AI-Generated Misinformation: A Literature Review

[1]Rafharum Fatimah, [2]Auziah Mumtaz, [3]Fauzan Muhammad Fahrezi, [4]Diky Zakaria
[1,2,3,4]Mechatronics and Artificial Intelligence Study Program, Universitas Pendidikan Indonesia, Indonesia
Email: [1]rafharumf@upi.edu, [2]auziahmumtaz@upi.edu, [3]rezifauzan303@upi.edu, [4]dikyzak@upi.edu

Article Info

Article history:
Received Dec 12th, 2023
Revised Mar 26th, 2024
Accepted Apr 20th, 2024

Keyword:
AI-generated
Artificial Intelligence
Detection Tools
Literature Review
Misinformation

ABSTRACT

The expansion of artificial intelligence (AI) technologies has signaled an entirely new era in which the creation and sharing of information, both correct and misleading, are becoming increasingly automated. This research of the literature explores the landscape of AI-generated misinformation, including its various manifestations, underlying technology, societal impact, and detection tools. This paper reviews articles from the Google Scholar database related to AI-Generated Misinformation focusing on the following research questions: the types, content distribution, detector variations, differences among the various tools, and strategies for developing AI-based tools. The result is to provide an absolute comprehension of this topic, underlining the importance of interdisciplinary collaboration, robust detection methods, and media literacy with the intention to solve the ethical and societal issues it poses in the age of digital technology.

Corresponding Author:
Diky Zakaria,
Mechatronics and Artificial Intelligence Study Program, Universitas Pendidikan Indonesia,
8 Veteran Street, Purwakarta, West Java 41115, Indonesia.
Email: dikyzak@upi.edu

DOI: http://dx.doi.org/10.24014/ijaidm.v7i2.26455

1. INTRODUCTION

Artificial intelligence (AI), the most exciting innovation that has revolutionized the way we live. It refers to the creation of computer systems that allow us to program machines to perform human-like tasks [1]. This term serves to deal with issues relating to people. AI systems include a variety of techniques and methods that allow computers to understand, learn, and make decisions based on the data provided [2]. The improved learning abilities have enabled AI to make advances into substantially more advanced decision-making contexts, such as those involving audio, speech, and object identification, or natural language processing [3].

The expansion of artificial intelligence algorithms has provided users with recommendations for a confusing variety of tasks. It is now possible to create written works, pictures. videos, and many other forms of content that are often indistinguishable from those created by humans [4]. However, some downsides include a lack of emotional and contextual awareness, as well as the risk of misinformation caused by job automation [5]. Having the aim of generating text, photos, and videos to create false material, the capabilities of AI are constrained in this case. Moreover, the rapid growth of social media during the last decade has enabled others to quickly share captured multimedia content, triggering an enormous expansion in multimedia content creation and ease of access [6]. It is becoming increasingly difficult to recognize the truth and trust information, which may have undesirable consequences. A countermeasure is needed to deal with the issues that are generated by the spread of AI-Generated misinformation.

Until now, three review articles have discussed specific aspects of AI-generated misinformation, but none have supplied a comprehensive overview of the topic. Salah et al [7] explain the potential threat of AI-generated misinformation, but it primarily focuses on the use of generative AI tools like ChatGPT in social psychology research and the challenges and opportunities that come with this technology. Adadi et al [8] explain a systematic umbrella review and foresight analysis on the role of AI in handling the challenges posed by the COVID-19 pandemic. Bahroun et al [9] give a comprehensive review of the ethical implications, responsible use, data privacy safeguards, biases, and academic integrity when exploring the assimilation of GAI (Generative Artificial Intelligence) within educational settings. Despite these contributions, these articles exclusively delve into specific aspects of AI-generated misinformation and do not provide a comprehensive overview of the entire topic. A review article specifically addressing AI-generated misinformation in a broad sense has not been written to date. Following this set of circumstances, the author aims to write a review article on AI-Generated Misinformation. To achieve this, the author put together the following research questions (RQs):

1. RQ1: What types of false or misleading information are commonly generated by AI systems?
2. RQ2: How does AI-generated misinformation content spread and influence online communities?
3. RQ3: What tools are available for detecting and analyzing AI-generated misinformation, and what are their variations?
4. RQ4: What are the key differences and strengths among the various tools available for detecting and countering AI-generated misinformation?
5. RQ5: What are the most promising strategies for developing AI-based tools that can find and combat the proliferation of AI-generated misinformation?

The method used is systematic literature review (SLR), and the articles are through Google Scholar. Multiple articles have employed the systematic literature review (SLR) approach in their research. Vargas-Murillo et al [10] conducted an SLR to acquire specific insights into the utilization of ChatGPT in the educational field. They analyzed 16 articles retrieved from various databases, including Corpus, Science Direct, ProQuest, IEEE Xplore, and ACM Digital Library. The search was performed using these keywords ("ChatGPT") AND ("education" OR "learning" OR "e-learning" OR "teaching") AND ("AI" OR "artificial intelligence"). An SLR is chosen as an approach because it enables investigators to compile existing studies and articles on the subject, creating a curated collection of literature that employs diverse approaches to assess the same topic, ensuring a comprehensive, unbiased, and up-to-date research foundation. Guerrero et al [11] conducted a SLR to explore the current research literature on detecting synthetic text. The authors selected 50 out of 1211 articles (4.12%) from Google Scholar, using it as their primary search engine. The SLR enabled the authors to supply a comprehensive overview of the current state of research in this field, show gaps in the literature, and prepare for future research.

As a result of that, SLR will be used in this article for getting information relating to the previously described RQs and furthermore to supply a comprehensive overview of the issue of AI-generated misinformation. It will examine the techniques used to generate AI-generated misinformation, the challenges in detecting it, the potential impact of AI-generated misinformation on society, and the tools used for evaluating and critically assessing AI-generated content. The review will highlight the need for further research in this area to develop effective strategies to combat AI-generated misinformation.

