use the term disinformation to describe the

[1] https://osf.io/hfyue/?view_only=9d2375b789a04df8a28d7c5e299bdfd2

intentional spread of false information, various synonyms and related terms (e.g., fake news, misinformation) were included. Similarly, to grasp a complete picture of AI, various terms used in computer science were added. The search string was tested in different databases and adapted multiple times to ensure robust results for the final search query. The final search string was as follows: (disinformation OR misinformation OR "fake news" OR propaganda) AND (AI OR "Artificial Intelligence" OR "machine learning" OR "deep learning" OR "deep fakes" OR "generative AI" OR Automation OR bots OR "natural language processing" OR NLP). The search string yielded a total of 1,351 records. After removing duplicates and applying the formal criteria of publication date, language, and peer review to the output, 1,084 articles remained for abstract screening according to preset inclusion and exclusion criteria. Records were eligible for further analysis if they were related to political communication, disinformation, and AI. Political communication thus includes communication by political actors, addressed to political actors, and communication about political actors and events. Political actors include individuals, parties, non-party organizations (including NGOs and corporate lobby groups), pressure groups, terrorist organizations, receivers of political communication ("audiences"), and media organizations (McNair, 2017). We define disinformation according to Bontridder & Poullet (2021). Only articles that address disinformation are included in the sample, even if the terminology varies (e.g., defining "misinformation" as false information spread to deceive). AI is defined as technology that performs "tasks characteristic of human intelligence, such as learning and decision making" (Kertysova, 2018, p. 57). AI must be applied in this context to align with the study's focus. This can be in the form of the dissemination, generation, detection, and countering of disinformation by AI. However, we acknowledge that AI is an umbrella term encompassing diverse techniques, which creates challenges for empirical synthesis.

During the abstract screening stage, 50 records were double-coded by two individual coders (Krippendorff's α = .8) to assess whether a record should be included or excluded. Reliability could be determined at this stage because coding was numerical (0 = exclude, 1 = include). In the subsequent phases of coding, this was not the case, as the codebook contained open codes for the different categories. To further ensure reliability throughout the coding process, the two coders frequently met to discuss results. After reviewing the differences in coding, the inclusion and exclusion criteria were revised accordingly, and the coding proceeded. A total of 264 reports were deemed eligible for full-text analysis. Four articles could not be retrieved, so the final sample for eligibility consisted of 260 records.

In the second step of data collection, a backward (BWS) and forward search (FWS) were conducted with the database sample to identify important works at the intersection of computer and social sciences and fully grasp the interdisciplinary nature of AI-disinformation research. For BWS, articles were identified via reference lists, and a citation network was created (n = 970) to identify potential gaps in the sample (see supplementary material). The FWS was conducted by viewing the articles that cited the records in the main sample (n = 3400). All results from the BWS and FWS that were cited at least three times (n = 1490) were included for further abstract screening according to the inclusion and exclusion criteria, of which 64 remained for the full-text analysis (two articles could not be retrieved).

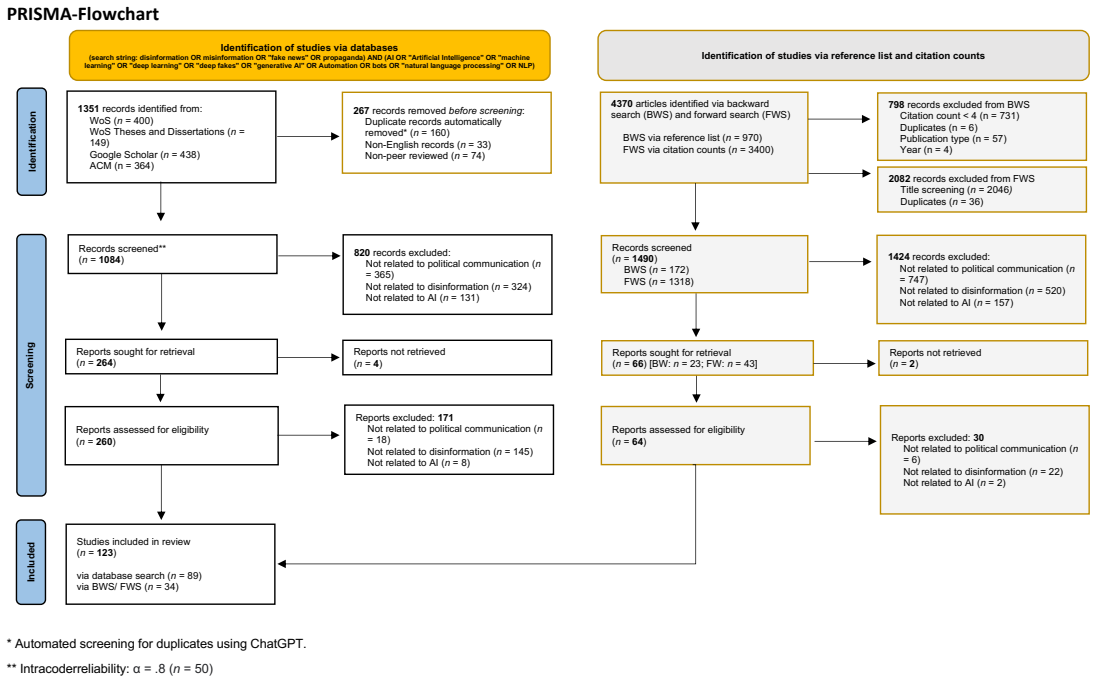

Figure 1. PRISMA-flowchart.

Data Analysis

260 articles were included for full-text analysis. Before data analysis, we developed a codebook (see supplementary material) to define the variables necessary to answer the research questions and provide coding instructions. The codebook consists of seven modules: Module A (Descriptives) includes basic details, such as the author, year, and affiliation. Module B (Contextualization) summarizes the article, focusing on the research questions, hypotheses, and main findings. Module C (Methodology) describes the type of paper (e.g., empirical, conceptual, data), the methodological approach (e.g., quantitative, qualitative, computational), and the applied method, sample, and communication channel. Module D (Theory) covers the theoretical approach, definitions of disinformation, AI's role (e.g., dissemination, detection), and central measures used in the study. Modules E and F address articles based on whether they focus on dissemination (E) or detection (F). Module E explores dissemination and generation, looking at target groups, disseminators, strategic use, impact, effect, and countermeasures. Module F focuses on detection, covering stakeholders, models, practical applications, reliability, language use, platforms, and limitations. Finally, Module G (Research Assessment) gathers the article's outlook on future research and assesses its quality.

Initially, two individual coders coded 10% of the articles (n = 25) to assess the quality of the codebook and the included articles. Unfit articles were already excluded at this stage, and the codebook was adjusted where needed. Afterward, the coders split the remaining sample for coding. Because of the open nature of the codebook, reliability was ensured through regular meetings of the coders during the coding

process to discuss results and potential ambiguities. In the final step, the coding results were sighted and systematically structured. After the coding was completed, 123 articles remained in the final sample. Further exclusions were necessary because of the inclusion and exclusion criteria; for instance, many articles employ disinformation or "fake news" as buzzwords in their publications, despite the actual research not addressing disinformation theoretically. All findings presented in the following section resulted from the coding process following the codebook (see supplementary material).

