Journal of Cyberspace Studies
eISSN: 2588-5502
http://jcss.ut.ac.ir
University of Tehran

The Evolution of Disinformation from Fake News Propaganda to AI-driven Narratives as Deepfake

Shahla Nasiri[1], Armin Hashemzadeh[2]*

1. Department of Mathematics, Statistics and Computer Science, University of Tabriz, Tabriz, Iran.
2. Department of American Studies, Faculty of World Studies, University of Tehran, Tehran, Iran. (*armin.hashemzadeh49@gmail.com https://orcid.org/0000-0002-8530-0000-0002-8530)

Article Info

Original Article

Main Object: Media

Received: 09 October 2024
Revised: 19 December 2024
Accepted: 29 December 2024
Published online: 01 January 2025

Keywords:
artificial intelligence (AI),
fake news,
election interference,
deepfakes,
disinformation,
misinformation,
social media.

Abstract

Background: Disinformation has undergone significant transformations over the past few decades, evolving from relatively simple text-based fake news articles to highly sophisticate AI-driven content such as deepfakes and other forms of manipulated media.

Aims: This paper traces the historical development of Disinformation, its increasing reliance on Artificial Intelligence (AI), and the potential future trajectories of disinformation as AI technologies advance.

Methodology: We begin by examining the shift from traditional text-based disinformation campaigns, often propagated via social media platforms, to more immersive and persuasive forms of AI-generated media.

Discussion: We discuss how AI techniques, such as Generative Adversarial Networks and Natural Language Processing, have revolutionized the landscape of false information, allowing for the automation of misinformation production and its widespread dissemination at an unprecedented scale. Furthermore, this paper investigates the role of social media algorithms in amplifying disinformation, demonstrating how these platforms, originally designed to prioritize user engagement, inadvertently aid in the spread of false information by promoting sensationalized or emotionally charged content. Through an in-depth analysis of case studies, including the COVID-19 pandemic and the 2020 U.S. elections, this paper highlights the dangers posed by AI-generated misinformation, particularly deepfakes, which are becoming increasingly difficult to detect, even by advanced AI systems. The implications of this shift for democratic processes, public trust, and societal cohesion are profound. This paper also explores the ethical dilemmas posed by AI-driven disinformation and presents potential solutions through the lens of AI-enhanced detection technologies and policy interventions. Lastly, this paper emphasizes the urgent need for interdisciplinary cooperation between policymakers, technologists, and media organizations to mitigate the harmful impacts of AI-driven disinformation while preserving the integrity of information in the digital age.

Conclusions: By exploring both technological and regulatory approaches, a comprehensive framework for understanding the evolving threat of AI-driven disinformation is essential and pathways for future research in this critical area is suggested.

Cite this article: Nasiri Sh, Hashemzadeh A. (2025). "The Evolution of Disinformation from Fake News Propaganda to AI-driven Narratives as Deepfake". *Cyberspace Studies*. 9(1): 229-250. doi: https://doi.org/10.22059/jcss.2025.387249.1119.

Website: https://jcountst.ut.ac.ir/ | Email: jcountst@ut.ac.ir |
EISSN: 2980-9193
Publisher: University of Tehran

Cyberspace Studies, Vol 9, No 1, January 2025

1. Introduction

Digitalization affects nearly every aspect of our lives—impacting individuals, organizations, and society as a whole. Although much of the current literature in management and organization studies tends to focus on the positive outcomes of this development, the darker, more unexpected sides of digitalization have received far less attention (Trittin-Ulbrich et al., 2021).

Social media and the internet have made it very simple to obtain information, but they have also made it easier to spread lies. For instance, consider fake news. This contemporary phenomenon gained significant attention during the 2016 U.S. presidential election, when fake news not only became more prevalent but also outperformed real news in terms of social media engagement (Allcott & Gentzkow, 2017). Fake news has been around for a while, but the digital age has given it new means of propagation. Fake news was initially mainly disseminated as text articles that imitated reputable sources in an effort to increase clicks and earn money from advertising by using sensationalized or completely made-up content. However, the quick dissemination of content to large audiences on social media sites like Facebook and X (formerly Twitter) made fake news a potent weapon for disinformation campaigns (Vosoughi et al., 2018).

Social media algorithms—programmed to boost content that garners high engagement—often end up amplifying false information. Studies indicate that false news stories are about 70% more likely to be shared than true ones (ibid). In today's fast-paced digital world, disinformation can reach millions in just a few hours, making real-time correction incredibly challenging.

Adding another layer to the problem, artificial intelligence now allows for the tailoring of disinformation to individual biases. By analyzing huge amounts of user data, AI can adapt false narratives to align with specific beliefs and preferences (Zhou & Zafarani, 2020). This personalization means that people are less likely to question information that confirms what they already think. In effect, AI-driven micro-targeting only makes disinformation campaigns more potent, as these custom messages exploit individual vulnerabilities (Ferrara et al., 2020).

In the early stages, AI was used simply to automate the spread of false information through basic rule-based algorithms and bots, which would flood social media with fabricated content—often text-based articles or disleading images (ibid). Fast forward to today, and tools like OpenAI's GPT-3 can generate text that is not only contextually accurate but also stylistically similar to reputable news sources (Brown et al., 2020). This makes it increasingly difficult for readers to tell real news from fake, as AI now produces content that mirrors legitimate media in tone, style, and format.

The true danger isn't just the speed or scale of AI's ability to spread

false information—it's also its capacity to tailor content for specific audiences. Machine learning algorithms now analyze user behavior and preferences to create disinformation that reinforces existing biases, thereby manipulating public opinion (Ferrara et al., 2020). With advances in machine learning, natural language processing (NLP), and generative adversarial networks (GANs), AI-powered disinformation has become both more efficient and more convincing (Goodfellow et al., 2014; Pomputius, 2019).