2. RESEARCH METHOD

The current article is a literature review, and the method employed is outlined in full listed below:

2.1. Article Selection

Relevant articles were chosen using the Google Scholar database. The keywords used in the search were based on (“AI-Generated Misinformation” OR “AI-Generated Fake News”). The chosen articles had to have been published throughout the past five years (2019-2023) and be unrestricted access. Materials such as books, online blog postings, essays, theses, dissertations, reviewed and irrelevant articles were not considered throughout the selection process. The search was conducted on September 22, 2023, and the following information was provided (Table 1):

Table 1. Article Selection Process on Google Scholar Database

No.	Process	Number of Article
1	In the Google Scholar search area, enter the terms “AI-Generated Misinformation” OR “AI-Generated Fake News” based on title, abstract, and keywords.	81 articles
2	Exclude books, online blog posts or essays, theses and dissertations, and review articles	62 articles
3	Exclude irrelevant article	22 articles
4	Article is closed access	15 articles
	The total number of articles used	15 articles

According to Table 1, there are 81 Scholar documents based on the keyword. We only used 15 items (18%) of the entire 81 after applying the exclusion criteria.

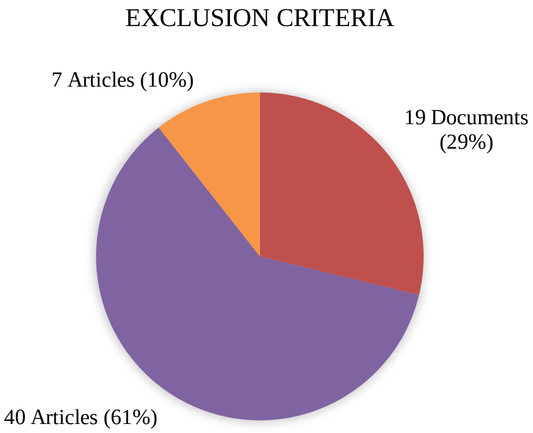

Figure 1. Exclusion Criteria on Google Scholar Database

2.2. Article Review Process

Once the article selection is done, the author thoroughly reviews the chosen articles, focusing on answering the 5 predetermined Research Questions (RQs) with meticulous diligence. The author downloads these articles from various sources to ensure accessibility and carefully examines each article, discussing any limitations or gaps in the studies. The discussion phase entails synthesizing commonalities, differences, and trends among the articles to supply well-founded conclusions based on the collective findings and addressing the RQs.

3. RESULTS AND ANALYSIS

3.1. Article Metadata

The metadata of the 15 articles analyzed are displayed in Table 2.

Table 2. Metadata of the Articles Analyzed

No.	Author (s)	Year	Paper Title	Conference/ Journal Source
1	Bhattacharjee et al [12]	2023	ConDA: Contrastive Domain Adaptation for AI-generated Text Detection	arXiv preprint arXiv:2309.03992.
2	Dalkir et al [13]	2021	Fake News and AI: Fighting Fire with Fire?	AIofAI'21: 1st Workshop on Adverse Impacts and Collateral Effects of Artificial Intelligence Technologies
3	Fung et al [14]	2021	InfoSurgeon:Cross-Media Fine-grained Information Consistency Checking for Fake News Detection	Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing
4	Heumann et al [15]	2023	ChatGPT and GPTZero in Research and Social Media: A Sentiment-and Topic-based Analysis	Twenty-ninth Americas Conference on Information Systems, Panama, 2023
5	Iceland [16]	2023	How Good Are SOTA Fake News Detectors?	arXiv preprint arXiv:2308.02727.
6	Koplin [17]	2023	Dual-use implications of AI text generation	Ethics and Information Technology (2023) 25:32
7	Kreps et al [18]	2022	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.
8	Kulkarni et al [19]	2023	Exploring Semantic Perturbations on Grover	arXiv preprint arXiv:2302.00509
9	Kumarage et al [20]	2023	Stylometric Detection of AI-Generated Text in Twitter Timelines	arXiv preprint arXiv:2303.03697
10	Baecker et al [21]	2019	Threats provided by artificial intelligence that could disrupt the democratic system.	Scientific Paper Faculty of Economics University of Applied Science Brandenburg, 15(2), 9–25.
11	Shay et al [22]	2020	The Ethics of Generative AI in Tax Practice	Tax Notes Federal, Volume 180, Number 5
12	Shu et al [23]	2021	A Pilot Study Investigating STEM Learners' Ability to Decipher AI-generated Video	2021 ASEE Virtual Annual Conference

No.	Author (s)	Year	Paper Title	Conference/ Journal Source
13	Wang et al [24]	2023	From Human-Centered to Social- Centered Artificial Intelligence: Assessing ChatGPT's Impact through Disruptive Events	arXiv preprint arXiv:2306.00227
14	Yang et al [25]	2023	Anatomy of an AI-powered malicious social botnet	arXiv preprint arXiv:2307.16336.
15	Zhou et al [26]	2023	Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions	Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1-20).

Following the data presented in Table 2, the analysis of the sources supplied shows that most references are drawn from conference proceedings and arXiv preprints, constituting a large 12 out of the 15 sources (80%). In contrast, only two references are from journal articles, specifically "Ethics and Information Technology (2023) 25:32" and "Journal of experimental political science, 9(1), 104-117" (13.33%). Furthermore, one source is a scientific paper from the Faculty of Economics University of Applied Science Brandenburg (6.67%). This distribution underscores the prevalence of conference-related sources in the field of artificial intelligence and technology ethics, showing potential opportunities for more extensive coverage of these topics in peer-reviewed journals. Researchers and scholars may consider this trend when selecting publication venues for their work, aiming to contribute to a more balanced representation in academic literature.

