Disinformation and Generative AI: Risks, Challenges, and Possible Solutions

Scientific paper DOI: 10.18485/judig.2025.1.ch9

Ana Kovačević[1] *0000-0003-4928-9848*

Abstract

The disinformation phenomenon has been present since the very dawn of human communication, with repeated exposure to false information often resulting in its acceptance as truth. In contemporary society, however, disinformation has become increasingly alarming due to the simple, inexpensive, and convincing way it can be created using generative artificial intelligence and large language models in particular. These models, which continue to make significant advancements, enable the generation of large volumes of content that appear credible, even in the national language. In addition, such models facilitate the creation of personalised content tailored to certain groups or individuals, citing seemingly credible sources, and thus serving to further strengthen the impact of disinformation. In a world inundated with information, this problem has become even more pronounced. One of the key channels for the spread of disinformation is social media, whose primary goal is to capture and maintain its users' attention with content that reinforces their existing beliefs. This paper will present the potential uses of generative AI in the creation of disinformation, as well as possible solutions to address this problem.

Key words: generative artificial intelligence, disinformation, the risks of artificial intelligence, large language models, social media

1. Introduction

Disinformation has long challenged societies throughout history, from early propaganda to modern disinformation campaigns. Today,

[1] Faculty of Security Studies, University of Belgrade, kana@fb.bg.ac.rs

generative artificial intelligence (AI) amplifies these challenges by enabling the large-scale, automated production of highly convincing false narratives. Unlike traditional manual propaganda, AI-driven tools are able to rapidly produce text, images, and deepfake videos almost entirely automatically with minimal human intervention (Goldstein et al., 2023), thereby influencing political discourse and individual decision-making.

According to Gartner's (2024) CIO and Technology Executive Survey, AI adoption across industries stands at 43%, while generative AI implementation has reached 37%. Projected investment growth recorded from 2024 to 2025 is also significant: 84% in AI, 86% in generative AI, and 87% in cybersecurity (Gartner, 2024). These figures highlight not only the growing reliance on AI, but also the urgency of addressing the security and ethical concerns this entails. Reflecting this need, numerous organisations now prioritise solutions to counter disinformation: Alaverz (2024) predicts that by 2028, 50% of enterprises will deploy specialised products or services for this very purpose—up from under 5% in 2024.

A development of particular concern is generative AI's ability to automate deceptive content creation, reducing or even eliminating the need for human oversight. Despite hypothetical misuse scenarios being widely debated in scientific literature, real-world incidents confirm the growing scope of these threats. Analysing over 200 such cases, Marchal et al. (2024) concluded that the majority of documented misuse involves accessible AI tools requiring minimal technical expertise, with 26% of instances aimed at manipulating public opinion, primarily through disinformation.

Even trusted software platforms, including those from Adobe, Microsoft, or Google, are incorporating AI features, making sophisticated manipulation increasingly accessible to general users (Schneier & Sanders, 2024). As these capabilities expand, concerns about privacy, security, and the erosion of public trust continue to grow.

2. Disinformation and Social Media

Disinformation refers to false information that is intentionally shared to mislead others (Aimeur et al., 2023). One of the subsets of disinformation is 'fake news,' which specifically presents false information as news. Other related terms in the literature include misinformation and malinformation. Misinformation is false information shared without the intent to mislead or cause harm, whereas malinformation involves sharing genuine information with the intent to cause harm (Aimeur et al., 2023).

Social media plays a pivotal role in spreading disinformation because of its affordability, accessibility, and rapid dissemination of content. According to a 2024 report, the number of active social media user accounts worldwide totals 5.22 billion, with users spending an average of 2 hours and 19 minutes a day on these platforms (DataReportal, 2024). Given this massive user base, the creation of text, images, and videos via generative AI, along with the ease of information sharing, has led to a dramatic rise in disinformation. This trend poses serious consequences for both individuals and society. Considering both the number of users and newly created posts, the challenge lies in delivering information to the most relevant audiences while ensuring its credibility. The main goal of social media is profit, often achieved by maximizing user engagement and satisfaction. Accordingly, algorithms are designed to present "relevant" content. In the pursuit of this objective, users are often presented with material which reinforces their existing beliefs and interests, creating so-called "echo chambers." Certain posts also gain higher visibility, a process termed "boosting", to increase user interactions such as likes, shares, comments, and clicks. This approach often promotes the creation of controversial or sensational topics which trigger emotional reactions, often at the expense of accuracy and relevance (Kovačević & Demić, 2023).