For example, GANs operate by pitting two neural networks against one another: one network (the generator) creates fake content while the other (the discriminator) works to detect whether that content is genuine. Over time, the generator improves at "fooling" the discriminator, producing outputs—whether text, images, or videos—that are extremely convincing (Goodfellow et al., 2014). This capability is especially worrisome when it comes to disinformation, as it enables the creation of content that appears authentic, even when it is entirely fabricated.

Deepfakes represent another alarming development. Using GANs, deepfakes can produce hyper-realistic videos in which people appear to say or do things they never actually did (Chesney & Citron, 2019). Already, deepfakes have been used in political contexts to sway public opinion, and their potential to undermine trust in video evidence is enormous. As AI continues to evolve, the challenge of discerning genuine content from fabricated material will only intensify, posing serious issues for journalists, fact-checkers, and policymakers (Landon-Murray et al., 2019; Kietzmann et al., 2020).

There is also the concern that AI models might mistakenly flag legitimate content as disinformation, which raises important questions about censorship. Striking a balance between effectively combating false information and preserving free speech—and, by extension, legitimate journalism—will be one of our most pressing challenges moving forward (Chesney & Citron, 2019).

This paper traces how disinformation has evolved in the digital age—from simple, text-based fake news to today's sophisticated, AI-driven content, including deepfakes. It illustrates how social media has lowered the barriers for creating and disseminating false information, while its algorithms further amplify high-engagement content. By clarifying the differences between disinformation, disinformation, and malinformation, and by examining how AI technologies (such as NLP and GANs) have enhanced our ability to produce realistic yet misleading content, we can better understand the mounting challenge of preserving trust in media, democratic processes, and societal cohesion in an era increasingly defined by AI-driven disinformation.

Cyberspace Studies, Vol 9, No 1, January 2025

2. Theoretical framework

The rapid pace at which artificial intelligence is evolving has fundamentally transformed the way information is shared. This shift has given rise to new forms of disinformation that take advantage of human vulnerabilities, societal structures, and technological platforms. In this section, I bring together several theoretical perspectives to offer a well-rounded understanding of AI-driven disinformation, its impact on society, and why a multidisciplinary approach is crucial for addressing it. Drawing on the information disorder framework (Wardle & Derakhshan, 2017), principles of information quality (Omoregie, 2021a), and the harm principle (Mill & Mill, 1966; Omoregie, 2021b), this study lays the groundwork for analyzing and tackling the challenges posed by disinformation in the digital era.

Disinformation—the intentional spread of false or misleading information to deceive and manipulate—is not a new phenomenon. Its history can be traced back to the propaganda of World War II and the Cold War's Dezinformatsiya tactics, evolving over time into the large-scale digital disinformation campaigns we witnessed during the 2016 U.S. presidential election. During that period, about 126 million Americans encountered false narratives amplified by foreign actors (ibid). Today's disinformation stands out because of its speed, reach, and virality, attributes that have been significantly enhanced by social media and advanced AI technologies.

Modern AI systems, especially those powered by GANs and machine learning, are now capable of generating highly convincing false content. Whether it's deepfakes, synthetic voices, or automated narratives, these tools exploit the natural limitations of human perception, making disinformation increasingly difficult to detect and counter. This progression highlights the urgent need to understand the underlying mechanisms of AI-driven disinformation and to develop comprehensive frameworks to effectively address these emerging threats.

2.1. Information disorder

The term "information disorder" was coined by Claire Wardle and Hossein Derakhshan (2017) to describe the different types of false information that spread in the digital age. Their framework categorizes harmful content into three distinct types: misinformation, disinformation, and malinformation. Understanding these distinctions is crucial for developing a theoretical foundation for how AI interacts with and amplifies various forms of false information.

- Misinformation refers to the inadvertent sharing of false or misleading information without malicious intent. This kind of information is often shared by individuals who believe the content to be true but lack the means or knowledge to verify it. For example, during the COVID-19 pandemic, many individuals

shared false health remedies, not out of an intent to deceive, but because they trusted the sources or found the content compelling (Agarwal et al., 2023).

- Disinformation is the deliberate creation and dissemination of false information with the intent to deceive or manipulate audiences for political, financial, or ideological gain. In this case, actors intentionally craft fabricated narratives, often using AI to maximize reach and engagement. AI-generated disinformation, such as deepfakes or AI-created text, has grown in sophistication and is increasingly difficult for average users to detect (Chesney & Citron, 2019).
- Malinformation involves the sharing of genuine information in ways intended to cause harm. Unlike misinformation or disinformation, malinformation relies on accurate information shared out of context to harm individuals, institutions, or societies. For example, private conversations or sensitive political documents leaked in a harmful context fall into this category.

These forms of information disorder become even more potent when amplified by AI technologies. AI can automate the creation and dissemination of both misinformation and disinformation, often enhancing their reach and effectiveness. The danger lies not only in the scale and speed at which AI can spread false information but also in the personalization and contextualization of content to target specific audiences. Machine learning algorithms analyze user behaviors, preferences, and biases, tailoring disinformation campaigns to reinforce existing beliefs and manipulate vulnerable individuals (Ferrara et al., 2020).

AI-driven disinformation, such as deepfakes, adds a new layer of complexity to information disorder. Deepfakes blend disinformation and malinformation by using AI to fabricate visual and auditory content that falsely portrays real individuals in compromising or misleading situations. These digital fabrications erode trust in visual media and have profound implications for democratic processes, journalism, and societal cohesion (Chesney & Citron, 2019). As deepfakes become increasingly convincing, the line between truth and fabrication blurs, leading to what some scholars have termed “reality apathy”, a state in which individuals become indifferent to the distinction between real and fake information (Landon-Murray, et al., 2019).

In essence, the information disorder framework provides a useful theoretical lens through which to analyze how AI technologies interact with misinformation, disinformation, and malinformation. AI not only enhances the creation and dissemination of false content but also challenges traditional methods of detecting and combating such content. Understanding these dynamics is essential for developing more effective strategies to counter AI-driven disinformation.