3.2. Article Review Summary

The results of the summary finding from the completed article reviews have been outlined in Table 3.

Table 3. Article Review Results

No.	References	RQ1	RQ2	RQ3	RQ4	RQ5
1	Bhattacharjee et al [12]	Generating human-like text for specific tasks such as summarization, translation, paraphrasing, creative writing, explanation of ideas and concepts, code generation and correction, and solving mathematical proofs etc	Spread through social media platforms and potentially influence online communities	ConDa	Effectively solves the problem of label scarcity, and achieves ultramodern performance for unsupervised detection	Developing ConDa that solves human-like text and comparing with human based text
2	Dalkir et al [13]	AI systems can generate a wide range of false or misleading information, including text-based fake news, multimedia (images and video) content, deep fakes, propaganda, and phishing campaigns	AI-generated misinformation content spreads on social media because anyone can share it easily. This is made worse by echo chambers, where people only listen to those who think like them. The fake news from AI fits right in, as it's designed to match what people already believe. This means it spreads fast and can change what people think and do online, making it harder for trusted sources to be heard and affecting real-world decisions	Tools including software that detects fake news using linguistic analysis and linguistic comparison of the language and using sentiment analysis. Others developed machine-learning tools to analyze and show potential fake news	These tools encompass linguistic analysis, which identifies deceptive language patterns; text similarity analysis, which detects inconsistencies between headlines and articles; sentiment analysis, focusing on emotional language in fake news; bot detection through the analysis of multiple features; fact-checking and evidence-based corrections to enhance content reliability; user reaction analysis by examining	Developing tools that combine human ability and artificial intelligence ability because of the plethora of different contexts in which false information flows online makes it tricky for AI to work on its own, absent human knowledge. Therefore, human ability is needed to complement AI-based solutions. To conclude, it suggests that the more promising approaches in using AI to detect AI do not rely on AI alone

No.	References	RQ1	RQ2	RQ3	RQ4	RQ5
					emotional responses; dissemination pattern analysis to identify spread patterns; and visual content analysis to distinguish between fake and real news based on multimedia elements presence of images and videos	
3	Fung et al [14]	Generative neural network models have been used to generate "realistic-looking AI-generated 'fake news'" that can easily deceive humans. These types of misinformation can be generated by manipulating, misusing, exaggerating, or falsifying only a small part of the true information, namely the knowledge element		Cross-media consistency checking, which is based on a benchmark dataset that includes both human-written and machine-generated fake news	-	A general approach to ensure proper, rather than malicious, application of dual-use technology should incorporate ethical considerations as the first-order principles in every step of the system design, as well as support a high degree of transparency and interpretability of data, algorithms, models, and functionality throughout the system. In addition, it is also important to create an interpretable approach so that users of the system can understand which parts of the article have been falsified. To conclude, the authors intend to make their misinformation detector software available as open source and share docker containers for public verification and auditing so it can be used to combat fake news
4	Heumann et al [15]	Potential misuse of plagiarism and false information.	From anyone who use the ChatGPT	GPTZero	GPTZero is a special tool to detect any content from ChatGPT, GPT-4 and more	If you make and AI language model, also make the countermeasure
5	Iceland [16]	AI-Generated Fake News	Social media platforms enable fake news to be generated and spread very easily	Automatic fake news detection with machine learning can prevent the dissemination of false statements before they gain wide attention	Machine learning model were used for fake news detection, and CT-BERT in particular stands out as one of the best performers and unlike the traditional models, the deep models can achieve near-perfect performance	To classify fake news reliably, an ensemble of traditional machine learning classifiers may achieve robust performance. Traditional models would also be easier to ensemble compared to large language models because of their fast

No.	References	RQ1	RQ2	RQ3	RQ4	RQ5
						performance and low training cost. Alternatively, continual learning could be used as a strong fake news detection framework in which models can incrementally get knowledge over extended periods of time and adapt to changing data distributions
6	Koplin [17]	AI text generators could be used to generate "synthetic" fake news to support specific viewpoints, discredit particular political regimes, or praise or slander particular products, people, or companies. This could include false political news, as well as online disinformation	Supported by a study referenced in the same text, which found that online disinformation including false political news, appears to spread further and faster than the truth because people are more likely to retweet false claims than true ones	There are multiple teams are developing AI tools to distinguish between human- and machine-generated text, based on certain peculiarities of machine-generated language	-	Developing technological solutions to the problem of machine-generated fake news that could be implemented at the bottom of the dual-use pipeline. Multiple teams are developing AI tools to distinguish between human- and machine-generated text, based on certain peculiarities of machine-generated language
7	Kreps et al [18]	The practice of online hyper-targeting, which involves delivering specific news stories or advertisements to particular demographic groups in an attempt to create polarization and influence political leanings, has prompted responses, leading platforms like Twitter to limit political advertising	The content are made by GPT-2 and the information are spread on twitter and other online platform	-	-	-
8	Kulkarni et al [19]	the current deep learning methods in natural language processing can automatically generate texts, and they can be applied to fake news generation	AI-generated misinformation content can spread through social media platforms and potentially influence online communities	Grover, which is an incredibly robust model that has proven to be great at both generating and detecting neural fake news	-	-
9	Kumarage et al [20]	Exploiting AI model to generate human like text or article at large scale to make misinformation	A potential threat scenario is that an adversary takes over the legitimate user account and incorporates a language generator to create fake news	Bot Detection on Twitter	Such as follower counts, also likes, and retweets	AI-Generated Text Detection. Early investigations into finding generated text included methods like bag-of-word and tf–idf encoding followed by standard classifiers such as