The lack of transparency in the algorithms used by large corporations also remains a major concern. Partial insights have been gained by cross-referencing publicly available information from social media platforms with statements from former employees, revealing how different algorithmic approaches can significantly alter the types of content users encounter (Oremus et al., 2021). These algorithms may also serve to shape users' perceptions of disinformation.

One important mechanism is "demotion," also referred to as downranking, reduction, or suppression. It is considered a form of "soft" content moderation, whereby problematic content is shown to a smaller audience rather than being removed outright (Gillespie, 2022). According to Narayanan (2023), a 10% reduction in visibility has a minor impact on a post's reach, whereas a 20% reduction can lower its reach tenfold, confining it mostly to the poster's immediate network. Since posts still appear in followers' feeds, —and because reach often fluctuates naturally, demotion often goes unnoticed by both content creators and consumers alike (Narayanan, 2023).

Facebook's algorithm ranks posts according to a user's interaction history, relationship with the content creator, preferred content formats, recency, and source quality (Narayanan, 2023). The objective is to prioritise those posts users are most likely to find engaging or relevant, placing them at the top of their News Feed, based on the likelihood of interaction. As user interests and behaviours evolve, the algorithm continuously adjusts these rankings and personalises the feed accordingly.

2.1. Why Disinformation Spreads

Many people struggle to differentiate between genuine and fabricated news and are susceptible to cognitive biases such as naïve realism and confirmation bias, which accelerate the spread of disinformation (Kovačević & Demić, 2023). Individuals have a tendency to judge a source's credibility based on the opinions of their peers, and repeated exposure to the same content can make false information seem true. These challenges are further compounded by information overload and the advent of generative artificial intelligence.

Once a misconception takes root, correcting it becomes extremely complex. A study by Nyhan and Reifler (2010) found that attempts to refute false information by presenting accurate facts can, at times, even reinforce the original misunderstanding. Similarly, research conducted by Schwalbe et al. (2024) demonstrated that both supporters and opponents of Trump were more likely to believe fabricated headlines such as "Trump Beats Grandmaster Chess Champion" or "Trump Attended Private Halloween Gala with Sex Orgies Dressed as the Pope" when these headlines aligned with their preexisting views, compared to true headlines that did not.

Changing people's opinions is inherently difficult, regardless of how persuasive the counterarguments might be. It is much easier, however, to manipulate existing emotions, reinforce preexisting beliefs, and incite animosity (Helming & Marsh, 2023). Designed to maximise user engagement, social media platforms typically reward outrageous or emotionally charged content with higher interaction levels (Helming & Marsh, 2023).

To boost engagement, algorithms may promote polarising content, creating a cycle in which political campaigners produce inflammatory material to attract media attention. As a result, regular users may unwittingly propagate those false statements that resonate with them, thereby perpetuating the cycle of misinformation (Helming & Marsh, 2023).

This dynamic is further illustrated by research carried out by McLoughlin et al. (2024), who examined the relationship between disinformation and moral outrage, a mix of anger and disgust triggered by perceived moral transgressions (Salerno & Peter-Hagene, 2013). In their study analysing Facebook and Twitter data alongside behavioural experiments, they found that misinformation sources tend to evoke more heightened moral outrage than reliable news outlets, which in turn promotes further sharing of misinformation (McLoughlin et al., 2024). In numerous cases, users share outrage-inducing content without even reading it, suggesting that interventions based solely on the desire for factual accuracy may be ineffective.

Moreover, outrage is highly engaging, and social media posts that provoke anger typically receive more likes and shares, further intensifying emotional responses through ranking algorithms (Brady et al., 2021). Consequently, disinformation designed to trigger outrage is particularly resistant to countermeasures like fact-checking or accuracy prompts, which assume that users are primarily motivated by truth (Osmundsen et al., 2021; Rathje et al., 2023). Additionally, individuals who express outrage are often perceived as more trustworthy (Jordan & Rand, 2019), making outrage an effective strategy for those seeking to spread disinformation.