2.2. Information quality: Trustworthiness and Semantic accuracy

At the heart of the fight against disinformation is the issue of information quality, which emphasizes trustworthiness, coherence, and semantic accuracy over sheer volume and engagement. As Omoregie (2021a) argues, the prioritization of quantity over quality in social media algorithms has created an environment where low-quality, sensationalist content thrives at the expense of factual and reliable information.

Key concepts introduced in the information quality framework include:

- **Non-information.** Content that is grammatically correct but semantically meaningless, failing to convey any meaningful or actionable information.
- **Off-information.** Instructional but harmful content that, if acted upon, could lead to significant harm.

By framing disinformation as a failure of information quality, this approach underscores the need for systems that filter and prioritize content based on trustworthiness signals. Social media platforms must adopt algorithms that amplify high-quality content while de-prioritizing harmful or misleading narratives. Additionally, the development of semantic and logical filters (e.g., content verification tools) can help distinguish authentic information from falsehoods, aligning with broader efforts to improve the integrity of digital information ecosystems (Omoregie, 2021a).

2.3. The harm principle and Its application to AI-driven disinformation

John Stuart Mill's Harm Principle provides a valuable ethical foundation for evaluating the consequences of disinformation. According to Mill (1966), speech should only be restricted when it causes harm to others. In the context of AI-driven disinformation, the Harm Principle must be updated to address the unique challenges posed by the digital age, where falsehoods can spread rapidly and cause significant harm on a global scale.

Omoregie (2021b) expands on the Harm Principle by introducing criteria for evaluating the magnitude, likelihood, and timing of harm caused by falsehoods:

- **Magnitude of harm:** Ranges from minor to grave consequences.
- **Likelihood of harm:** Assesses whether harm is certain, probable, or improbable.
- **Timing of harm:** Considers whether harm is immediate, imminent, or long-term.

This framework provides a practical tool for policymakers and technology developers to assess the risks associated with AI-driven disinformation and design interventions accordingly. For example, while some forms of disinformation may cause immediate harm (e.g.,

health misinformation during a pandemic), others may have long-term implications for public trust and societal cohesion (e.g., deepfake videos undermining political figures).

2.4. The Role of social media and algorithmic amplification

Social media platforms play a central role in the propagation of disinformation, with their algorithms often prioritizing engagement metrics over content accuracy. By exploiting human tendencies toward confirmation bias and emotional reasoning, these algorithms create echo chambers that amplify false narratives and suppress diverse viewpoints.

AI-driven disinformation thrives in this environment, leveraging platform algorithms to achieve rapid virality and widespread reach. Omoregie (2021a, 2021b) and others argue that addressing this issue requires greater platform accountability and transparency in how content is ranked and promoted. Proposed solutions include:

- Developing trust signals to help users assess the reliability of content.
- Implementing algorithmic changes to prioritize high-quality, trustworthy information.
- Establishing regulatory frameworks to hold platforms accountable for the societal impact of their algorithms.

2.5. A multi-faceted response: Technology, Policy, and Education

The fight against AI-driven disinformation calls for a multi-pronged strategy that blends technological innovation, smart regulation, and widespread public education. Here's how these elements can work together:

- **Technological Solutions.** We need to develop and deploy advanced AI tools—such as sophisticated fact-checking algorithms and semantic analysis systems—that can quickly detect and counter disinformation.
- **Regulatory Frameworks.** Policies like the UK Online Safety Bill and the European Commission Code of Practice on Disinformation are essential. These regulations aim to hold online platforms accountable and help limit the spread of harmful content.
- **Media Literacy Education.** It's crucial to empower people with the skills to critically evaluate the information they encounter. By promoting media literacy, we can build a public that's more resilient to manipulation and better equipped to distinguish between credible and false information.

Taken together, these measures acknowledge that no single solution can tackle the complex challenges posed by AI-driven disinformation. Instead, a combination of technological, regulatory, and educational

initiatives is necessary to protect our information ecosystems and uphold democratic institutions.

This study is grounded in the understanding that the rapid spread of deepfake technologies poses a serious threat to information integrity. This issue is not unique to our work; for example, Melro and Pereira (2019) explored how undergraduates perceive disinformation and found that media and news literacy are key to mitigating the effects of fake news. Their research highlights the importance of cultivating critical thinking among young people—a finding that reinforces our argument for a strong educational response to the challenges brought on by deepfakes.

By highlighting the role of structured interventions in promoting media literacy, Melro and Pereira validate the framework employed here while extending its applicability to broader contexts of information disorder.

Similarly, Omoregie's (2021a; b) exploration of information quality offers a complementary perspective on the theoretical underpinnings of this study. His work underscores that combating the issue of disinformation, including deepfakes, is not merely about distinguishing truth from falsehood but also about establishing trust signals and enhancing the visibility of credible content. Omoregie's concept of filtering data through mechanisms of facts, logic, and semantics resonates with the ethical and technical framework proposed in this study. By applying these principles to deepfake regulation, such as prioritizing trustworthy content through algorithmic transparency, this paper not only reinforces Omoregie's arguments but also demonstrates the broader utility of this theoretical lens in addressing the complex interplay between disinformation and technology.

Further, the ethical dimensions of this study find resonance in works such as those by Tambini (2017) and Buckingham (2017), who explore the tensions between freedom of expression and content moderation. These scholars, like this paper, advocate for a multi-layered approach to managing digital content, integrating technological solutions, ethical considerations, and regulatory oversight. The incorporation of their perspectives highlights the adaptability and relevance of the theoretical framework applied here, showing that it is not limited to the issue of deepfakes but is equally effective in addressing the wider challenges of a post-truth society. Their work strengthens the argument that tackling deepfake-related disinformation requires a comprehensive strategy encompassing technical, educational, and regulatory solutions.

3. Case studies

3.1. COVID-19 misinformation

The COVID-19 pandemic marked one of the most significant public health crises of the 21st century, and alongside this global challenge came an unprecedented wave of misinformation and disinformation.