Results

Descriptive Findings

Studies first emerged in 2016 following the U.S. presidential elections (Bessi & Ferrara, 2016) and the Macron Leaks campaign in 2017 (Ferrara, 2017). Publications increased notably in 2019 and again in 2022 (Figure 2), likely driven by technological advancements such as the emergence of "deepfakes" and the global launch of ChatGPT in 2022, which raised public awareness of AI (Nah et al., 2023). Of the sample, 60.16% (n = 74) are data-driven studies that train and test statistical models but often lack theoretical grounding as assessed in the codebook (see Module D), complicating evaluation of detection model effectiveness. Conceptual (n = 11), review (n = 9), and simulation studies (n = 5) are fewer, highlighting the need for interdisciplinary approaches. Empirical studies measuring the real-world impact of AI-driven disinformation have increased in recent years (n = 25), with a focus on deepfakes (n = 10) and bot-driven dissemination (n = 5).

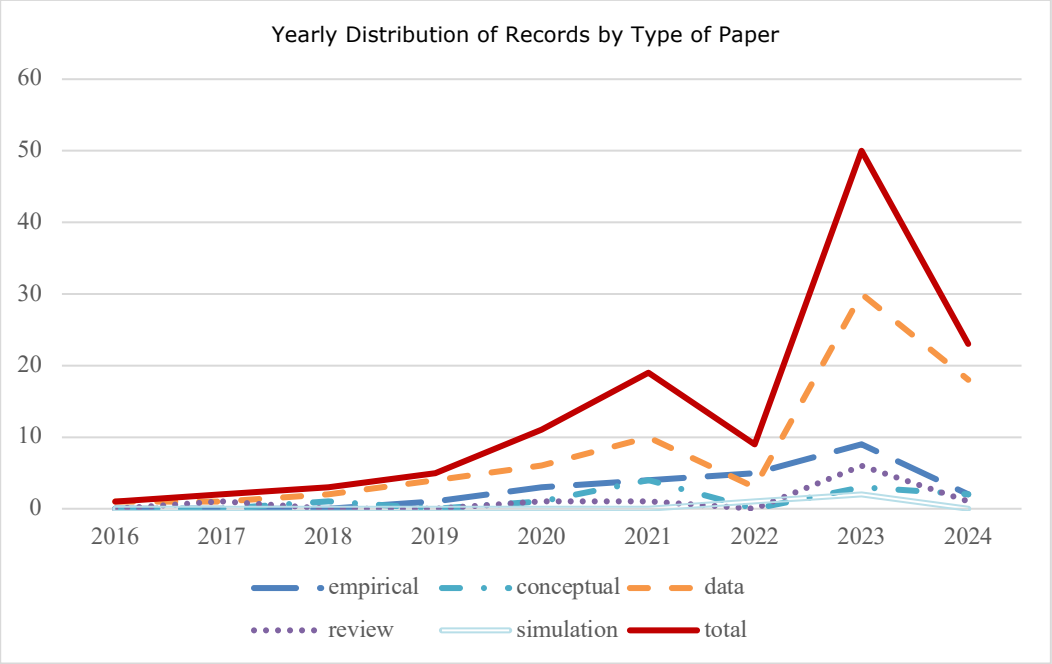

Figure 2. Yearly distribution of records by type of paper.

Computational methods (*n* = 75), particularly model testing of automated detection systems (*n* = 72), dominate (Table 1). Among quantitative approaches, surveys (*n* = 11) are the most common. Mixed-method designs (*n* = 10) traditionally combine quantitative and qualitative methods, but recent studies have integrated computational techniques with traditional methods, such as content analysis and bot detection.

Table 1. Methodological Approaches.

	Method	*n*	%
Computational	model testing	72	58.53%
(*n* = 75; 60.98%)	other (topic modeling, network analysis)	3	2.44%
Quantitative	survey	11	8.94%
(*n* = 15; 12.19%)	systematic review	2	1.63%
	simulation	2	1.63%
Qualitative	unsystematic review	7	5.69%
(*n* = 12; 9.76%)	interview	2	1.63%
	other (content analysis, case study, hermeneutic analysis)	3	2.44%
Mixed-Methods	quantitative & qualitative	2	1.63%
(*n* = 10; 8.13%)	computational & qualitative	3	2.44%
	computational & quantitative	5	4.06%
Conceptual (*n* = 11; 8.94%)		11	8.94%
	Total	123	100%

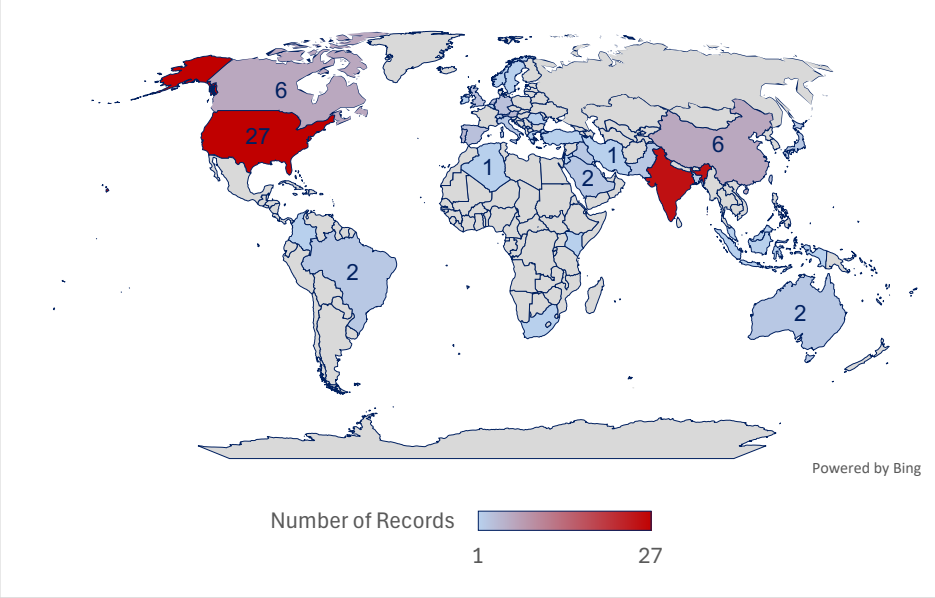

Figure 3. Geographical distribution.

The sample includes peer-reviewed records published in English from 37 countries (Figure 3), with 65 records originating from nations in the Global North (e.g., USA, Portugal) and 58 from countries in the Global South (e.g., India, Singapore). While the geographical distribution indicates a global interest in AI-driven disinformation, African (n = 3) and South American (n = 3) publications are underrepresented. Globally, detection is the most researched area of interest; however, the Global South emphasizes it more (77.59%) than the North (52.31%), which also features work on generation and multi-dimensional studies (e.g., combining generation, detection, and correction) (see Table 2). Research from both the Global South and North emphasizes U.S.-centric events. Articles from the Global North further explore AI regulations and the impact of deepfakes (e.g., Reisach, 2021).

Table 2. Distribution of Dissemination, Generation, and Detection by Country.