3. Community Notes

Since taking over X (formerly Twitter), Elon Musk has dismissed over 80% of the platform's trust and safety engineers and a third of related staff, creating a "perfect storm" for the spread of abusive content online, according to Australia's online safety commissioner (Brewster, 2024). The goal was to reduce censorship by leveraging crowdsourced intelligence, specifically through *Community Notes*. On the X platform, Community Notes have been implemented to allow registered contributors to fact-check tweets and add contextual notes. Other contributors then rate these notes as helpful or not helpful. An algorithm calculates a helpfulness score for each note based on ratings from contributors with diverse perspectives. Only those that achieve a high helpfulness score are labelled “Helpful” and then displayed on the tweet.

In their large-scale empirical study, which analysed the efficiency of Community Notes in reducing engagement with misinformation on X, Chuai et al. (2024) found no evidence that Community Notes effectively lowered engagement with misleading tweets. They suggest this may be due to the system’s slow response during the early, viral stages of misinformation, highlighting the need for the continuous evaluation and improvement of crowdsourced fact-checking strategies (Chuai et al., 2024).

However, starting in January 2025, Meta announced the discontinuation of its third-party fact-checking programme, adopting instead a Community Notes system similar to that used by X (Kaplan, 2025).

3.1. Information Labelling

Due to the lack of transparency in social media news-ranking algorithms and their impact on users, an experiment was designed to compare the effectiveness of algorithmic interventions with informational interventions (Kaiser & Mayer, 2023). Under the algorithmic approach (deamplification), the visibility of disinformation was reduced by modifying the content-ranking algorithm, while informational interventions included labels, fact-checks, and informational panels to provide users with additional context. In collaboration with the DuckDuckGo platform, the researchers gained insight into real user interactions via search results leading to websites containing disinformation. A total of 500,000 users participated in the experiment. The findings revealed that algorithmically reducing the visibility of disinformation significantly decreases its spread (by 50%), whereas informational interventions such as warning labels and fact-check tags have a statistically negligible effect on reducing user engagement with disinformation (Kaiser & Mayer, 2023). These results are in line with the Carnegie Endowment for International Peace report (Bateman & Jackson, 2024), which indicated that fact-checking and labelling false information yield only modest results. Moreover, they observed that top-down labelling tends to reinforce the very mindset that makes people more prone to conspiracy theories, namely, the belief that "elites" dictate what they should think (Bateman & Jackson, 2024).

4. Using gen AI for Creating Disinformation

There are numerous hypothetical and real examples of disinformation. As Marchael et al. noted, a large number of real-life cases analysed involve the misuse of general AI through disinformation, typically requiring little to no technical expertise (Marchael et al., 2024). The study conducted by Sallami et al. (2024) investigated the potential for generating disinformation using various large language models (LLMs), including *Phi-3, Gemma-1.1, Mistral, Llama-3, C4AI, Zephyr-ORPO, and GPT-4*. The findings revealed that although LLMs can easily create biased disinformation, some models implement safety protocols to prevent the generation of biased or misleading content. The following is an example of disinformation generated using ChatGPT-4.

Create texts and photography of Spathiphyllum (smaller one) for Facebook, explaining how the Spathiphyllum plant can purify the air from PM 2.5 particles. Text should have 4–5 sentences. Refer to a relevant public health institution that confirms this research as well as famous scientific institutions or journals.

Using LLMs makes it easy to automate the process of creating targeted disinformation, from scripts created for downloading publicly posted information to creating targeted news for a specific audience. The major problem lies in the sheer volume that can be created quickly.