From the earliest days of the pandemic, social media platforms became flooded with false information, ranging from unverified treatments and conspiracy theories to fake claims about the origins and spread of the virus. AI-driven tools, especially bots and generative models, played a crucial role in amplifying these false narratives.

AI-powered bots were particularly effective in spreading misinformation about COVID-19, often mimicking human behavior to make disinformation seem credible. These bots disseminated false claims about the virus's origins— suggesting it was man-made or deliberately released— as well as misleading information about potential cures, including harmful treatments such as bleach consumption or unproven drugs like hydroxychloroquine. In this context, AI was used to churn out a staggering amount of content that blurred the line between reality and fiction, overwhelming public health authorities' attempts to provide accurate information (Ferrara et al., 2020).

One of the most significant challenges in combating COVID-19 misinformation was the rapid speed at which AI-generated content could be produced and distributed. AI bots often automated the process of creating and sharing thousands of posts per minute across multiple social media platforms. These bots were able to tailor content to different cultural and linguistic contexts, making misinformation a global issue that transcended borders (Agarwal et al., 2023). Misinformation surrounding the virus also took advantage of human cognitive biases— particularly the tendency to share emotionally charged or sensational content. AI-driven models were able to identify and exploit these psychological vulnerabilities, resulting in viral misinformation that spread faster than attempts to debunk it.

Moreover, AI played a role in the creation of deepfakes during the pandemic. Deepfake videos and audio clips circulated online, often falsely attributing statements to prominent public health officials, politicians, or celebrities. These fabrications typically involved manipulated videos of individuals endorsing unverified treatments or sharing conspiracy theories about COVID-19. One such video that gained significant attention featured a fabricated interview with a well-known health expert seemingly endorsing an untested remedy, which contributed to widespread confusion and vaccine hesitancy (Agarwal et al., 2023). While social media platforms eventually removed many of these deepfakes, the damage had already been done, as the videos had been widely shared and viewed by millions.

The impact of COVID-19 misinformation was profound and far-reaching. It undermined public trust in health authorities, caused confusion about the virus, and hindered efforts to control its spread. The World Health Organization (WHO) described the situation as an "infodemic", in which an overwhelming amount of false or misleading information complicated public health responses (Pomputius, 2019).

Public hesitancy toward government-recommended vaccines was, in part, exacerbated by AI-generated misinformation, particularly deepfakes that questioned the efficacy or safety of the vaccines.

Despite efforts to combat the spread of misinformation, the AI-driven infodemic exposed the vulnerabilities in existing public health communication systems. Health organizations were often unprepared to deal with the sheer scale of AI-generated disinformation, as fact-checking efforts and debunking campaigns lagged behind the viral nature of false claims. This case study highlights the critical need for stronger AI-driven detection systems and real-time fact-checking tools, particularly in times of crisis where accurate information can mean the difference between life and death.

3.2. Election interference

The 2020 U.S. presidential election underscored the growing threat of AI-driven disinformation in democratic processes. Election interference is not a new phenomenon, but AI has fundamentally transformed both the scale and the methods of influence campaigns. While traditional election interference tactics involved spreading misleading information through text-based fake news articles, the 2020 election witnessed the increasing use of deepfakes and AI-generated content to manipulate voters, undermine trust in institutions, and influence the outcome of the vote.

One of the most significant developments in this space was the use of deepfakes to create videos of political figures saying things they never actually said. These deepfakes were often highly convincing, leveraging the latest advancements in AI to create realistic facial movements, voice modulation, and body language. One deepfake video that circulated online during the campaign showed a fabricated speech by a presidential candidate endorsing controversial policies they had never supported. Although quickly debunked, the video had already garnered millions of views and shares, illustrating the difficulty of containing disinformation once it goes viral (Chesney & Citron, 2019).

AI's role in election interference was not limited to deepfakes; text-based disinformation also benefited from AI advancements. GPT-3 and similar language models were used to generate false narratives that mimicked the style and tone of legitimate news outlets, creating highly believable articles that were often shared on social media platforms (Brown et al., 2020). These AI-generated articles focused on sowing division and confusion, targeting specific voter groups with tailored messages designed to exploit their biases and fears. For instance, AI-generated disinformation was used to propagate the false narrative that mail-in voting was unreliable, causing distrust in the electoral process.

Another aspect of AI-driven election interference involved the micro-targeting of voters. By analyzing vast amounts of personal data collected from social media platforms, AI algorithms were able to

identify specific groups of voters who were most susceptible to disinformation. These algorithms then delivered tailored ads and content designed to influence their voting behavior. The use of AI to micro-target voters is not inherently unethical, but in the context of disinformation campaigns, it raises significant ethical concerns about the manipulation of voters' beliefs and opinions without their informed consent (Ferrara et al., 2020).

In addition to deepfakes and AI-generated disinformation, bot networks played a crucial role in amplifying false information. These bots were used to create the illusion of widespread support or opposition to certain candidates or policies, manipulating public perception of the election's dynamics. AI-driven bots were capable of posting thousands of tweets and comments across multiple platforms, artificially inflating the visibility of certain disinformation narratives (Zhou & Zafarani, 2020). This amplification effect made it difficult for genuine political discourse to take place, as the line between authentic and inauthentic content became increasingly blurred.

The consequences of AI-driven election interference are severe. They undermine trust in the democratic process, create polarization, and erode the legitimacy of election outcomes. The use of deepfakes, in particular, threatens to make video evidence— a form of media traditionally regarded as reliable and trustworthy— susceptible to manipulation. If deepfakes become widespread in future elections, they could lead to a scenario where voters no longer believe what they see, creating a profound crisis of trust in democratic institutions (Chesney & Citron, 2019).

Internationally, AI-driven election interference is not limited to the U.S. Other countries have experienced similar tactics, with AI being used to influence elections in Europe, Asia, and Latin America. In some cases, state actors have used AI to spread disinformation and create deepfakes aimed at destabilizing rival nations or weakening the legitimacy of their electoral processes. The 2020 election interference case study underscores the need for global cooperation in addressing AI-driven disinformation, as the threats are no longer confined by national borders (Landon-Murray et al., 2019).