	Global South		Global North	
	n	%	N	%
Dissemination	7	12.07%	8	12.31%
Generation	3	5.17%	10	15.38%
Detection	45	77.59%	34	52.31%
Multiple	3	5.17%	13	20.00%
Total	58	100%	65	100%

Comparative designs are rare, with only a few cross-regional studies (e.g., Ahmed, 2023: Singapore – U.S). This finding highlights the need for more global research to understand how AI impacts disinformation across diverse political and media systems, and to adapt detection models for different contexts. Empirical work primarily focuses on deepfake perception, while other domains like sharing intention or social bot impact remain underexplored. A small number of expert studies, such as interviews with fact-checkers (Weikmann & Lecheler, 2023) and social media practitioners (Robertson & Meintjes, 2021), provide valuable but limited real-world insights.

A significant limitation is the dominance of English datasets (Table 3), which hinders testing for low-resource languages like Sindhi and Kannada, making external testing and model reliability difficult. Few studies test models across multiple languages (e.g., Al Ghamdi, Bhatti, Saeed, Gillani, & Almotiri, 2023, test in English, Arabic, and Urdu). It must be considered, though, that only English-language publications were included in the sample, so language distribution and geographical assessment might be biased because of sample selection. A complete overview of publicly available datasets, including language applications and platforms, can be found in the supplementary material.

Table 3. Language Application.

Language	Number of Records*
English	58
Arabic	5
Bengali/Bangla, French, Indonesian	3
German, Urdu	2
Algerian, Chinese, European Portuguese, Hindi, Kannada, Luxembourgish, Malaysian, Persian, Russian, Sindhi, Spanish, Swahili, Vietnamese	1

*Some records apply multiple languages

Twitter (X) is the most researched platform (n = 40), followed by Facebook (n = 10), with no notable differences between the Global North and South. Other platforms like YouTube (n = 4) or Instagram (n = 2) remain underexplored despite their association with mis- and disinformation (e.g., Jost & Dogruel, 2023). This may represent challenges in accessing data, as many detection models rely on publicly available Twitter (X) datasets. Accessing current data from various platforms remains a significant issue for researchers (de Vreese & Tromble, 2023). News website content is often derived from fact-checking websites like Politifact and global broadcasters like Reuters, BBC, and Google News, but also includes regional sources like *Prajavani* (Kannada-language newspaper in India) (e.g., Sanjana, Kuranagatti, Devisetti, Sharma, & Arya, 2023) and *The Herald Sun* (Australia) (e.g., Dao & Zettsu, 2023).

The selected records adhere to Bontridder and Poullet's (2021) definition of disinformation, highlighting intentional dissemination or generation for political, financial, or social motives to harm, erode trust, or influence public opinion. They also reflect the role of malicious actors and organized campaigns, with evolving definitions incorporating image manipulation (e.g., "Fauxtography," Ghai, Kumar, & Gupta, 2024). The interchangeable use of "Fake News" with disinformation complicates detection and leads to biased or oversimplified conclusions (Krämer, 2021). Conceptual clarity, as noted by Lange and Lechtermann (2021), is critical for adequate identification.

Theoretical frameworks span multiple fields, including communication science, psychology, behavioral science, and information systems. Established theories, such as Inoculation Theory, explain how exposure to counterarguments builds resistance to manipulation (Zerback, Töpfl, & Knöpfle, 2021), while Information Overload addresses the impact of excessive content on critical assessment (Abd, Mahdi, Jassim, & Hussain, 2023). Key frameworks include Sundar's (2008) Heuristic-Systematic Model, particularly the realism heuristic, which examines the perceived authenticity of fake content like deepfakes (e.g., Hameleers, 2024), and perceived news credibility, which explores how factors like source trustworthiness influence reliability (e.g., Weikmann, Greber, & Nikolaou, 2024). Framing Theory examines media influence on disinformation perceptions (e.g., Kuznetsova & Makhortykh, 2023), and the Theory of Planned Behavior (TPB) analyzes how attitudes and norms shape behavior toward disinformation (e.g., Lin, 2022).

Emerging theories include Explainable AI (XAI), which bridges computer and social science by investigating how to make AI models more transparent and user-friendly (Dikmen & Burns, 2022). Beyond

a few exceptions like "deepfake self-efficacy" (Weikmann et al., 2024), there is a crucial need to develop new theories to address the evolving dynamics of AI-driven disinformation.

Defining the Role(s) of AI in Disinformation Research (RQ1)

AI plays a multifaceted role in disinformation—detection, dissemination, generation, and correction of disinformation—through various AI-based tools and strategies (see Figure 4).

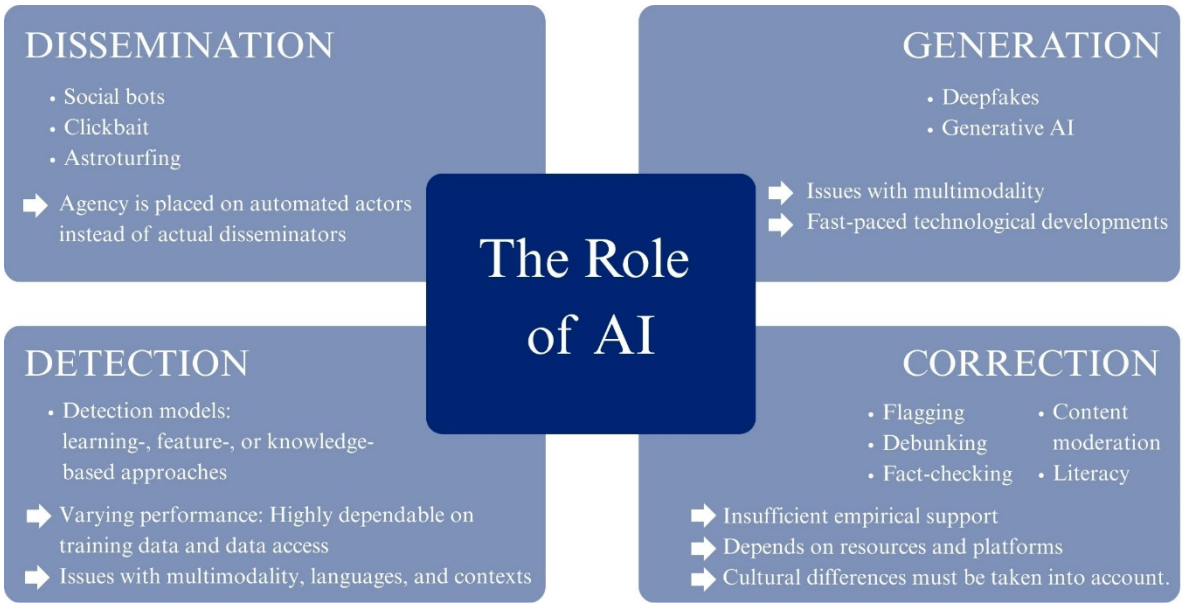

Figure 4. Taxonomy of the role of AI.