No.	References	RQ1	RQ2	RQ3	RQ4	RQ5
						logistic regression, random forest, and SVC
10	Baecker et al [21]	Misapplication of . surveillance and control of humans, manipulation through fake news, filter bubbles, algorithm bias, fake scientific article, manipulate citizen's opinions about political candidates, manipulate voter's opinions	Anonymous people can easily spread fake news that refers to misleading, distorted, manipulated and highly inaccurate content ; used social media accounts under false identities and disparaged candidate's campaigns	Everyone should increase their vigilance on any content they consume from the internet, they should always take information with their brain and soul. Then they also need to evaluate the status quo of the content they read, try to always look for contradictory arguments, so that get correct information as possible	-	Media literacy and research are important skills. The need for increased awareness and regulation of AI technology to ensure its responsible use and to protect democratic values, the purpose of AI is not to replace the role of humans, but to support their work. AI algorithms should be as transparent as possible, so that humans can also get the most objective results that everyone expects
11	Alarie et al [22]	AI generates untrue information that is not backed up by real-world data	Particularly popular recently in the media	-	-	Robust systems must be implemented to check and validate LLM outputs to ensure their accuracy. Recognizing the nascent nature of this technology, legal professionals bear the responsibility of thoroughly reviewing and confirming the analysis generated by LLMs before relying on it or providing it to clients. This approach not only safeguards against potential inaccuracies but also allows lawyers to harness their creative critical thinking skills when navigating complex legal issue
12	Shu et al [23]	Misleading information or data with AI generated video	COVID-19 pandemic increase the STEM learner that rely on online course that vulnerable to misinformation that can be made by AI	-	-	The pattern of the expression from AI generated video can still be seen by human, the pattern can be used to train AI model to detect AI generated video
13	Wang et al [24]	The tendency to generate inaccurate outputs unfaithful to the training data, cheating and plagiarism in higher education, and false allegations	Prioritizing groups and institutions compels us to confront how people, with their distinctive know-how and resources, use technologies in the wild, group of students or anyone who use ChatGPT	Turnitin, GPTZero	The plagiarism detection program Turnitin to detect AI writing calling such work "misconduct," and GPTZero is a classification model that predicts whether a document was written by ChatGPT by	Use a plagiarism detection service like Turnitin to check students' submissions against a database of previously given work and other digital sources. Turnitin's long-term usefulness may hinge

No.	References	RQ1	RQ2	RQ3	RQ4	RQ5
			spread it in online media		comparing the variation and complexity of sentence, AI team was working hard to keep detection services at pace with generative models	on its ability to incorporate new techniques to evolve along with dialogue agents (for example, to shift away from similarity checks to examine the "origin of content," When accessing ChatGPT's web interface, users are reminded that the system "may produce inaccurate information
14	Yang et al [25]	Fake content on twitter botnet that employ ChatGPT to generate human-like content, distort online conversations and spread misinformation in various contexts, from elections to public health crises	Promote suspicious websites and spread harmful comments in Twitter	GPT-2, Botometer, OpenAI's AI text detector, and GPTZero	GPT-2, Generate text and compare it with human-generated content. Botometer considers over 1,000 features covering account profiles, content, social networks, and more. It has been confirmed in various research projects under various contexts. OpenAI's Detector and GPTZero, Human- and machine-generated content that works for different LLMs including OpenAI's own model	Black-box and white-box approaches. Black-box detection methods are often framed as binary classification problems where classifiers are trained on texts generated by humans and machines. White-box owners embed specific signals or watermarks into generated content for later identification. Raise public awareness about the existence of LLM-powered bots and educate individuals on strategies for self-protection against such threats
15	Zhou et al [26]	Large language AI model can generate human like text to be use to generate persuasive misinformation	Spreading fake news about cures to conspiracy theories about COVID-19	CT-BERT	The model pretrained on COVID-19-related Twitter documents collected from January to April 2020	Compared to human based misinformation, AI based misinformation lacks emotion and cognitive processing. So, we can make a model based on that pattern

3.3. Answering RQs

To develop AI-based tools to detect and combat AI-generated misinformation, various strategies have been shown. These strategies include comparing AI text with human text, blending human and AI expertise, ensuring ethics and transparency, building countermeasures, using traditional classifiers, identifying machine-generated text, using AI text detection, promoting media literacy and rules, legal review, recognizing AI-generated video, using plagiarism detection, applying black-box and white-box approaches, and analyzing emotion and cognitive patterns. These techniques are described briefly below:

1. Comparing AI Text with Human Text: Create AI that can tell the difference between content made by AI and humans [12].
2. Blend Human and AI Expertise: Use both human and AI knowledge to combat online false info, as AI alone can struggle due to the variety of misleading content [13].
3. Ethics and Transparency: Ensure ethical design, transparency, and user understanding of misinformation detection tools. Make software open source for public review [14].
4. Build Countermeasures: If AI generates fake content, make AI tools to detect and counteract it [15].
5. Traditional Classifiers: Use traditional machine learning methods to classify fake news. Continual learning can also help adapt over time [16].

6. Identify Machine-Generated Text: Develop tools to spot machine-generated text by recognizing unique language patterns [17].
7. AI Text Detection: Early methods used bag-of-words and classifiers like logistic regression to detect AI-generated text [20].
8. Media Literacy and Rules: Raise awareness and set rules for using AI responsibly. AI should aid humans, not replace them. Be clear and open with how AI works [21].
9. Legal Review: Legal experts should carefully review AI-generated reports before using them. This ensures accuracy and lets lawyers apply their skills [22].
10. Recognize AI-Generated Video: Train AI models to spot AI-generated videos by identifying patterns [23].
11. Plagiarism Detection: Services like Turnitin can be used to check for AI-generated content, but may need to evolve to deal with AI-generated material [24].
12. Black-Box and White-Box Approaches: Detect AI content using either black-box (binary classification) or white-box (embedding signals) methods [25].
13. Emotion and Cognitive Patterns: AI-generated misinformation lacks human emotion and thought patterns, which can be used to build models for detection [26].