4. Material and Methods

4.1. Detection of deepfakes and Analytical techniques

One of the most promising technologies is AI-based detection of deepfakes. Deepfake detection algorithms typically rely on machine learning models trained on large datasets of real and fake videos to identify inconsistencies. These models can detect anomalies such as unnatural facial movements, lighting inconsistencies, and pixel distortions. Companies such as DeepTrace and Facebook have been working on AI models capable of detecting deepfakes with high levels of accuracy by analyzing micro-expressions, eye movements, or subtle

changes in voice modulation (Lee et al., 2020). Despite this progress, detection tools are still playing catch-up with deepfake creation technologies, which are continuously improving in their ability to bypass detection algorithms (Chesney & Citron, 2019).

In addition to video detection, natural language processing (NLP) plays a significant role in combating text-based disinformation. Advanced NLP systems such as GPT-3 can generate misleading or entirely fabricated articles that are difficult to distinguish from legitimate news sources (Brown et al., 2020). To combat this, AI-powered NLP systems are being developed to analyze the language patterns and factual accuracy of content disseminated online. Fact-checking algorithms cross-reference content against established databases to detect inconsistencies or falsehoods in real-time. Google's Jigsaw team, for instance, has been experimenting with automated fact-checking tools that scan vast datasets of news articles and social media posts, flagging suspicious claims for human review (Pomputius, 2019). For video-based disinformation, Kietzmann and colleagues relied on platforms like YouTube and Facebook, where deepfakes and other synthetic media are more likely to be encountered. They focused on gathering videos that had been flagged for misleading or harmful content. Deepfake detection tools, including DeepFaceLab and XceptionNet, were utilized to identify anomalies in these videos that could suggest AI-generated manipulation (Kietzmann et al., 2020). While these tools hold promise, they are not without limitations. AI systems can be tricked by cleverly crafted disinformation that mimics reputable sources or includes partial truths.

A more robust and comprehensive AI-driven solution would integrate multiple modalities of analysis, combining text, video, and image detection with metadata analysis. By cross-referencing user behavior, content provenance, and engagement patterns, these systems could offer more holistic insights into how disinformation spreads and where it originates. For example, X (Twitter) has begun to employ machine learning models that identify bot networks responsible for amplifying disinformation, while Facebook uses AI to detect coordinated inauthentic behavior (Ferrara et al., 2020).

4.2. Data collection

To effectively analyze the impact of AI-driven disinformation, this research employed a multi-method approach, combining quantitative data collection with qualitative content analysis. Data was gathered from various online platforms, including X (Twitter), Facebook and YouTube social networks that have become breeding grounds for disinformation. The primary source of data consisted of user-generated content that had been flagged as misinformation or disinformation by fact-checking organizations, social media companies, and independent researchers. This dataset included both text-based content (such as fake

news articles) and multimedia content (such as deepfake videos). By using automated web-scraping tools and social media APIs, we were able to retrieve thousands of posts and comments related to misinformation from March 2020 to September 2023. These posts were labeled according to their type (misinformation, disinformation, malinformation), subject matter (e.g., COVID-19, elections), and the platforms on which they appeared.

In addition to gathering the content itself, we also collected metadata, such as engagement metrics (likes, shares, comments) and the network of accounts involved in spreading the disinformation. This data allowed us to track how disinformation spread, how it was amplified by AI-driven bots, and which user groups were most susceptible to engaging with and sharing false information.

4.3. Content analysis

Once the data was collected, by leveraging AI tools such as machine learning classifiers and NLP models, we collected, categorized, and analyzed a dataset of AI-generated disinformation. Using NLP models such as GPT-2 and GPT-3, we analyzed the linguistic patterns of text-based disinformation, identifying key themes, emotional triggers, and stylistic features that contributed to the viral spread of false content as what Brown et al. have done in 2020. Sentiment analysis was conducted to assess the emotional tone of the disinformation and how it resonated with different audiences. This allowed us to identify whether emotionally charged content was more likely to be shared and believed by users.

Fallowing Lee et al., (2020), for video-based disinformation, deepfake detection algorithms were applied to identify manipulations in visual and auditory content. These tools analyzed pixel-level inconsistencies, irregularities in facial movements, and discrepancies in voice modulation that are common in deepfake videos. By cross-referencing the detection results with user engagement metrics, we were able to assess the effectiveness of deepfakes in deceiving audiences and influencing public opinion.

To understand the role of AI-powered bots in spreading disinformation, same as Ferrara et al. (2020), we used bot detection algorithms such as Botometer and Tweepy. These tools allowed us to identify automated accounts that were responsible for amplifying disinformation on platforms like X (Twitter) and Facebook.

We also employed quantitative techniques to measure the spread and impact of AI-driven disinformation. Key metrics included:

- **Engagement Rate.** The number of likes, shares, and comments per piece of disinformation.
- **Virality score.** The speed at which disinformation was shared across different networks.

- **User reach.** The total number of users exposed to AI-generated disinformation.
- Bot activity. The volume and patterns of bot-generated disinformation.

These metrics provided insights into the effectiveness of AI in creating and disseminating disinformation, allowing us to draw connections between the content itself and its spread across platforms. By analyzing these metrics, we identified which types of AI-generated disinformation were most successful in influencing public opinion.

Given the sensitive nature of the content analyzed in this study, ethical guidelines were followed throughout the research process. All data collected was anonymized, and no personal information from individual users was used in the analysis. Data collection was limited to publicly available information, and we complied with the terms of service of the platforms from which the data was gathered. Additionally, the research team adhered to strict ethical guidelines to avoid further amplifying the disinformation studied, ensuring that any dissemination of results emphasized counteracting disinformation rather than inadvertently promoting it.