Detection dominates research, though dissemination, generation, and correction are gaining attention because of technological advances and concerns over AI's political impact. Research on dissemination has grown since 2021, examining structures of bot networks (e.g., Guzmán Rincón, Carrillo Barbosa, Segovia-García, & Africano Franco, 2022) and strategies of astroturfing (e.g., Zerback et al., 2021). Interest in generation surged from 2020, focusing on advancements in deepfake technology (e.g., Vizoso, Vaz-Álvarez, & López-García, 2021). A newer role of AI, correction, emerges alongside dissemination and detection (see Figure 5). Correction involves countering disinformation through measures such as debunking, flagging, or moderating content (Repede, 2023).

Deepfakes involve creating or manipulating digital content using GANs to produce highly realistic and difficult-to-detect media, including video, audio, and images (Ahmed, 2023; Kertysova, 2018). Terms like "synthetic media" describe such content, while "cheapfakes" refer to simpler manipulations using basic tools (Dao & Zettsu, 2023; Hameleers, 2024). Social bots are algorithm-driven entities designed to mimic human users and manipulate public opinion, especially in political contexts (Abdirahman, 2023). They employ strategies such as smoke screening—distracting discussions with related hashtags—and misdirecting attention to other topics. Bots are also used in astroturfing to create the illusion of widespread support for

specific opinions (Guzmán Rincón et al., 2022). Clickbait involves using misleading or sensational headlines to attract users to websites that spread disinformation. This strategy serves financial motives, such as generating advertising revenue, and causes public or personal harm (Ma, Chen, Chen, & Huang, 2024).

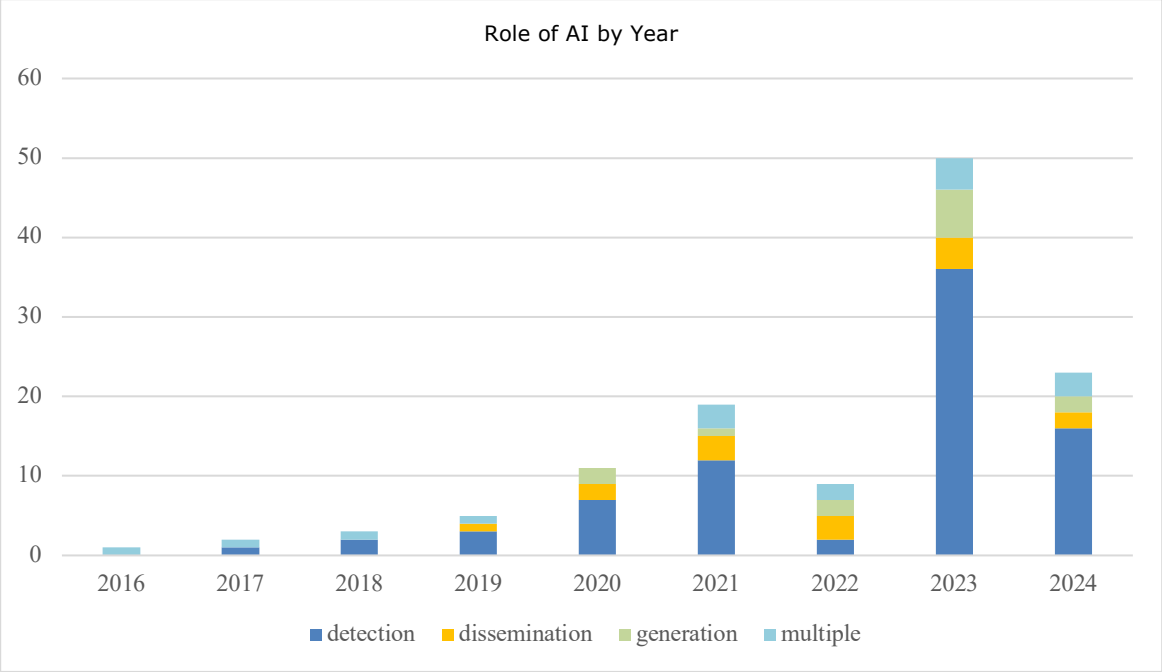

Figure 5. Role of AI by year.

Modalities of AI-driven Disinformation (RQ2)

Although AI-disinformation is generated and spread in various forms, most research focuses on textual data (n = 90), followed by multimodal data (n = 11), visual data (n = 8), and textual-visual data (n = 9). Textual analyses include the falsehood of tweets (e.g., Rasyid et al., 2023), information spread by social bots (e.g., Guzmán Rincón et al., 2022), online news articles and headlines (e.g., Afonso & Rosas, 2024), social media posts and comments (e.g., Zhang, Song, Koura, & Su, 2023), and political clickbait (e.g., Sukhramani, Kumre, Rasool, & Jadav, 2024).

Deepfakes are studied visually (e.g., Ahmed, 2023), audio-visually (e.g., Weikmann et al., 2024), and multimodally, incorporating text descriptions (e.g., Li & Wan, 2023). Despite interest in countering deepfake disinformation, detecting multimodal and audiovisual deepfakes remains limited because of a lack of training data and the challenges of integrating methods for textual, visual, and audio data. This highlights the need for more data and advanced multimodal detection algorithms.

Social bot analysis primarily focuses on detection and dissemination patterns, with limited research on bot influence and perception (e.g., Zerback et al., 2021). Phenomena like astroturfing and political clickbait remain underexplored in terms of dissemination, generation, and detection.

Dissemination and Generation of AI-Driven Dissemination (RQ3)

The records identify various target groups and disseminators of AI-driven disinformation (Figure 6). We distinguished between individual target groups (e.g., social media users and political actors), targeted institutions (e.g., governmental institutions and academic institutions), and media discourse influenced by AI-driven disinformation (Jungherr, 2025). AI-driven disinformation most often targets individuals, with political actors being most affected (e.g., Patel et al., 2023), but also includes examples of users of specific social media (e.g., Kuznetsova & Makhortykh, 2023), celebrities (e.g., Dourado, 2023), or specific demographics such as the Dutch (e.g., Hameleers, Van Der Meer, & Dobber, 2022) or Russian-speaking population (e.g., Starbird, Arif, & Wilson, 2019). The influence of media discourse includes articles analyzing the occurrence of "fake news" in established media (e.g., Al Ghamdi et al., 2023).

Records name disseminators less frequently than target groups. They include individuals like political actors, corporate entities (e.g., third-party companies), media actors (e.g., journalists, media outlets), foreign actors (e.g., the Russian Federation), or automated agents such as social bots. However, while social bots are presented as key disseminators, the origin of bot campaigns is often not transparent (Pedrazzi & Oehmer, 2020). The publications usually refer to disseminators and target groups on a conceptual level rather than assessing their strategies and impact empirically, despite evidence that AI-driven disinformation influences political debates (Bessi & Ferrara, 2016) and shapes public perceptions of issues (Pedrazzi & Oehmer, 2020). Results must be interpreted carefully, and more research is needed in this regard.