3.3.1 Answering RQ1

Based on the article review result to answer RQ1 about "What types of false or misleading information are commonly generated by AI systems?". The most common type of AI-generated misinformation identified was text-based, accounting for 40% of the reviewed articles [12][13][17][18][19][20]. This includes human-like text [12], fake news [13], synthetic news supporting viewpoints [17], hyper-targeted online content [18], deep learning for fake news generation [19], and large-scale fake article generation [20]. Multimedia misinformation, through manipulated images/videos and misleading AI-generated video, was discussed in 13% of the articles [14][23]. 20% of the articles focused on unethical or biased misuse of AI to generate misinformation via plagiarism, control, and inaccurate outputs [15][21][24]. Persuasive misinformation and social media misinformation campaigns accounted for 13% each of the reviewed literature, covering fake news [16], persuasive misinformation [26], untrue AI-generated information [22], and AI chatbot misinformation campaigns [25].

In summary, the literature analysis revealed text-based misinformation as the most prevalent application of AI for generating false or misleading information. However, AI also enables a wide range of multimedia, persuasive, unethical, and social media-based misinformation. Further research is needed to comprehensively address the rapidly evolving landscape of AI's capabilities for information manipulation.

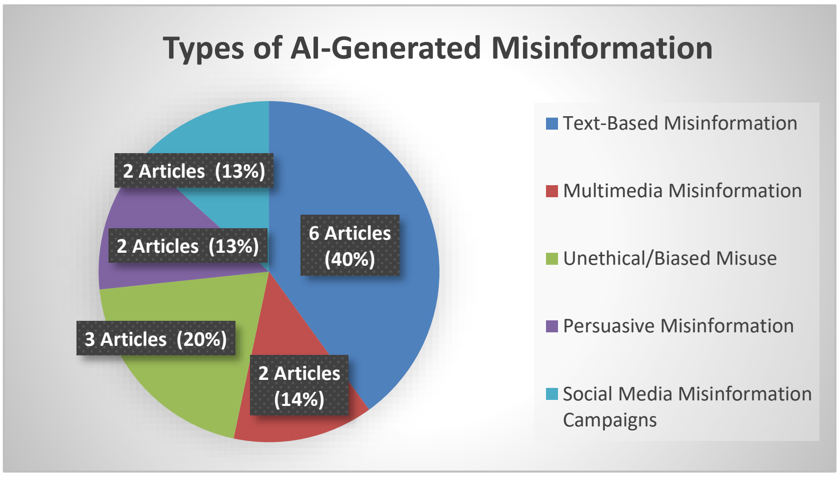

Figure 2. Types of false or misleading information that commonly generated by AI

3.3.2 Answering RQ2

This literature review provided a comprehensive analysis of the spread methods and impact resulting from artificial intelligence (AI)-generated misinformation in online platforms. The majority high-weighted research support the general consensus that social media encourages the widespread distribution of misleading artificial intelligence (AI) content [12], [13], [15], [16], [18], [19], [21]-[24]. The content has the ability to proliferate through social media platforms and affect online communities. AI-generated fake news and conspiracy theories may rapidly propagate across online communities, alter perceptions, and influence public opinion by exploiting the immense reach and viral appeal of popular social media platforms [19].

Furthermore, direct messaging applications allow the unrestricted dissemination of falsified statements among users, which can quickly disseminate false information across communities of peers [17], [20], [24], [26]. Suspicious websites also use the absence of content verification to spread AI-fabricated misinformation to innocent online communities [25]. However, there is minimal evidence of effective countermeasures.

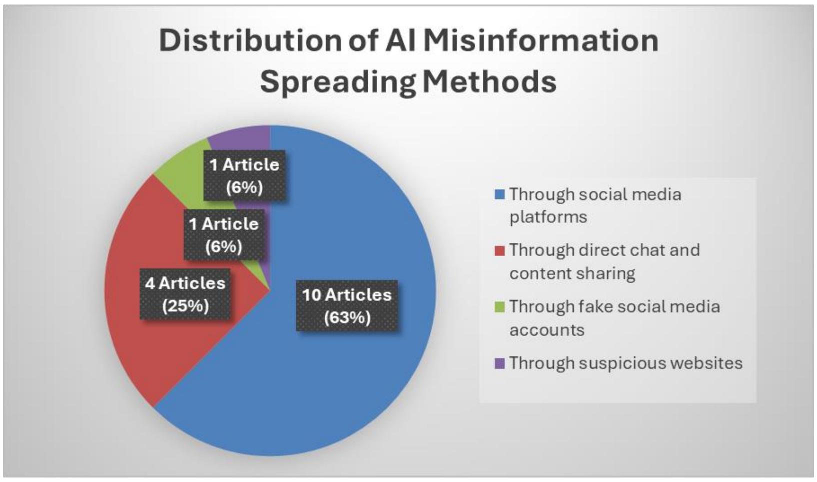

Figure 3. Distribution of AI Misinformation Spreading Methods

Overall, it seems clear that online social interactions are the key means for AI-generated lies to spread quickly and influence people [12]. Although the damage may be mitigated by stricter content management and increased user awareness, there are presently insufficient methods for identifying deceptive AI content [13]. Further research is desperately needed to develop comprehensive methods for preventing the dissemination and impacts of AI misinformation in online spaces.

3.3.3 Answering RQ3

There are a lot of tools developed to detect AI generated text. the most widely utilized AI text detection tools encompass Turnitin, GPTZero, and OpenAI’s AI text detector. Turnitin is a plagiarism detection tool commonly used to detect plagiarism including those involving AI-generated text[24]. While GPTZero and OpenAI’s AI text detector work similarly to determine how much of a document is generated by AI by doing multilevel classification which analyze the text on sentence, paragraph and document levels[25]. While it is impossible to reliably detect all AI written text. But with those tools we can aware for false claims that AI-generated text was written by a human: for example, running automated misinformation campaigns, using AI tools for academic dishonesty, and positioning an AI chatbot as a human.