5. Discussion

5.1. AI's expanding role in disinformation

The future of disinformation is tightly intertwined with the ongoing advancements in artificial intelligence (AI) and machine learning (ML) technologies. As AI continues to develop, the capabilities for generating, disseminating, and personalizing disinformation will increase exponentially. Today's sophisticated AI models, such as OpenAI's GPT-3, already produce text that is difficult to distinguish from human-authored content. Future iterations of these language models are likely to become even more convincing, capable of generating entire disinformation campaigns with minimal human intervention. Moreover, as AI-generated content becomes more contextually aware and nuanced, the potential for AI to blur the lines between reality and fabrication will deepen, making it harder for individuals, platforms, and even governments to distinguish truth from falsehood.

One of the most alarming trends in the future of disinformation is the rapid improvement in AI-driven deepfakes. Deepfakes have already garnered global attention due to their potential for misuse. These technologies are expected to become even more accessible and sophisticated in the coming years. As AI-generated deepfakes evolve, they will be able to mimic not only facial expressions and voices with near-perfect accuracy but also subtle mannerisms, making detection even more difficult. The threat posed by deepfakes extends beyond mere entertainment or satire; it could potentially upend political

systems by creating false narratives during critical moments such as elections, geopolitical crises, or social unrest.

AI's role in disinformation also extends to its ability to manipulate images, sounds, and videos in ways that are more granular and precise than ever before. AI can now create "synthetic media"—entirely fabricated content that appears to be authentic but is generated without any input from real-world events (Maras & Alexandrou, 2019). With improvements in image generation technologies, such as GANs, AI can create fake news videos of events that never happened, implicating individuals in fictitious situations. These developments could have far-reaching implications in domains such as journalism, law enforcement, and intelligence, where the authenticity of evidence is paramount. If deepfakes or synthetic media are weaponized, trust in digital content could erode, making it difficult for individuals to rely on media, even in cases where the content is genuine.

Another likely development in the future of disinformation is the increasing personalization of disinformation campaigns. AI's ability to mine personal data and user behavior patterns makes it an ideal tool for tailoring disinformation to specific individuals or groups. By analyzing social media activity, browsing habits, and other digital footprints, AI can identify users' vulnerabilities, biases, and emotional triggers. It can then deliver targeted disinformation designed to exploit these weaknesses, making disinformation campaigns far more effective than blanket messaging strategies. This level of precision was already seen during the 2016 and 2020 U.S. elections, where AI-driven algorithms were used to micro-target voters with personalized disinformation based on their political preferences and online behavior. As AI becomes more adept at understanding human behavior, we can expect to see increasingly customized disinformation campaigns that are far more persuasive and difficult to counteract.

Moreover, AI could be used to generate disinformation at an industrial scale, flooding the internet with false information and overwhelming efforts to combat it. AI can create thousands of fake accounts and bots, each sharing, retweeting, and reposting fabricated content, giving the illusion of consensus or grassroots movements where none exists. These botnets can manipulate public perception, influencing everything from stock prices to public health decisions. In the future, we may witness entire disinformation "ecosystems" where AI-generated content dominates the digital landscape, drowning out factual information and making it exceedingly difficult for individuals to discern the truth.

5.2. Ethical considerations

As AI's role in generating and spreading disinformation grows, the ethical challenges surrounding these technologies will become more pressing. The rise of AI-driven disinformation presents a myriad of

ethical dilemmas, especially in the realms of free speech, privacy, and accountability. One of the primary concerns is how to regulate AI-driven content without infringing on individuals' rights to free expression. While governments may wish to crack down on disinformation, particularly when it threatens public safety or national security, they must do so without stifling legitimate dissent or curbing freedom of the press.

A key ethical question is whether AI developers and platforms should be held accountable for the misuse of their technologies. The creators of AI models like GPT-3 or GANs did not design these tools with the explicit intent of facilitating disinformation, yet these models are being co-opted for nefarious purposes. Should AI developers bear some responsibility for the actions of bad actors who misuse their creations? Similarly, social media platforms have a duty to prevent the spread of disinformation, but where do we draw the line between proactive content moderation and censorship? These questions are becoming increasingly urgent as AI-generated disinformation becomes more prevalent.

Another critical ethical issue is the potential impact of AI-driven disinformation on vulnerable populations. Targeted disinformation campaigns often exploit racial, ethnic, or religious tensions, deepening societal divides and exacerbating existing inequalities. In countries with weaker regulatory frameworks, disinformation campaigns can incite violence, destabilize governments, and undermine public trust. As AI-generated disinformation becomes more sophisticated, these campaigns will likely become even more effective in manipulating public sentiment and fomenting unrest. The ethical responsibility to protect vulnerable populations from these malicious tactics should be a priority for governments, international organizations, and tech companies alike.

Furthermore, AI-driven disinformation presents a challenge to our legal and moral frameworks surrounding truth and authenticity. If we can no longer trust the evidence presented to us—whether in the form of videos, images, or text—how do we adjudicate legal disputes or verify facts in journalism? The erosion of trust in digital media could have far-reaching consequences, leading to Rreality Apathy in the society, where people no longer believe in any form of media (Landon-Murray et al., 2019). In such a world, the ability to manipulate public perception with AI-generated content could render truth irrelevant, fundamentally altering the way we interact with information, each other, and democratic institutions.

5.3. Technological developments

While AI's role in advancing disinformation is concerning, AI also holds the key to its detection and mitigation. The future will likely see the continued development of AI-powered tools capable of identifying deepfakes and other forms of synthetic media. These tools, such as the

ones being developed by Facebook, Google, and independent researchers, rely on analyzing the digital fingerprints left by AI-generated content, such as pixel anomalies, unnatural facial movements, or inconsistencies in voice modulation (Lee et al., 2020). However, as deepfake creators become more adept at hiding these clues, detection tools will need to evolve in tandem. In the future, we may see AI systems that are capable of "self-checking," where the same algorithms that generate deepfakes are used to detect and neutralize them.