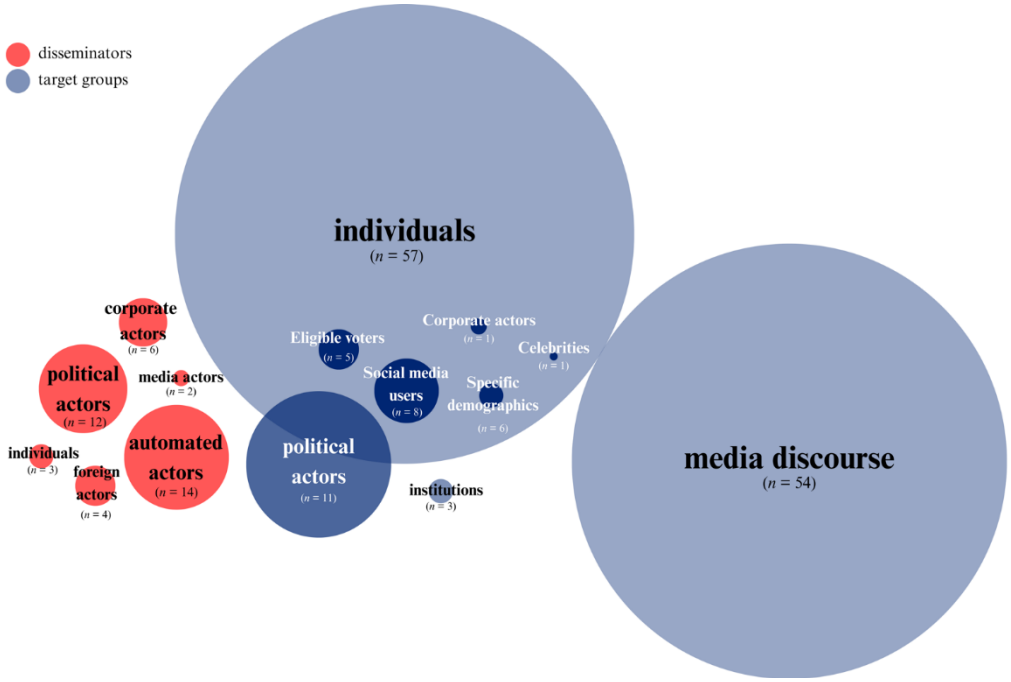

Figure 6. Target groups and disseminators of AI-driven disinformation based on conceptual occurrence.

AI-driven disinformation is spread to achieve financial, geopolitical, and economic goals (Reisach, 2021), destabilize political situations (Kertysova, 2018), and influence elections (Abdirahman, 2023). Current research primarily focuses on dissemination strategies and networks, such as the distribution of deepfakes during wars (Twomey et al., 2023) or their spread during election campaigns (e.g., Kuznetsova & Makhortykh, 2023; Santana, Nunes, & Silva, 2021).

Effects and Impact of AI-Driven Disinformation on Citizens and Society (RQ4)

Empirical findings remain limited, and generalizations must be approached cautiously. Research on deepfakes primarily focuses on news credibility, self-efficacy, and awareness. While deepfakes do not directly mislead users (Vaccari & Chadwick, 2020), their exposure reduces trust in audiovisual media (Weikmann et al., 2024) and increases uncertainty about social media's truthfulness (Vaccari & Chadwick, 2020). Though not perceived as more credible or persuasive than textual disinformation (Hameleers et al., 2022), deepfake videos enhance content vividness and boost user engagement (Lee & Shin, 2022). Awareness of deepfakes correlates with higher political interest and media consumption, while individuals often perceive others as more vulnerable to their effects than themselves (Ahmed, 2023).

Evidence shows that organized social bot networks are reused across campaigns. For example, the same botnet active during Brazil's 2014 and 2018 elections reappeared in later global disinformation campaigns (Santana et al., 2021). Emotionally bot-generated content also drives higher sharing rates

(Daume, Galaz, & Bjersér, 2023), and both astroturfing and inoculation messaging can partially influence political opinions (Zerback et al., 2021). However, research on bot perception in political contexts and its impact on election outcomes remains insufficient. Few studies empirically test countermeasures against AI-disinformation, such as false-tags (Lee & Shin, 2022), inoculation messages (Zerback et al., 2021), or "good bots" designed to counter disinformation with accurate content (Zhang et al., 2023). Most countermeasures remain theoretical, including literacy interventions (Kertysova, 2018), recovery strategies (Robertson & Meintjes, 2021), stricter regulations by governments or corporations (O'Donnell, 2021), and advancements in automated solutions like fact-checking (Santos, 2023), content moderation (Pedrazzi & Oehmer, 2020), or flagging (Repede, 2023). There is an urgent need for empirical, interdisciplinary research combining social and computer sciences to develop and evaluate effective AI-based countermeasures.

Functionalities and Effectiveness of Detection Models (RQ5–6)

Detection studies primarily use data from Twitter (X), Facebook, and news websites. Datasets often come from repositories (e.g., kaggle.com, Github) or are scraped using tools like CrowdTangle. Access to data and training models remains a challenge, primarily because of changes in the social media landscape and the decentralized power of social media corporations, which increasingly hinder access to free and unbiased data (Pasquetto et al., 2020). While textual data are most common, some studies incorporate images and videos (e.g., Ghai et al., 2024). Popular datasets include LIAR (fake vs. real news), Cresci (bot labels), and Botometer for bot detection. Overall, the real-world application of detection models remains underexplored, making it difficult to assess their effectiveness.

Most detection models utilize learning-based approaches (n = 57), followed by hybrid feature-learning methods (n = 12). Solely feature-based models (n = 4) and knowledge-based models (n = 1) are rare in the sample. Learning-based models, which rely on deep learning or unsupervised machine learning (ML) algorithms, offer high performance (e.g., precision, F1-scores) and can detect previously unseen disinformation. However, they require extensive training time, large and balanced datasets, and often have reduced accuracy when predicting accurate information (Babu, Lung, & Zaman, 2023). Feature-based approaches, commonly used in bot detection (e.g., Bessi & Ferrara, 2016), analyze factors like sentiment, user data, metadata, and networks. While they support replication and integrate user knowledge during optimization, they face scalability issues and low generalizability. Knowledge-based approaches, incorporating human input (e.g., expert annotators), improve transparency and semantic understanding but are time-intensive and prone to bias.

Detection models predominantly use deep learning classifiers, such as Bidirectional Encoder Representations from Transformers (BERT) and Long Short-Term Memory (LSTM), which outperform ML classifiers like Support Vector Machines (SVM) or Gradient Boost. Average performance[2] metrics highlight LSTM (accuracy = 0.8999, F1 = 0.8384) and BERT (accuracy = 0.8948, F1 = 0.8754) as top performers.

[2] Average performance was calculated according to the metrics provided by the articles. Only articles that stated accuracy and F1 for their models were included in the calculation. Hence, the average performance is the average of all metrics for different models provided by the articles included in the sample. The numbers are to be interpreted cautiously and can only be understood as broad indicators of model effectiveness.

However, variations like IndoBERT for Indonesian datasets demonstrate significant performance drops, indicating a lack of language and context diversity in automated models. Deep learning approaches, while effective, sacrifice explainability for performance (Gongane et al., 2024). Choosing the right approach depends on context and requires balancing performance with interpretability. Innovative methods such as combining knowledge graphs, convolutional neural networks (CNNs), and deep learning under explainable AI (XAI) show promise (Kundu & Nguyen, 2024), though interpretability challenges persist.