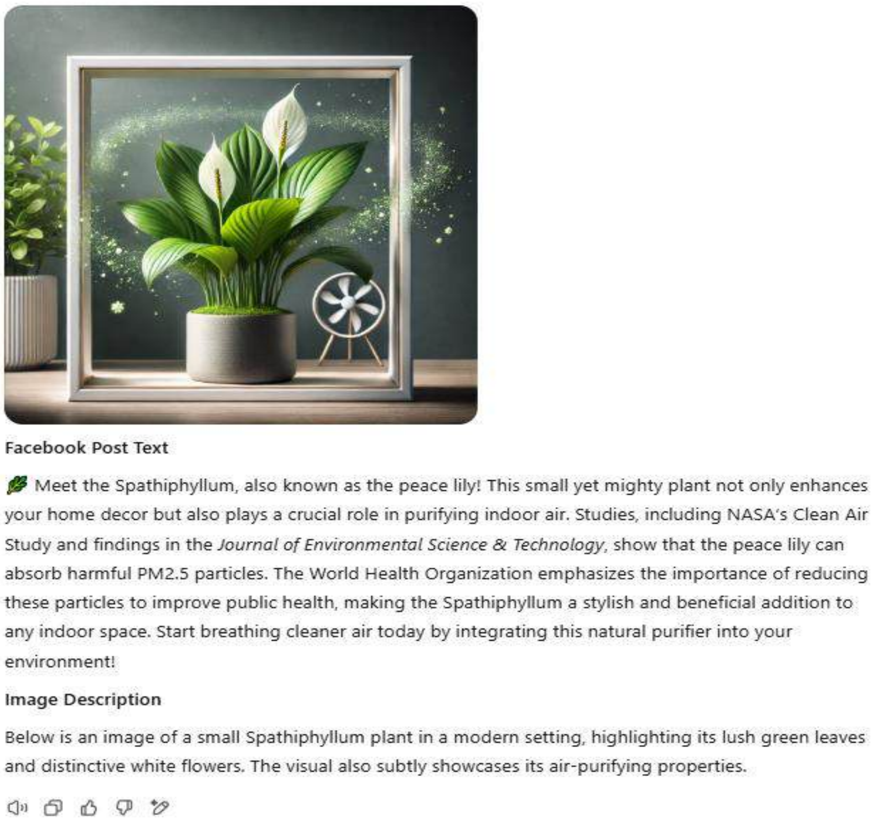

Figure 1. Example of creating disinformation

5. Solutions

As we have already mentioned, crowdsourcing interventions on platform X ("Community notes"), as well as information labelling, have had mixed success in preventing the spread of disinformation. Automatic disinformation detection poses a significant challenge because the content is crafted to closely mimic the truth, making it difficult for AI to assess its accuracy independently. As a result, it is essential to explore more effective strategies to address the issue of disinformation on social media. One possible solution is to use generative AI in disinformation detection.

Sallami et al. (2024) analysed the capabilities of LLMs to detect disinformation and assess the quality of their explanations. The study showed that larger models generally demonstrated superior performance in disinformation detection. However, some smaller models, such as Gemma-1.1, exhibited unexpectedly high efficiency due to sophisticated training methods, particularly Reinforcement Learning with Human Feedback (RLHF). According to human evaluators, GPT-4 and Llama-3 provided the highest quality explanations of disinformation, scoring highly on criteria such as clarity, relevance, accuracy, and comprehensiveness. Additionally, the study also showed that LLMs encountered difficulties in recognising disinformation generated by other LLMs, especially when these texts included fabricated citations from scientific studies. In such cases, the models often provided indecisive responses, acknowledging potential issues but failing to confidently label the content as fake. The evaluation indicated that while LLMs outperform traditional disinformation detectors, they also exhibit certain weaknesses, particularly when faced with AI-generated disinformation. The authors suggest that although LLMs hold significant potential for disinformation detection, improvements in transparency and the reliability of result interpretation are essential. The authors also note that after reading the LLM explanations, 40% of users changed their views regarding the authenticity of the news, while 20% became even more uncertain than before. This finding suggests that certain solutions might inadvertently introduce additional confusion (Sallami et al., 2024). Moreover, the issue may also stem from the biases inherent in the disinformation itself.