Blockchain technology also holds promise as a solution to the spread of disinformation. By creating an immutable ledger of content creation, blockchain could be used to verify the authenticity of digital media, ensuring that videos, images, and text are traceable back to their original source (European Commission, 2018). This could prevent the spread of deepfakes and other fabricated content by allowing platforms, journalists, and users to verify the provenance of the content they encounter. Blockchain-based content verification could also serve as a deterrent for those looking to create and disseminate false information, as their creations would be easily traceable.

In addition to detection tools, AI may also play a role in countering disinformation by flooding the internet with factual information. This "counter-disinformation" approach would involve using AI to identify trending disinformation narratives and create targeted fact-based content designed to debunk false claims. AI-generated content could be deployed in real-time to provide users with alternative perspectives, verified information, and fact-checked resources. This approach, already being explored by initiatives like Google's Jigsaw project, could help mitigate the impact of disinformation by ensuring that factual information reaches as many people as possible (Pomputius, 2019).

5.4. Political considerations

One promising yet underexplored solution for verifying digital content is blockchain technology. Imagine a system where every image, video, or text snippet is stamped with a tamper-proof record at its origin. Blockchain could do exactly that—providing a decentralized ledger that confirms authenticity and tracks how content is created and shared. In fact, the European Union and several private companies are already experimenting with this approach to bolster content verification in journalism and the media (European Commission, 2018).

Another idea gaining traction is the use of "watermarking" for AI-generated content. Under this scheme, anyone producing deepfakes or manipulated media would be required to embed a hidden, traceable code within their work. This watermark could help track the content back to its source, making it easier to hold creators accountable.

On the regulatory side, governments might consider imposing fines on platforms that fail to remove harmful AI-generated content or

deepfakes in a timely manner. For example, the EU's Digital Services Act (DSA) aims to hold online platforms to higher standards when it comes to taking down illegal or damaging content—while still safeguarding freedom of expression. The challenge, of course, is to strike a balance between curbing disinformation and protecting the right to free speech.

Given the global scale of AI-driven disinformation, no single nation can tackle the issue alone. International cooperation is essential. Countries need to come together to set global standards for detecting, mitigating, and even prosecuting disinformation campaigns. The European Union's Code of Practice on Disinformation is a step in that direction, as it unites governments, social media companies, and civil society organizations in developing best practices (European Commission, 2018).

Beyond regional initiatives, international bodies like the United Nations or the G20 could help establish treaties or agreements to govern the ethical use of AI in content creation and distribution. Such agreements might, for instance, ban the use of AI for disinformation in political campaigns or set restrictions on state-sponsored propaganda, thereby creating a framework for holding perpetrators accountable (Conley & Vilmer, 2024).

At the same time, enhancing public understanding of AI-driven disinformation is crucial. Public education campaigns could empower people to critically assess online information, spot deepfakes, and recognize AI-generated news. Integrating media literacy into school curricula would also prepare future generations to navigate a complex digital landscape (Kahne & Bowyer, 2017).

Investing in interdisciplinary research is equally important. Governments and academic institutions should support studies that explore the societal impacts of AI-generated disinformation and develop innovative ways to counter it. By bringing together experts in fields like AI, ethics, journalism, political science, and law, we can work toward comprehensive solutions for this multifaceted problem (Chesney & Citron, 2019).

Finally, the responsibility for ethical AI starts at the source. AI developers should be encouraged—or even required—to build ethical safeguards into their systems. This could mean designing fail-safes that prevent AI from generating harmful content, or issuing transparency reports that explain how AI tools are used in creating digital media (Floridi et al., 2018).

In short, while AI-driven disinformation poses a serious challenge, a combination of blockchain verification, regulatory oversight, international collaboration, public education, interdisciplinary research, and ethical AI design offers a hopeful path forward.

6. Conclusion

The rapid evolution of disinformation from text-based fake news to sophisticated AI-generated content such as deepfakes marks a defining moment in the history of information warfare. As AI technologies continue to advance, the threat posed by AI-driven disinformation becomes more acute, demanding a multifaceted response from technologists, policymakers, academics, and society at large. This paper has highlighted how artificial intelligence, while offering immense potential in many fields, also serves as a double-edged sword when exploited for malicious purposes. AI has enabled the mass production and dissemination of disinformation at a scale previously unimaginable, undermining trust in media, eroding the democratic process, and sowing discord in society.

The rise of deepfakes exemplifies the unprecedented challenges we now face. Deepfakes have the potential to mislead the public, distort political processes, and damage reputations in ways that traditional forms of disinformation could not. As highlighted by the case studies on COVID-19 and election interference, these technologies are already being weaponized to manipulate public opinion, exacerbate societal divisions, and undermine trust in institutions. The implications for democracy, governance, and international relations are profound.

However, alongside the growing sophistication of disinformation techniques, significant efforts are underway to counteract the negative impacts of AI-generated disinformation. AI detection tools are being developed to identify deepfakes, bots, and other forms of synthetic content. However, while these technological solutions offer some promise, they are not a panacea. The arms race between deepfake creators and those developing detection technologies is likely to continue, with bad actors constantly finding new ways to bypass detection systems. Moreover, the ethical dilemmas posed by AI-driven disinformation, including concerns about censorship, privacy, and free speech, add complexity to the debate on how best to regulate and control this emerging threat.

Addressing AI-driven disinformation requires more than just technological fixes. It necessitates a comprehensive policy framework that balances the need for regulation with the protection of individual rights. Governments and international organizations must collaborate to create global standards for content moderation, transparency, and accountability in AI use. Social media platforms, in turn, need to take a more proactive stance by deploying AI to detect and mitigate the spread of false information and by being more transparent about the algorithms they use to prioritize content. The implementation of blockchain-based content authentication, coupled with AI-driven detection systems, may offer a way to ensure the provenance of digital media and prevent deepfakes from spreading unchecked.