Additionally, for Twitter, specialized bot detectors serve to discern whether an account is spreading false information, thus contributing to the identification of fake accounts [20]. The tools for detecting and analyzing AI-generated misinformation are diverse and continually evolving, including linguistic and sentiment analysis, machine learning techniques for fake news detection, and cross-media consistency checking with benchmark datasets that feature both human-written and machine-generated fake news.

Everyone should also increase their vigilance on any content they consume from the internet, they should always take information with their brain and soul. Then they also need to test the status quo of the content they read, try to always look for contradictory arguments, so that get accurate information as much as possible[21].

In response to RQ3, the most widely utilized AI text detection tools encompass ConDa, Turnitin, GPTZero, and CT-BERT. These tools primarily consist of AI models trained to identify AI-generated text. Additionally, for Twitter, specialized bot detectors serve to discern whether an account is spreading false information, thus contributing to the identification of fake accounts [20]. The tools for detecting and analyzing AI-generated misinformation are diverse and continually evolving, including linguistic and sentiment analysis, machine learning techniques for fake news detection, and cross-media consistency checking with benchmark datasets that feature both human-written and machine-generated fake news. Notable models and tools, such as GPTZero, Grover, Bot Detection on Twitter, Turnitin, GPT-2, Botometer, and OpenAI's AI text detector, are developed to distinguish between human and machine-generated text by leveraging language peculiarities.

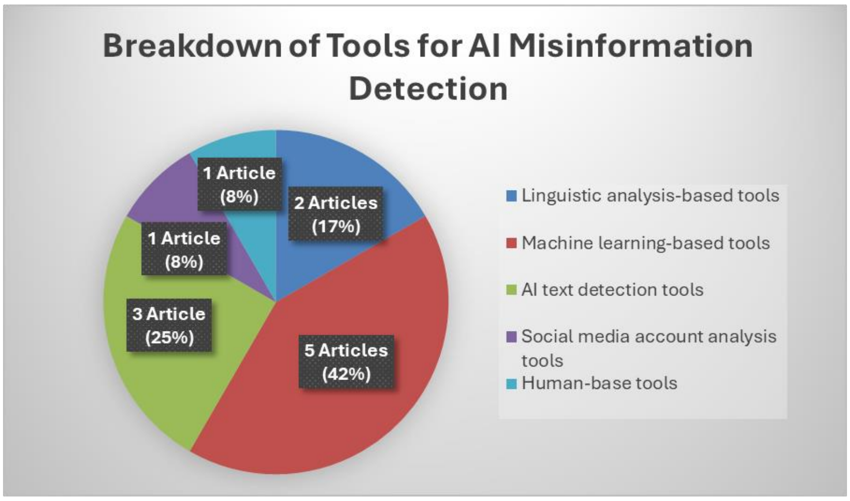

Figure 4. Breakdown of Tools for AI Misinformation Detection

1. Linguistic analysis-based tools
   References: [12], [13]
   One of Linguistic analysis-based tools is ConDa. ConDa is a fake news detection software using linguistic analysis method. These tools identify misinformation based on linguistic patterns and inconsistencies.

2. Machine learning-based tools
   References: [13], [14], [15], [16], [19]
   Automatic fake news detection models made using machine learning algorithms. These tools can classify texts as misinformation by training on labelled datasets.

3. AI text detection tools
   References: [17], [24], [25]
   Using AI Models to distinguish machine-generated vs human-written text. These tools identify unique linguistic signatures of AI systems.

4. Social media account analysis tools
   Reference: [20]
   Botometer for detecting Twitter bot accounts. Useful for identifying automated misinformation spreaders.

5. Human-based tools
   Reference: [21]
   Increasing human vigilance and verification. Humans may still outperform AI detectors in some cases.

3.3.4 Answering RQ4

Regarding RQ4, The tools utilize different approaches to identify AI-generated misinformation, with respective strengths. ConDa [12] effectively addresses label scarcity and achieves state-of-the-art unsupervised detection performance. The tools in [13] encompass multi-faceted analysis like linguistic patterns, text similarity, sentiment analysis, bot detection, fact checking, user reactions, spread patterns, and multimedia content analysis. GPTZero [15] specifically targets content from ChatGPT and other LLMs. CT-BERT [16] excels as a deep learning fake news detector, outperforming traditional models. For social media, tools examine user statistics like followers and engagement [20]. Turnitin [24] focuses on plagiarism detection by comparing texts, while GPTZero analyzes linguistic complexity. Botometer [25] considers over 1000 account and content features to identify AI bots. The model in [26] is specialized for COVID-19 Twitter misinformation. In summary, while strengths vary, combining different tools can allow comprehensive detection across texts, accounts, and media types [13]. Tailored tools like GPTZero [15] also help counter specific AI models' misinformation output. An ensemble approach leverages these tools' complementary strengths for robust AI misinformation detection.

3.3.5 Answering RQ5

For RQ5, to effectively combat AI-generated misinformation, the best strategies involve creating tools to detect AI-generated text and incorporating them into social media to limit its spread [17]. Here are the most exciting possibilities for building solutions based on artificial intelligence to counteract misinformation.

1. Combining AI with human expertise
   Human contextual expertise is required to enhance the ability of AI to recognize misleading data. Domain-specific human expertise is critical for detecting misleading claims. Human expertise and experience are required to interpret nuanced language and complicated the preferences, which AI currently struggles with. Misinformation detection systems may be made more precise and trustworthy by integrating artificial intelligence (AI) capabilities with human expertise [13].

2. Integrating ethical principles and transparency into AI systems
   Applying ethical standards and transparency in AI systems is crucial to ensuring that they work effectively and ethically. Clear ethical principles are necessary to avoid damaging or destructive AI solutions, as is transparency, so that AI system processes and logic can be evaluated and reported for. With ethics and transparency, public trust in AI systems could have increased [14].