Few studies test models on external datasets or assess their effectiveness in real-world applications. For instance, browser extensions (e.g., Afonso & Rosas, 2024) and tools like the FactFinder web app (Pathak et al., 2021) aim to verify content but lack extensive evaluation. Correction measures, including debunking, flagging, fact-checking, and content moderation, remain theoretical or underdeveloped. The findings underscore the urgent need for interdisciplinary efforts and real-world testing to improve detection models' effectiveness and generalizability.

Discussion

This review highlights how AI operates in the disinformation ecosystem not only as a technological driver but also as a field-defining concept that reconfigures methodological, theoretical, and normative debates in political communication, with an overview provided in Table 4. Despite a growing body of literature, our analysis reveals persistent theoretical fragmentation. Many studies, particularly in computer science, operate without reference to established theories of media effects, credibility, or political behavior. Conversely, communication studies often treat AI as a contextual factor rather than a system with its logic and epistemology. Bridging these traditions requires not just methodological integration, but also theoretical innovation that acknowledges AI as both actor and structure in the disinformation landscape.

A key theoretical gap lies in the limited engagement with the agency of AI systems. While some communication scholars apply heuristic models—such as the realism heuristic (e.g., Hameleers, 2024) and source credibility (e.g., Lee & Shin, 2022)—these approaches are not tailored to the unique characteristics of AI-generated content. Future research should explore how classic models of media effects (e.g., cultivation theory, spiral of silence) adapt—or fail—in environments shaped by generative AI and algorithmic dissemination. Similarly, Explainable AI (XAI) offers a promising bridge for developing user-centered theories of trust, interpretability, and resistance to manipulation (Gongane et al., 2024).

Furthermore, most studies frame disinformation as a content-level problem, overlooking relational and networked dimensions, such as how users interact with detection tools, how trust is co-constructed, or how countermeasures (e.g., good bots [Zhang et al., 2023] and inoculation messaging [e.g., Zerback et al., 2021]) evolve in feedback loops with platform policies. These perspectives require a shift from single-cause analysis toward systems thinking, including cross-platform dynamics, multimodal environments, and cross-linguistic contexts.

Methodologically, the dominance of textual data and English-language sources narrows the scope of generalizability. Studies addressing underrepresented languages and regions—particularly in the Global South—are urgently needed to assess whether existing detection models are applicable beyond well-

resourced contexts (e.g., Sanjana et al., 2023). The same applies to underexamined modalities like audio and video deepfakes, which present unique perceptual and verification challenges.

Practically, our findings suggest that many proposed countermeasures remain either untested or lack real-world validation. Future work should explore interdisciplinary pilot studies—for instance, embedding detection tools in real platform environments to observe user behavior and unintended consequences (e.g., Afonso & Rosas, 2024). Theoretical work must go hand-in-hand with practical experimentation, especially when informing policy debates and public education initiatives.

In sum, while AI opens new avenues for combating disinformation, it also demands a rethinking of how we conceptualize agency, credibility, and intervention. By synthesizing insights across disciplines, this review contributes to the foundations of a more coherent, comparative, and critically reflective research agenda on AI and disinformation.

Table 4. Overview of Research Gaps and Future Research Avenues.

Descriptive Research Gaps
- Research on underrepresented regions (e.g., Africa, South America) - A wider application of detection models, especially on low-resource languages, and multimodal applications - Establishment of new datasets that include non-English, esp. non-U.S.-centric contexts, and different platforms (e.g., YouTube, Instagram, Reddit, Telegram, Twitch)
Methodological Research Gaps
- Empirical assessment of the impact of AI-driven disinformation and countermeasures - New mixed-method approaches combining traditional methods of communication and computer science (e.g., computational–qualitative approaches) to explore the effects of AI-driven disinformation and the research multimodality of AI-driven disinformation - Comparative designs to explore country-specific differences - Real-world testing of detection models across different settings and contexts (e.g., browser extensions, content moderation on social media)
Theoretical and Conceptual Research Gaps
- Need for advanced theoretical frameworks and development of classic models of media effects to assess the impact of AI-driven disinformation (e.g., cultivation theory, spiral of silence) - Exploration of emerging theories and frameworks at the intersection of communication and computer science (e.g., XAI) - This includes conceptual papers and discussions on AI regulation. - Expand research on disseminators of disinformation beyond conceptual considerations, also discussing the role of automated actors and actor agency

Limitations

This review provides important insights but has several limitations. The focus on English-language, peer-reviewed studies may have excluded relevant work in other languages or non-academic formats, limiting regional and practical perspectives. The emphasis on textual data reflects trends in the field,

potentially underrepresenting emerging formats like deepfakes. Additionally, reliance on existing databases and systematic coding, while robust, may introduce selection and interpretation biases.

References

Abd, D. H., Mahdi, M. F., Jassim, M. A., & Hussain, A. (2023). Arabic fake news detection using ensemble technique. *16th International Conference on Developments in eSystems Engineering (DeSE)*, 292–297. doi:10.1109/DeSE60595.2023.10469046

Abdirahman, A. I. (2023). Exploring co-regulation as a solution to automated disinformation in Kenya. *Journal of Intellectual Property and Information Technology Law (JIPIT), 3*(1), 201–256. doi:10.52907/jipit.v3i1.262

Afonso, R., & Rosas, J. (2024). Development of a smartphone application and chrome extension to detect fake news in English and European Portuguese. *IEEE Latin America Transactions, 22*(4), 294–303. Retrieved from https://latamt.ieeer9.org/index.php/transactions/article/view/8547

Ahmed, S. (2023). Examining public perception and cognitive biases in the presumed influence of deepfakes threat: Empirical evidence of third person perception from three studies. *Asian Journal of Communication, 33*(3), 308–331. doi:10.1080/01292986.2023.2194886

Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: A review. *Social Network Analysis and Mining, 13*(1), 1–36. doi:10.1007/s13278-023-01028-5

Al Ghamdi, M. A., Bhatti, M. S., Saeed, A., Gillani, Z., & Almotiri, S. H. (2023). A fusion of BERT, machine learning and manual approach for fake news detection. *Multimedia Tools and Applications, 83*(10), 30095–30112. doi:10.1007/s11042-023-16669-z

Babu, R. N., Lung, C.-H., & Zaman, M. (2023). Performance evaluation of transformer-based NLP models on fake news detection datasets. *IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, 316–321. doi:10.1109/COMPSAC57700.2023.00049

Benevenuto, F., & Melo, P. (2024). Misinformation campaigns through WhatsApp and Telegram in presidential elections in Brazil. *Communications of the ACM, 67*(8), 72–77. doi:10.1145/3653325

Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday, 21*(11). doi:10.5210/fm.v21i11.7090

Bontridder, N., & Poullet, Y. (2021). The role of artificial intelligence in disinformation. *Data & Policy*, *3*, e32. doi:10.1017/dap.2021.20

Boulianne, S., & Humprecht, E. (2023). Perceived exposure to misinformation and trust in institutions in four countries before and during a pandemic. *International Journal of Communication, 17*, 2024–2047.