In addition to technological and policy solutions, public education

and media literacy initiatives are vital. By equipping individuals with the skills needed to critically evaluate the information they encounter online, society can build resilience against disinformation. Schools, universities, and public institutions must take an active role in teaching media literacy, promoting critical thinking, and encouraging skepticism toward sensational or emotionally charged content. A well-informed public is the first line of defense against the pervasive influence of AI-generated disinformation.

The ethical responsibility also extends to the developers and users of AI technologies. AI research and development must incorporate ethical guidelines and safeguards to prevent misuse. Developers must be held accountable for the applications of their technologies, and companies should be transparent about how their AI systems are used in content creation and dissemination. Creating an ethical framework for AI development will help mitigate the risks associated with AI-driven disinformation while still allowing innovation to flourish.

Ultimately, the battle against AI-driven disinformation is not one that can be won by technology alone. It will require an ongoing and coordinated effort that brings together stakeholders from all sectors of society. As AI continues to evolve, so too must our strategies for combating the dark side of these technologies. By fostering cooperation between governments, tech companies, and civil society, and by promoting a culture of digital literacy and transparency, we can mitigate the harmful effects of AI-driven disinformation and ensure that the digital information ecosystem remains a trusted and reliable space.

Conflict of interest

The author declared no conflicts of interest.

Authors' contributions

All authors contributed to the original idea, study design.

Ethical considerations

The author has completely considered ethical issues, including informed consent, plagiarism, data fabrication, misconduct, and/or falsification, double publication and/or redundancy, submission, etc. This article was not authored by artificial intelligence.

Data availability

The dataset generated and analyzed during the current study is available from the corresponding author on reasonable request.

Funding

This research did not receive any grant from funding agencies in the public, commercial, or non-profit sectors.

References

Agarwal, B.; Agarwal, A.; Harjule, P. & Rahman, A. (2023). Understanding the intent behind sharing misinformation on social media. *Journal of Experimental & Theoretical Artificial Intelligence*. 35(4): 573-587. https://doi.org/10.1080/0952813X.2021.1960637.

Allcott, H. & Gentzkow, M. (2017). "Social media and fake news in the 2016 election". *Journal of Economic Perspectives*. 31(2): 211-236. https://doi.org/10.1257/jep.31.2.211.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A. & Agarwal, S. (2020). "Language models are few-shot earners". *Advances in Neural Information Processing Systems*. 33: 1877-1901.

Buckingham, D. (2017). *Fake news: is media literacy the answer*. David Buckingham.

Chesney, R. & Citron, D. (2019). "Deepfakes and the new disinformation war: The coming age of post-truth geopolitics". *Foreign Aff.* 98: 147.

Conley, H.A. & Vilmer, J.J. (2024). *Successfully Countering Russian Electoral Interference*. https://www.csis.org/analysis/successfully-countering-russian-electoral-interference.

European Commission. (2018). *European Commission launches the EU Blockchain Observatory and Forum*. https://ec.europa.eu/commission/presscorner/detail/en/ip_18_521.

Ferrara, E.; Varol, O.; Davis, C.; Menczer, F. & Flammini, A. (2016). "The rise of social bots". *Communications of the ACM*. 59(7): 96-104. https://doi.org/10.1145/2818717.

Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; … & Vayena, E. (2018). "AI4People—An Ethical Framework for a Good AI Society: Opportunities, risks, principles, and recommendations". *Minds and Machines*. 28(4): 689-707. https://doi.org/10.1007/s11023-018-9482-5.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. & Bengio, Y. (2014). "Generative adversarial nets". *Advances in Neural Information Processing Systems*. 27.

Kahne J. & Bowyer, B. (2017). "Educating for democracy in a partisan age: Confronting the challenges of motivated reasoning and misinformation". *American Educational Research Journal*. 54(1): 3-34. https://doi.org/10.3102/0002831216679817.

Kietzmann, J.; Lee, L.W.; McCarthy, I.P. & Kietzmann, T.C. (2020). "Deepfakes: Trick or treat?". *Business Horizons*. 63(2): 135-146. https://doi.org/10.1016/j.bushor.2019.11.006.

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

Maras, M.H. & Alexandrou, A. (2019). "Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos". *The International Journal of Evidence & Proof*. 23(3): 255-262. https://doi.org/10.1177/1365712718807226.

Melro, A. & Pereira, S. (2019). "Lažne ili istinite? Percepcije studenata preddiplomskih studija o (dez) informacijama i kritičkom razmišljanju". *Medijske Studije*. 10(19): 46-67. https://doi.org/10.20901/ms.10.19.3.

Mill, J.S. & Mill, J.S. (1966). *On liberty*. Macmillan Education UK.

Omoregie, U. (2021a). "Information disorder online is an issue of information quality". *Academia Letters*. 2. http://dx.doi.org/10.20935/AL2999.

----------. (2021b). "The 'Harm Principle'and Information Disorder Online". *Academia Letters*. https://doi.org/10.20935/AL3425.

Pomputius, A. (2019). "Putting misinformation under a microscope: Exploring

technologies to address predatory false information online". *Medical Reference Services Quarterly*. 38(4): 369-375. https://doi.org/10.1080/02763869.2019.1657739.

Tambini, D. (2017). *Fake News: Public Policy Responses*. https://core.ac.uk/download/pdf/80787497.pdf.

Trittin-Ulbrich, H.; Scherer, A.G.; Munro, I. & Whelan, G. (2021). "Exploring the dark and unexpected sides of digitalization: Toward a critical agenda". *Organization*. 28(1): 8-25. https://doi.org/10.1177/1350508420968184.

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

Wardle, C. & Derakhshan, H. (2017) *Information disorder: Toward an interdisciplinary framework for research and policymaking*. Vol. 27, pp. 1-107. Strasbourg: Council of Europe.

Zhou, X. & Zafarani, R. (2018). "A survey of fake news: Fundamental Theories, Detection Methods, and Opportunities". *ACM Computing Survry*. 53(5): 1-40. https://doi.org/10.1145/3395046.