Is the following text disinformation " Meet the Spathiphyllum, also known as the peace lily! This small yet mighty plant not only enhances your home decor but also plays a crucial role in purifying indoor air. Studies, including NASA's Clean Air Study and findings in the Journal of Environmental Science & Technology, show that the peace lily can absorb harmful PM2.5 particles. The World Health Organization emphasizes the importance of reducing these particles to improve public health, making the Spathiphyllum a stylish and beneficial addition to any indoor space. Start breathing cleaner air today by integrating this natural purifier into your environment!"

a) Prompt

Reasoned for 4 seconds >

The text contains several misleading claims and overstatements that could qualify it as disinformation, particularly in the context of how it represents scientific studies and the plant's air purifying capabilities. Here are the key points explaining why:

1. **Misrepresentation of NASA's Clean Air Study:**
   NASA's Clean Air Study, conducted in the 1980s, focused on the ability of certain plants to remove volatile organic compounds (VOCs) from sealed environments—not on particulate matter like PM2.5. Claiming that this study shows the peace lily absorbs PM2.5 particles is inaccurate.
2. **Exaggerated Efficacy in Removing PM2.5:**
   While some studies have explored the role of plants in improving indoor air quality, the research does not support the idea that a peace lily can significantly reduce PM2.5 levels in real-world, everyday settings. The conditions in controlled laboratory studies differ greatly from those in typical indoor environments, where the amount of plant biomass needed to impact PM2.5 levels would be impractically large.
3. **Improper Association with WHO Recommendations:**
   The text links the World Health Organization's emphasis on reducing PM2.5 (due to its health risks) with the purported ability of the peace lily to act as a solution. This association is misleading because WHO recommendations are based on comprehensive public health measures and industrial solutions—not on the use of houseplants.
4. **Potential Consumer Misinformation:**
   By suggesting that simply adding a peace lily to your home can lead to significantly cleaner air, the text may lead consumers to underestimate the importance of proven air purification methods (such as mechanical ventilation and filtration systems). This oversimplification can divert attention from more effective strategies for improving indoor air quality.

Conclusion:

While the peace lily is indeed a popular and aesthetically pleasing plant that may have some minor benefits for indoor air quality, the claims made in the text are not fully supported by scientific evidence. The misrepresentation of studies and the exaggeration of the plant's capabilities contribute to a misleading narrative. Therefore, the text can be considered a form of disinformation as it misleads readers about the true benefits of the peace lily in purifying indoor air.

b) answer by ChatGPT-03-mini-high

Accurate Elements

1. **NASA's Clean Air Study**: The study (1989) did identify plants, including the peace lily, as capable of removing **volatile organic compounds (VOCs)** like benzene and formaldehyde in controlled lab settings.
2. **WHO's stance on PM2.5**: The WHO does highlight the dangers of PM2.5 (fine particulate matter) and urges its reduction for public health.

Misleading/Inaccurate Claims

1. **PM2.5 Absorption by Peace Lilies**:
   - There is **no credible evidence** that peace lilies (or most houseplants) significantly reduce PM2.5 particles indoors. PM2.5 refers to airborne particulates (e.g., dust, smoke), not gases like VOCs. Plants are not effective at filtering these particles, which require mechanical air purifiers or improved ventilation.
   - NASA's study focused on **VOCs, not PM2.5**, and even its conclusions about VOCs are debated in real-world settings due to the unrealistic lab conditions (e.g., extreme plant density and airflow).
2. **Misleading Citation of the Journal of Environmental Science & Technology**:
   - No specific study from this journal is cited, and there is no peer-reviewed research confirming the peace lily's ability to absorb PM2.5. This vague reference creates a false impression of scientific consensus.
3. **Overstatement of Benefits**:
   - The text implies peace lilies are a "crucial" or "natural purifier" for PM2.5, which is false. While plants may marginally improve air quality, they are not a practical solution for particulate pollution.

Why This Is Disinformation

- **False Cause-Effect Relationship**: Linking NASA's study (about VOCs) to PM2.5 reduction misrepresents the science.
- **Exploiting Trusted Institutions**: Citing NASA and the WHO gives the claims undue credibility, even though neither organization endorses houseplants as a solution for PM2.5.
- **Potential Harm**: Overstating the plant's benefits could lead people to neglect proven solutions (e.g., air purifiers, reducing pollution sources).

Conclusion

The text is **disinformation** because it:

- Falsely attributes PM2.5 reduction capabilities to peace lilies.
- Misuses reputable studies (NASA) and institutions (WHO) to legitimize unsupported claims.
- Exaggerates the plant's air-purifying effects, which are marginal at best in real-world conditions.

c) answer by DeepThinkR1

Based on the previous example, it is clear that LLMs show promise in detecting disinformation, though questions remain about their overall effectiveness.