Dao, M.-S., & Zettsu, K. (2023). Leveraging knowledge graphs for CheapFakes detection: Beyond dataset evaluation. *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 99–104. doi:10.1109/ICMEW59549.2023.00024

Daume, S., Galaz, V., & Bjersér, P. (2023). Automated framing of climate change? The role of social bots in the Twitter climate change discourse during the 2019/2020 Australia bushfires. *Social Media + Society, 9*(2), 1–16. doi:10.1177/20563051231168370

de Vreese, C., & Tromble, R. (2023). The data abyss: How lack of data access leaves research and society in the dark. *Political Communication, 40*(3), 356–360. doi:10.1080/10584609.2023.2207488

Dikmen, M., & Burns, C. (2022). The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies*, *162*, 102792. doi:10.1016/j.ijhcs.2022.102792

Dourado, T. (2023). Who posts fake news? Authentic and inauthentic spreaders of fabricated news on Facebook and Twitter. *Journalism Practice, 17*(10), 2103–2122. doi:10.1080/17512786.2023.2176352

Ekpang, J. E., Iyorza, S., & Ekpang, P. O. (2023). Social media and artificial intelligence: Perspectives on deepfakes' use in Nigeria's 2023 general elections. *Kampala International University Interdisciplinary Journal of Humanities and Social Sciences, 4*(2), 109–124. doi:10.59568/kijhus-2023-4-2-08

Ferrara, E. (2017). Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday, 22*(8). doi:10.5210/fm.v22i8.8005

Gagrčin, E., Naab, T. K., & Grub, M. F. (2024). Algorithmic media use and algorithm literacy: An integrative literature review. *New Media & Society*, 1–25. Advance online publication. doi:10.1177/14614448241291137

Ghai, A., Kumar, P., & Gupta, S. (2024). A deep-learning-based image forgery detection framework for controlling the spread of misinformation. *Information Technology & People, 37*(2), 966–997. doi:10.1108/ITP-10-2020-0699

Gil De Zúñiga, H., Goyanes, M., & Durotoye, T. (2024). A scholarly definition of artificial intelligence (AI): Advancing AI as a conceptual framework in communication research. *Political Communication, 41*(2), 317–334. doi:10.1080/10584609.2023.2290497

Gongane, V. U., Munot, M. V., & Anuse, A. D. (2024). A survey of explainable AI techniques for detection of fake news and hate speech on social media platforms. *Journal of Computational Social Science, 7*(8), 587–623. doi:10.1007/s42001-024-00248-9

Guba, B. (2008). Systematische Literatursuche [Systematic literature search]. *Wiener Medizinische Wochenschrift, 158*(1–2), 62–69. doi:10.1007/s10354-007-0500-0

Guzmán Rincón, A., Carrillo Barbosa, R. L., Segovia-García, N., & Africano Franco, D. R. (2022). Disinformation in social networks and bots: Simulated scenarios of its spread from system dynamics. *Systems, 10*(2), 1–11. doi:10.3390/systems10020034

Hajli, N., Saeed, U., Tajvidi, M., & Shirazi, F. (2022). Social bots and the spread of disinformation in social media: The challenges of artificial intelligence. *British Journal of Management, 33*(3), 1238–1253. doi:10.1111/1467-8551.12554

Hameleers, M. (2024). Cheap versus deep manipulation: The effects of cheapfakes versus deepfakes in a political setting. *International Journal of Public Opinion Research, 36*(1), 1–9. doi:10.1093/ijpor/edae004

Hameleers, M., Van Der Meer, T. G. L. A., & Dobber, T. (2022). You won't believe what they just said! The effects of political deepfakes embedded as vox populi on social media. *Social Media + Society, 8*(3), 1–12. doi:10.1177/20563051221116346

Humprecht, E., Esser, F., & Van Aelst, P. (2020). Resilience to online disinformation: A framework for cross-national comparative research. *The International Journal of Press/Politics, 25*(3), 493–516. doi:10.1177/1940161219900126

Jost, P., & Dogruel, L. (2023). Radical mobilization in times of crisis: Use and effects of appeals and populist communication features in telegram channels. *Social Media + Society, 9*(3), 1–12. doi:10.1177/20563051231186372

Jungherr, A. (2025). Political disinformation: "Fake news," bots, and deep fakes. In M. Powers (Ed.), *Oxford research encyclopedia of communication*. Oxford, UK: Oxford University Press.

Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications, 80*(8), 11765–11788. doi:10.1007/s11042-020-10183-2

Kertysova, K. (2018). Artificial intelligence and disinformation: How AI changes the way disinformation is produced, disseminated, and can be countered. *Security and Human Rights, 29*(1–4), 55–81. doi:10.1163/18750230-02901005

Kessler, S. H., Mahl, D., Schäfer, M. S., & Volk, S. C. (2025). All eyez on AI: A roadmap for science communication research in the age of artificial intelligence. *Journal of Science Communication, 24*(2), Y01. doi:10.22323/2.24020401

Kirby, P., & Thorpe, N. (2024, December 6). *Romania's cancelled presidential election and why it matters*. BBC. Retrieved from https://www.bbc.com/news/articles/cx2yl2zxrq1o

Krämer, B. (2021). Stop studying "fake news" (we can still fight against disinformation in the media). *Studies in Communication and Media, 10*(1), 6–30. doi:10.5771/2192-4007-2021-1

Kundu, A., & Nguyen, U. T. (2024). Automated fact checking using A knowledge graph-based model. *2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 709–716. doi:10.1109/ICAIIC60209.2024.10463196

Kuznetsova, E., & Makhortykh, M. (2023). Blame it on the algorithm? Russian government-sponsored media and algorithmic curation of political information on Facebook. *International Journal of Communication*, *17*, 971–992. Retrieved from https://ijoc.org/index.php/ijoc/article/view/18687/4035

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. doi:10.1057/s41304-024-00482-9

Landon-Murray, M., Mujkic, E., & Nussbaum, B. (2019). Disinformation in contemporary U.S. foreign policy: Impacts and ethics in an era of fake news, social media, and artificial intelligence. *Public Integrity, 21*(5), 512–522. doi:10.1080/10999922.2019.1613832

Lange, B., & Lechterman, T. M. (2021). Combating disinformation with AI: Epistemic and ethical challenges. *IEEE International Symposium on Technology and Society (ISTAS)*, 1–5. doi:10.1109/ISTAS52410.2021.9629122

Lee, J., & Shin, S. Y. (2022). Something that they never said: Multimodal disinformation and source vividness in understanding the power of AI-enabled deepfake news. *Media Psychology, 25*(4), 531–546. doi:10.1080/15213269.2021.2007489

Li, M., & Wan, Y. (2023). Norms or fun? The influence of ethical concerns and perceived enjoyment on the regulation of deepfake information. *Internet Research, 33*(5), 1750–1773. doi:10.1108/INTR-07-2022-0561

Lin, T. (2022, June). *Investigating the relationship of disguised socialbots and disinformation threat in Taiwan*. Presented at the 31st European Conference of the International Telecommunications Society (ITS), Gothenburg, Sweden. Retrieved from https://hdl.handle.net/10419/265654

Ma, Y.-W., Chen, J.-L., Chen, L.-D., & Huang, Y.-M. (2024). Intelligent clickbait news detection system based on artificial intelligence and feature engineering. *IEEE Transactions on Engineering Management*, 1–10. doi:10.1109/TEM.2022.3215709

Makhortykh, M., Sydorova, M., Baghumyan, A., Vziatysheva, V., & Kuznetsova, E. (2024). Stochastic lies: How LLM-powered chatbots deal with Russian disinformation about the war in Ukraine. *Harvard Kennedy School (HKS) Misinformation Review*. doi:10.37016/mr-2020-154

McNair, B. (2017). *An introduction to political communication* (6th ed.). London, UK: Routledge.