3. Developing flexible incremental machine learning models
   Machine learning (ML) models that are incrementally and adaptively trained are useful. Traditional machine learning models can perform better when integrated than when they are used separately. Also, incremental machine learning models are simpler to integrate and modify for new data. AI systems are more adaptable and able to modify outputs in response to changing data and information when they use incremental models [16].

4. AI text detection using linguistic peculiarities
   Different from human writing, machine-generated language has observable patterns and peculiarities. Compared to human writing, artificial intelligence text is typically less diversified, repetitious, and deficient the contextualization. Machine-authored material may be reliably detected by algorithmic training to detect linguistic abnormalities in AI writing; linguistic-based detection is essential to distinguishing authentic from misleading information [17].

5. Using human verification when reviewing AI systems
   Human verification of AI systems is required to assure output correctness and solve algorithmic errors [22]. For the purpose of evaluating and improving AI performance, human validation is still necessary. In ways that robots find difficult, humans are able to recognize the flaws, prejudices, and restrictions of AI systems. AI systems may be improved and made more reliable by frequent human checks [24].
6. Comparing AI outputs to human text samples.
   Identification of machine-generated material may be achieved effectively by comparing AI text outputs with instances of human writing. It is possible to identify pattern discrepancies that point to fake writing by using systems like ConDa that can produce language that is human-like and compare it to actual text. For training systems to detect AI text anomalies, benchmarking against authentic human text is essential [12].

7. Applying machine learning-based classification algorithms
   The use of machine learning-based classification algorithms has shown to be successful in the identification of incorrect data. Training datasets and black-box and white-box approaches may be used to train algorithms to recognize authentic and false text and video. With additional training data, machine learning improves the capacity of a system to distinguish between material created by AI and authentic information. Detection speed and scalability are increased by automatic machine learning classification [25].

8. Training identification of falsely created video/imagery.
   Training detection of machine-generated visual content is critical since synthetic media contains observable artifacts and patterns. Deepfakes and synthetic pictures have unique characteristics such as artifacts, noise, and inconsistencies that distinguish them from actual video. Fake sights may be consistently spotted by training pattern recognition to detect false signs. Detecting synthetic content contributes to the prevention of disinformation dissemination [23].

9. Taking use of emotional and cognitive patterns
   Distinguishing between writing written by AI and human beings can be achieved by utilizing variations in emotional and cognitive patterns. The emotional and cognitive complexity of language produced by current AI is still inferior to that of human writers. AI writing often has little cognitive depth, is literal, and lacks emotional complexity [26].

4. CONCLUSION

This article was made to learn the expansion of artificial intelligence (AI) as a tool to generate misinformation that looks like a human made text. The method used for this article is SLR(Systematic literature review) where relevant articles were chosen from the Google Scholar database. The keywords used in the search were based on ("AI-Generated Misinformation" OR "AI-Generated Fake News"). The chosen article were published throughout the past five years (2019-2023) and are open access. Materials such as books, online blog postings, essays, theses, dissertations, reviewed and irrelevant articles were not considered throughout the selection process. Also, closed access article from google scholar are not discussed in this article. The search was carried out on September 22, 2023, and obtained 16 articles. After the review process is done, it can be concluded that AI is a double edge sword where it can be useful to make things easier to do but can also be misused to generate human-like misinformation text. This misinformation are spread throughout social media where information spread very quick. In order to fight AI generated misinformation, there are already several tools that can detect AI generated text such as ConDa, GPTZero, Turnitin, and CT-BERT. The best strategy to fight AI generated misinformation is to make a tool or program that can distinguish AI made text and human made text that integrated to social media.Also, there are two things we can do to prevent AI generated misinformation from spreading further. The first one is to develop a tool that can detect AI generated text. And the second one is to teach humans about the importance of digital literature and crosscheck every piece of information that comes to them. With the ability to differentiate AI generated text, we can prevent misinformation from spreading further on social media.

REFERENCES

[1] S. Bhbosale, V. Pujari, and Z. Multani, "Advantages and Disadvantages of Artificial Intelligence," in *Aayushi International Interdisciplinary Research Journal*, 2020, pp. 227–230. [Online]. Available: https://www.researchgate.net/profile/Vinayak-Pujari-2/publication/344584269_Advantages_And_Disadvantages_Of_Artificial_Intellegence/links/5f81b70192851c14bcbc1d96/Advantages-And-Disadvantages-Of-Artificial-Intellegence.pdf%0Awww.aiirjournal.com

[2] Y. Duan, J. S. Edwards, and Y. K. Dwivedi, "Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda," *Int. J. Inf. Manage.*, vol. 48, pp. 63–71, 2019, doi: 10.1016/j.ijinfomgt.2019.01.021.

[3] N. Berente, B. Gun, J. Recker, and R. Santhanam, "MANAGING ARTIFICIAL INTELLIGENCE," vol. 45, no. September 2021, pp. 1–18, 2021, doi: 10.25300/MISQ/2021/16274.

[4] Y. Yang, Y. Zhuang, and Y. Pan, "Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies," *Front. Inf. Technol. Electron. Eng.*, vol. 22, no. 12, pp. 1551–1558, 2021, doi: 10.1631/FITEE.2100463.

[5] J. Paschen, "Investigating the emotional appeal of fake news using artificial intelligence and human contributions," *J. Prod. Brand Manag.*, vol. 29, no. 2, pp. 223–233, 2020, doi: 10.1108/JPBM-12-2018-2179.

[6] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward," *Appl. Intell.*, vol. 53, no. 4, pp. 3974–4026, 2023, doi: 10.1007/s10489-022-03766-z.