Additionally, LLMs exhibit inherent limitations and biases and often struggle with tasks requiring common sense and extensive world knowledge, potentially reinforcing existing stereotypes present in their training data (Eigner et al., 2024). Furthermore, since LLMs are unable to fully capture the subtleties of human language, they may sometimes generate responses that are not entirely accurate or relevant (Kojima et al., 2022). Consequently, while improving the quality of training data, restricting user access to generative AI tools, and employing detection methods (such as watermarking) may help to mitigate the misuse of generative AI, these measures alone are insufficient to combat disinformation.

Urman and Makhortykh (2023) underscore the need for more robust and detailed transparency in the way online platforms handle content, arguing that current reports often fail to provide the public with a clear understanding of platform practices and policies. This analysis is crucial in the context of the ongoing debate about disinformation and the role of tech companies in content moderation. The importance of algorithmic transparency was highlighted at the ECAT Research Workshop 2024 in Brussels, where experts from academia, policy, and industry discussed critical issues such as algorithmic auditing, automated content moderation, online gender-based violence, and recommender systems (ECAT, 2024).

In addition, the spread of false information through fake profiles must be addressed by identifying and removing such accounts using network analysis, the examination of behavioural and interaction patterns, as well as image verification.

Educating users about the risks of generative AI and increasing public understanding of recommendation algorithms are essential for enabling more responsible navigation of these platforms. Raising awareness and fostering critical thinking and fact-checking skills are vital steps toward reducing the spread of disinformation. A 2023 study among students in Serbia revealed that most had read little or nothing about the algorithms social media companies use to detect false information on their platforms, thus underscoring the importance of incorporating these topics into university courses (Kovačević & Demić, 2023).

Collaboration between researchers, industry, and regulators is critical to ensure that risks are detected, discussed, and mitigated. Legislative measures such as the Digital Services Act (DSA), which establishes legal frameworks to regulate very large online platforms and search engines, and the AI Act (AIA), which mandates risk assessments prior to the launch of new systems, further highlight the need for accountability. Finally, as Pronin and Kugler (2007) argue, it is vital to teach people to critically evaluate both news sources and their own thought processes.

6. Conclusion

Generative AI is reshaping society and is increasingly being applied across various industries. However, if the influx of AI-generated content into publicly accessible data goes unchecked, it may impede effective information retrieval and distort our collective understanding of socio-political realities and scientific consensus. The majority of instances of generative AI misuse are not sophisticated attacks on AI systems themselves. They merely exploit openly accessible features (e.g. automated disinformation, fake content creation) which require no specialised technical expertise. Social media platforms further amplify these challenges by playing a significant role in spreading disinformation.

As generative AI becomes more deeply rooted in daily life and business, the fight against disinformation must be both rigorous and forward-looking. By investing in AI detection tools, encouraging collaboration among tech firms, policymakers, and civil society, and increasing public awareness, the risks posed to both individuals and society at large may be mitigated.

Large Language Models (LLMs) generally excel at detecting disinformation, but they often struggle when confronting content created by other LLMs. While community-based fact-checking approaches, such as Community Notes on the X platform, have shown limited effectiveness in curbing engagement with misleading content, experts emphasise the importance of transparency and algorithmic accountability. Along with technical measures like improved training data, watermarking, and user-access restrictions, fostering critical thinking and educating the public about recommendation algorithms are also key steps for curbing the spread of disinformation. Such endeavours underscore the need for interdisciplinary collaboration, legislative frameworks such as the DSA and AI Act, and a proactive stance from tech companies in managing online content.