Nah, F.-H., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research, 25*(3), 277–304. doi:10.1080/15228053.2023.2233814

O'Donnell, N. (2021). Have we no decency? Section 230 and the liability of social media companies for deepfake videos. *University of Illinois Law Review, 2*, 701–740. Retrieved from https://www.illinoislawreview.org/wp-content/uploads/2021/03/ODonnell.pdf

Paschen, J., Kietzmann, J., & Kietzmann, T. C. (2019). Artificial intelligence (AI) and its implications for market knowledge in B2B marketing. *Journal of Business & Industrial Marketing, 34*(7), 1410–1419. doi:10.1108/JBIM-10-2018-0295

Pasquetto, I., Swire-Thompson, B., Amazeen, M. A., Benevenuto, F., Brashier, N. M., Bond, R. M., . . .Yang, K.-C. (2020). Tackling misinformation: What researchers could do with social media data. *Harvard Kennedy School (HKS) Misinformation Review*. doi:10.37016/mr-2020-49

Patel, Y., Tanwar, S., Gupta, R., Bhattacharya, P., Davidson, I. E., Nyameko, R., . . . Vimal, V. (2023). Deepfake generation and detection: Case study and challenges. *IEEE Access, 11*, 143296–143323. doi:10.1109/ACCESS.2023.3342107

Pathak, A., Srihari, R. K., & Natu, N. (2021). Disinformation: Analysis and identification. *Computational and Mathematical Organization Theory, 27*(3), 357–375. doi:10.1007/s10588-021-09336-x

Pedrazzi, S., & Oehmer, F. (2020). Communication rights for social bots?: Options for the governance of automated computer-generated online identities. *Journal of Information Policy, 10*, 549–581. doi:10.5325/jinfopoli.10.2020.0549

Rasyid, H. R. P., Sibaroni, Y., & Ihsan, A. F. (2023). Classification of disinformation tweets on the 2024 Presidential election in Indonesia using optimal transformer based model. *International Conference on Data Science and Its Applications (ICoDSA)*, 191–196. doi:10.1109/ICoDSA58501.2023.10277101

Reisach, U. (2021). The responsibility of social media in times of societal and political manipulation. *European Journal of Operational Research, 291*(3), 906–917. doi:10.1016/j.ejor.2020.09.020

Repede, Ș. E. (2023). Researching disinformation using artificial intelligence techniques: Challenges. *Bulletin of "Carol I" National Defence University, 12*(2), 69–85. doi:10.53477/2284-9378-23-21

Robertson, R., & Meintjes, C. (2021). Towards an online risk mitigation framework for political brands subject to computational propaganda. *Communicatio, 47*(1), 95–121. doi:10.1080/02500167.2021.1884578

Sanjana, S., Kuranagatti, S., Devisetti, J. G., Sharma, R., & Arya, A. (2023). Intersection of machine learning, deep learning and transformers to combat fake news in Kannada language. *6th International Conference on Contemporary Computing and Informatics (IC3I)*, 2264–2270. doi:10.1109/IC3I59117.2023.10398034

Santana, C., Nunes, A., & Silva, F. (2021). The role of bots in the disinformation process in Brazilian politics between 2014 and 2018. *Libri, 71*(4), 321–333. doi:10.1515/libri-2020-0071

Santos, F. C. C. (2023). Artificial intelligence in automated detection of disinformation: A thematic analysis. *Journalism and Media, 4*(2), 679–687. doi:10.3390/journalmedia4020043

Schäfer, M. S. (2023). The notorious GPT: Science communication in the age of artificial intelligence. *Journal of Science Communication, 22*(2), Y02. doi:10.22323/2.22020402

Starbird, K., Arif, A., & Wilson, T. (2019). Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), 1–26. doi:10.1145/3359229

Strömbäck, J., Tsfati, Y., Boomgaarden, H., Damstra, A., Lindgren, E., Vliegenthart, R., & Lindholm, T. (2020). News media trust and its impact on media use: Toward a framework for future research. *Annals of the International Communication Association, 44*(2), 139–156. doi:10.1080/23808985.2020.1755338

Sukhramani, K., Kumre, H., Rasool, A., & Jadav, A. (2024). Binary classification of news articles using deep learning. *2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, 1–9. Retrieved from https://www.proceedings.com/content/074/074117webtoc.pdf

Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger & A. J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 73–100). Cambridge, MA: The MIT Press.

Tomar, M., Nihal, R., Singh, S., Marwaha, S. S., & Tiwani, M. (2023). The role of AI-driven tools in shaping the democratic process: A study of Indian elections and social media dynamics. *Industrial Engineering Journal, 52*(11), 143–153. Retrieved from http://www.journal-iiie-india.com/1_nov_23/20.3_nov.pdf

Twomey, J., 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), 1–22. doi:10.1371/journal.pone.0291668

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society, 6*(1), 1–13. doi:10.1177/2056305120903408

Valente, A., Holanda, M., Mariano, A. M., Furuta, R., & Da Silva, D. (2022). Analysis of academic databases for literature review in the computer science education field. *IEEE Frontiers in Education Conference (FIE)*, 1–7. doi:10.1109/FIE56618.2022.9962393

Vizoso, Á., Vaz-Álvarez, M., & López-García, X. (2021). Fighting deepfakes: Media and Internet giants' converging and diverging strategies against hi-tech misinformation. *Media and Communication, 9*(1), 291–300. doi:10.17645/mac.v9i1.3494

Weikmann, T., Greber, H., & Nikolaou, A. (2024). After deception: How falling for a deepfake affects the way we see, hear, and experience media. *The International Journal of Press/Politics, 30*(1), 187–210. doi:10.1177/19401612241233539

Weikmann, T., & Lecheler, S. (2023). Cutting through the hype: Understanding the implications of deepfakes for the fact-checking actor-network. *Digital Journalism, 12*(10), 1505–1522. doi:10.1080/21670811.2023.2194665

Whyte, C. (2020). Deepfake news: AI-enabled disinformation as a multi-level public policy challenge. *Journal of Cyber Policy, 5*(2), 199–217. doi:10.1080/23738871.2020.1797135

Zerback, T., Töpfl, F., & Knöpfle, M. (2021). The disconcerting potential of online disinformation: Persuasive effects of astroturfing comments and three strategies for inoculation against them. *New Media & Society, 23*(5), 1080–1098. doi:10.1177/1461444820908530

Zhang, Y., Song, W., Koura, Y. H., & Su, Y. (2023). Social bots and information propagation in social networks: Simulating cooperative and competitive interaction dynamics. *Systems, 11*(4), 1–19. doi:10.3390/systems11040210