[7] M. Salah, H. Al Halbusi, and F. Abdelfattah, "May the force of text data analysis be with you: Unleashing the power of generative AI for social psychology research," *Comput. Hum. Behav. Artif. Humans*, vol. 1, no. 2, p. 100006, 2023, doi: 10.1016/j.chbah.2023.100006.

[8] Z. Bahroun, C. Anane, V. Ahmed, and A. Zacca, "Transforming Education: A Comprehensive Review of Generative Artificial Intelligence in Educational Settings through Bibliometric and Content Analysis," *Sustain.*, vol. 15, no. 17, p. 12983, 2023, doi: 10.3390/su151712983.

[9] A. Adadi, M. Lahmer, and S. Nasiri, "Artificial Intelligence and COVID-19: A Systematic umbrella review and roads ahead," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 8, pp. 5898–5920, 2022, doi: 10.1016/j.jksuci.2021.07.010.

[10] A. R. Vargas-Murillo, I. N. M. de la Asuncion Pari-Bedoya, and F. de Jesús Guevara-Soto, "Challenges and Opportunities of AI-Assisted Learning: A Systematic Literature Review on the Impact of ChatGPT Usage in Higher Education," *Int. J. Learn. Teach. Educ. Res.*, vol. 22, no. 7, pp. 122–135, 2023, doi: 10.26803/ijlter.22.7.7.

[11] J. Guerrero and I. Alsmadi, "Synthetic Text Detection: Systemic Literature Review," pp. 1–9, 2022, [Online]. Available: http://arxiv.org/abs/2210.06336

[12] A. Bhattacharjee, T. Kumarage, R. Moraffah, and H. Liu, "ConDA: Contrastive Domain Adaptation for AI-generated Text Detection," pp. 1–13, 2023, [Online]. Available: http://arxiv.org/abs/2309.03992

[13] K. Dalkir, "Fake news and AI: Fighting fire with fire?," *CEUR Workshop Proc.*, vol. 2942, pp. 112–115, 2021.

[14] Y. R. Fung *et al.*, "InfoSurgeon: Cross-media fine-grained information consistency checking for fake news

detection," in *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2021, pp. 1683–1698. doi: 10.18653/v1/2021.acl-long.133.

[15] M. Heumann and T. Kraschewski, "ChatGPT and GPTZero in Research and Social Media: A Sentiment-and Topic-based Analysis Optimal end-of-life strategies for aging wind turbines View project," 2023.

[16] M. Iceland, "How Good Are SOTA Fake News Detectors," pp. 1–13, 2023, [Online]. Available: http://arxiv.org/abs/2308.02727

[17] J. J. Koplin, "Dual-use implications of AI text generation," *Ethics Inf. Technol.*, vol. 25, no. 2, pp. 1–11, 2023, doi: 10.1007/s10676-023-09703-z.

[18] S. Kreps, R. M. McCain, and M. Brundage, "All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation," *J. Exp. Polit. Sci.*, vol. 9, no. 1, pp. 104–117, 2022, doi: 10.1017/XPS.2020.37.

[19] P. Kulkarni, Z. Ji, Y. Xu, M. Neskovic, and K. Nolan, "Exploring Semantic Perturbations on Grover," pp. 1–15, 2023, [Online]. Available: http://arxiv.org/abs/2302.00509

[20] T. Kumarage, J. Garland, A. Bhattacharjee, K. Trapeznikov, S. Ruston, and H. Liu, "Stylometric Detection of AI-Generated Text in Twitter Timelines," pp. 1–13, 2023, [Online]. Available: http://arxiv.org/abs/2303.03697

[21] C. Baecker, G. P. Yogiputra, T. D. Nguyen, and O. Alabbadi, "Threats provided by artificial intelligence that could disrupt the democratic system," 2023.

[22] B. Alerie and R. McCreight, "The Ethics of Generative AI in Tax Practice," 2023.

[23] D. Shu *et al.*, "A Pilot Study Investigating STEM Learners' Ability to Decipher AI-generated Video," in *ASEE Annual Conference and Exposition, Conference Proceedings*, 2021, pp. 1–20. doi: 10.18260/1-2--36601.

[24] S. Wang, N. Cooper, M. Eby, and E. S. Jo, "From Human-Centered to Social-Centered Artificial Intelligence: Assessing ChatGPT's Impact through Disruptive Events," pp. 1–23, 2023, [Online]. Available: http://arxiv.org/abs/2306.00227

[25] K.-C. Yang and F. Menczer, "Anatomy of an AI-powered malicious social botnet," pp. 1–27, 2023, [Online]. Available: http://arxiv.org/abs/2307.16336

[26] J. Zhou, Y. Zhang, Q. Luo, A. G. Parker, and M. De Choudhury, "Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions," *Conf. Hum. Factors Comput. Syst. - Proc.*, pp. 1–20, 2023, doi: 10.1145/3544548.3581318.

BIBLIOGRAPHY OF AUTHORS

Rafharum Fatimah, an undergraduate student majoring in mechatronics and artificial intelligence, is dedicated and driven, with a passion for technical excellence. She aims to craft innovative solutions that align with the digital world, contributing to a smarter future through integrating technology and intelligence.

Auziah Mumtaz is an enthusiastic undergraduate student majoring in Mechatronics and Artificial Intelligence at Indonesia University of Education West Java, Indonesia. She is enthusiastic about exploring the fusion of mechanics, electronics, and computer science to create intelligent systems. Auziah's path is all about learning, innovating, and shaping the future of technology.

Fauzan Muhammad Fahrezi is an undergraduate student who majored in mechatronics and artificial intelligence. He is extremely interested in artificial intelligence. He wants to create AI to make everyday life much easier.

Diky Zakaria, a lecturer at the Indonesia University of Education's Purwakarta Campus, is dedicated to sharing his passion for electrical engineering and mechatronics. With undergraduate and graduate degrees in electrical engineering from UPI and ITB, he brings deep technical knowledge to his teaching.