Acknowledgement: The work was supported by the Science Fund of the Republic of Serbia within the framework of the 'IDEAS' Program – Management of New Security Risks Research and Simulation Development, NEWSiMR&D, #7749151

References

[1] AIA. (2025, February 7). *AI Act | Shaping Europe's digital future*. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

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

[3] Alvarez, G. (n.d.). *Gartner Top 10 Strategic Technology Trends for 2025*. Gartner. Retrieved 15 January 2025, from https://www.gartner.com/en/articles/top-technology-trends-2025

[4] Bateman, J., & Jackson, D. (n.d.). *Countering Disinformation Effectively: An Evidence-Based Policy Guide*. Carnegie Endowment for International Peace. Retrieved 1 February 2025, from https://carnegieendowment.org/research/2024/01/countering-disinformation-effectively-an-evidence-based-policy-guide?lang=en

[5] Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, *7*(33), eabe5641. https://doi.org/10.1126/sciadv.abe5641

[6] Brewster, T. (n.d.). *Musk's X Fired 80% Of Engineers Working On Trust And Safety, Australian Government Says*. Forbes. Retrieved 1 February 2025, from https://www.forbes.com/sites/thomasbrewster/2024/01/10/elon-musk-fired-80-per-cent-of-twitter-x-engineers-working-on-trust-and-safety/

[7] Chuai, Y., Tian, H., Pröllochs, N., & Lenzini, G. (2024). Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter? *Proc. ACM Hum.-Comput. Interact.*, *8*(CSCW2), 428:1-428:52. https://doi.org/10.1145/3686967

[8] Datareportal. (n.d.). *Global Social Media Statistics*. DataReportal – Global Digital Insights. Retrieved 30 January 2025, from https://datareportal.com/social-media-users

[9] DSA. (n.d.). *DSA: Very Large Online Platforms and Search Engines* [Text]. European Commission - European Commission. Retrieved 9 February 2025, from https://ec.europa.eu/commission/presscorner/detail/en/ip_23_2413

[10] ECAT. (n.d.). *ECAT Research Workshop: Shaping the Future of Algorithmic Transparency - European Commission*. Retrieved 31 January 2025, from https://algorithmic-transparency.ec.europa.eu/news/ecat-research-workshop-shaping-future-algorithmic-transparency-2024-11-28_en

[11] Eigner, E., & Händler, T. (2024). *Determinants of LLM-assisted Decision-Making* (arXiv:2402.17385). arXiv. https://doi.org/10.48550/arXiv.2402.17385

[12] Gartner, (2024). *2025 Gartner CIO and Technology Executive survey*. Gartner, Inc.

[13] Gillespie, T. (2022). Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society*, *8*(3), 20563051221117552. https://doi.org/10.1177/20563051221117552

[14] Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations* (arXiv:2301.04246). arXiv. https://doi.org/10.48550/arXiv.2301.04246

[15] Helming, C., & Marsh, O. (n.d.). *10 Questions about AI and elections*. AlgorithmWatch. Retrieved 11 December 2024, from https://algorithmwatch.org/en/10-questions-ai-elections/

[16] Jordan, J. J., & Rand, D. G. (2020). Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *Journal of Personality and Social Psychology*, *118*(1), 57.

[17] Kaiser, B., & Mayer, J. (2023). *It's the Algorithm: A large-scale comparative field study of misinformation interventions Emilie Flamme*. https://knightcolumbia.org/content/its-the-algorithm-a-large-scale-comparative-field-study-of-misinformation-interventions

[18] Kaplan, J. (2025, January 7). More Speech and Fewer Mistakes. *Meta*. https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

[19] Kovačević, A., & Demić, E. (2023). Artificial intelligence for detecting fake news on social media-attitude survey. *Medjunarodni problemi*, *75*(4), 685-710.

[20] Kojima, T., Gu, S. (Shane), Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems*, *35*, 22199–22213.

[21] Marchal, N., Xu, R., Elasmar, R., Gabriel, I., Goldberg, B., & Isaac, W. (2024). *Generative AI Misuse: A Taxonomy of Tactics and Insights from Real-World Data* (arXiv:2406.13843). arXiv. https://doi.org/10.48550/arXiv.2406.13843

[22] McLoughlin, K. L., Brady, W. J., Goolsbee, A., Kaiser, B., Klonick, K., & Crockett, M. J. (2024). Misinformation exploits outrage to spread online. *Science*, *386*(6725), 991–996. https://doi.org/10.1126/science.adl2829

[23] Narayanan, A. (2023). *Understanding Social Media Recommendation Algorithms*. https://doi.org/10.7916/khdk-m460

[24] Nyhan, B., & Reifler, J. (2010). When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior*, *32*(2), 303–330. https://doi.org/10.1007/s11109-010-9112-2

[25] Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. (2021). Partisan Polarization Is the Primary Psychological Motivation behind Political Fake News Sharing on Twitter. *American Political Science Review*, *115*(3), 999–1015. https://doi.org/10.1017/S0003055421000290

[26] Oremus, W., Alcantara, C., Merrill, J. B., & Galocha, A. (n.d.). *How Facebook shapes your feed*. Washington Post. Retrieved 10 August 2023, from https://www.washingtonpost.com/technology/interactive/2021/how-facebook-algorithm-works/

[27] Pronin, E., & Kugler, M. B. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *Journal of Experimental Social Psychology*, *43*(4), 565–578. https://doi.org/10.1016/j.jesp.2006.05.011

[28] Rathje, S., Roozenbeek, J., Van Bavel, J. J., & Van Der Linden, S. (2023). Accuracy and social motivations shape judgements of (mis) information. *Nature Human Behaviour*, *7*(6), 892–903.

[29] Salerno, J. M., & Peter-Hagene, L. C. (2013). The Interactive Effect of Anger and Disgust on Moral Outrage and Judgments. *Psychological Science*, *24*(10), 2069–2078. https://doi.org/10.1177/0956797613486988

[30] Sallami, D., Chang, Y.-C., & Aïmeur, E. (2024). *From Deception to Detection: The Dual Roles of Large Language Models in Fake News* (arXiv:2409.17416). arXiv. https://doi.org/10.48550/arXiv.2409.17416

[31] Schneier, B., & Sanders, N. (2024, December 2). *The apocalypse that wasn't: AI was everywhere in 2024's elections, but deepfakes and misinformation were only part of the picture*. The Conversation. http://theconversation.com/the-apocalypse-that-wasnt-ai-was-everywhere-in-2024s-elections-but-deepfakes-and-misinformation-were-only-part-of-the-picture-244225

[32] Schwalbe, M. C., Joseff, K., Woolley, S., & Cohen, G. L. (2024). When politics trumps truth: Political concordance versus veracity as a determinant of believing, sharing, and recalling the news. *Journal of Experimental Psychology: General*, *153*(10), 2524–2551. https://doi.org/10.1037/xge0001650

[33] Urman, A., & Makhortykh, M. (2023). How transparent are transparency reports? Comparative analysis of transparency reporting across online platforms. *Telecommunications Policy*, *47*(3), 102477. https://doi.org/10.1016/j.telpol.2022.102477

Dezinformacije i generativna veštačka inteligencija: rizici, izazovi i moguća rešenja

Ana Kovačević

Sažetak

Fenomen dezinformacija prisutan je od samih početaka ljudske komunikacije, pri čemu ponovljena izloženost lažnim informacijama često dovodi do njihovog prihvatanja kao istine. U savremenom društvu, međutim, dezinformacije postaju sve zabrinjavajuće zbog jednostavnog, jeftinog i uverljivog načina na koji se mogu kreirati koristeći generativnu veštačku inteligenciju, naročito velike jezičke modele. Generartna veštčka inteligencija, koji i dalje prave značajne napretke, omogućavaju generisanje velikih količina sadržaja koji izgledaju verodostojno, čak i na nacionalnom jeziku. Pored toga, ovi modeli olakšavaju kreiranje personalizovanog sadržaja prilagođenog određenim grupama ili pojedincima, navodeći prividno pouzdane izvore, čime dodatno jačaju uticaj dezinformacija. U svetu preplavljenom informacijama, ovaj problem postaje još izraženiji. Jedan od ključnih kanala za širenje dezinformacija su socijalni mediji, čiji je primarni cilj da privuku i zadrže pažnju korisnika sadržajem koji potvrđuje njihova postojeća uverenja. U radu je predstvljen potencijal upotrebe generativne veštačke inteligencije u kreiranju dezinformacija, kao i mogućnosti za suzbijanje ovog problema.

Ključne reči: generativna veštačka inteligencija, dezinformacije, rizici veštačke inteligencije, veliki jezički modeli, socijalni mediji.