Adapt Institute

AI-Generated Disinformation: Understanding and Fighting New Phenomenon

Jerguš Lajoš

Adapt Long Read

NATO OTAN This publication is supported by NATO's Public Diplomacy Division

The author is solely responsible for the content of this document. Opinions expressed in this publication do not purport to reflect the opinions or views of Adapt Institute or donors of this publication.

Author

Jerguš Lajoš
BA student student of Security and Strategic Studies at Masaryk University in Brno
Adapt Institute Junior Research Fellow

Editor

Matúš Jevčák
Editor-in-Chief at Adapt Institute

Adapt Long Read - Analytical Paper
©Adapt Institute
August 2023

AI-GENERATED DISINFORMATION: UNDERSTANDING AND FIGHTING NEW PHENOMENON

Jerguš Lajoš

SUMMARY AND RECOMMENDATIONS

- Disinformation has been a longstanding issue in information warfare, but the development of artificial intelligence (AI) has fundamentally transformed the landscape. It has made the creation, amplification, and dissemination of disinformation possible in more efficient and widespread ways.
- Disinformation campaigns powered by AI operate on an unprecedented scale, speed, and level of sophistication. They generate highly convincing false information and target specific users with tailored content, undermining democratic processes, dividing society, and eroding institutional trust.
- Conventional methods of detecting and countering disinformation are increasingly inadequate when faced with the volume and complexity of AI-generated disinformation. This necessitates continuous innovation and partnerships to confront this evolving threat effectively.
- AI-generated disinformation encompasses various types, including automated content generation, deepfakes, and manipulation of social media algorithms. Each presents unique challenges for detection and mitigation.
- The implications of AI-generated disinformation are extensive, impacting political manipulation and influence, eroding trust and social cohesion, and potentially leading to economic and security ramifications.
- The development of sophisticated AI detection methods, collaborative efforts among researchers, platforms, and policymakers, and the establishment of legal and regulatory frameworks that promote accountability, transparency, and responsible AI use are all essential components for formulating effective strategies to counter AI-generated disinformation.

INTRODUCTION

The swift advancement of artificial intelligence (AI) technology has led to a concerning development: the emergence of AI-generated disinformation. This phenomenon poses serious threats to the accuracy of information, public perception, and societal well-being. The ability of AI algorithms to create, amplify, and propagate disinformation on an unprecedented scale and with remarkable sophistication is blurring the line between authentic and fabricated content. This paper underscores the urgency and significance of addressing the challenges posed by AI-generated disinformation. It also delves into the various implications, mechanisms, and potential defences against the evolution of this threat in the digital age.

Perhaps the first well-known instance of AI manipulating public opinion was during the 2016 US presidential election. This event provided tangible evidence of how digital transformation is impacting politics and democracy. The utilisation of algorithms, automation, and AI heightened the efficacy and reach of disinformation campaigns and associated cyber activities, influencing the opinions and voting decisions of American citizens.

As advancements in AI technology progressed, other tactics to confuse or genuinely mislead consumers of media content began to gain popularity. A pivotal year in AI development was 2018, when OpenAI introduced the GPT model, forming the foundation for the ChatGPT chatbot. This innovation opened access to an essentially boundless amount of text created by AI. Subsequently, additional forms of AI-generated content emerged. In 2021, the introduction of deepfake videos, such as the one featuring Tom Cruise, showcased an elaborate manipulation of video using a generative adversarial network (GAN). This network consists of a generator crafting images from random noise and a discriminator assessing the authenticity of an image produced by the generator. While initially challenging, producing such video content became more feasible in recent years, enabling ordinary users to employ these techniques. AI has also been employed to craft deepfake images and manipulate audio, including voice cloning.

Over the past couple of years, notable strides have been made in AI-based technologies and systems. Since its inception, ChatGPT has undergone multiple updates, enhancing its technical capabilities and information database.

Developers aim to facilitate societal functioning, yet these technologies can also yield significant problems when misused with malicious intent.

BACKGROUND

In the realm of information warfare and propaganda, the issue of disinformation has long been recognised. Nevertheless, the advent of artificial intelligence (AI) technologies has significantly transformed the landscape of disinformation. AI-driven disinformation campaigns operate on an unparalleled scale, speed, and complexity compared to traditional methods. These campaigns have the capability to generate incredibly convincing fabricated media like images, videos, and articles. They can also tailor content to specific individuals and exploit weaknesses in social media algorithms designed for content targeting (Hurst 2023).

As noted by Himelein-Wachowiak and her colleagues (2021), a pivotal role is played by AI-powered bots, automated accounts that amplify mis/disinformation. These bots distribute false content to wide audiences, manipulate social media platforms to boost engagement, participate in coordinated campaigns to enhance virality, drown out legitimate voices, and even manipulate trending topics and search results, all contributing to the spread of false narratives. The coordinated, speedy, and large-scale nature of bots makes it exceedingly challenging to halt the dissemination of false information online.

Furthermore, both conventional disinformation and that which is generated by AI have the potential to undermine democratic processes, disrupt social cohesion, and erode trust in institutions. AI can impact election outcomes, exacerbate societal divisions, and foster an atmosphere of distrust and uncertainty by influencing public opinion through targeted disinformation campaigns.

The complexities presented by this issue make AI-generated disinformation a significant challenge for societies worldwide. Detecting and effectively countering this evolving threat is becoming increasingly challenging for humans alone. The volume and intricacy of AI-generated disinformation may outpace traditional fact-checking and debunking methods. To proficiently identify, counteract, and mitigate the detrimental effects of AI-generated disinformation, staying ahead of its ever-evolving nature is essential. This entails exploring innovative techniques, advancements, and collaborations to effectively detect and thwart AI-powered disinformation campaigns (Bond 2023).

To develop a comprehensive grasp of this phenomenon, it is imperative to define and categorise AI-generated disinformation as false or misleading information intentionally created or amplified using AI technologies.

Several types of AI-generated disinformation have emerged in recent years. A prominent example is automated content generation, where AI algorithms churn out vast amounts of fabricated news articles, social media posts, and comments. These algorithms can replicate human-like language patterns and writing styles with the aid of natural language processing (NLP) techniques. Techniques like generative adversarial networks (GANs) and recurrent neural networks (RNNs) have been developed to enhance the quality and believability of AI-generated content (DemDigest 2018; Nasir et al. 2021). Detecting and countering automated content generation poses a considerable challenge due to the rapid pace and sheer volume at which it is produced.

Another form is the utilisation of deepfakes and synthetic media. Deepfakes involve manipulated videos or audio recordings using AI algorithms to substitute the appearance or voice of individuals with someone else's, creating highly convincing yet fabricated content (Rossler et al. 2019). Deepfake creation involves a GAN system where two AI programs collaborate. It's essentially a method to generate new types of data from existing datasets algorithmically. For example, it could analyse numerous images of Donald Trump and then create entirely new images similar to but not identical replicas of the original images (Walorska 2020). AI techniques like deep learning and computer vision enable the development of deep fakes that are challenging to differentiate from authentic videos or audio recordings. This carries the potential to spread false narratives, tarnish reputations, and manipulate public perception, posing a significant challenge to media authenticity and trust (Ram et al. 2022).

AI-generated disinformation campaigns often manipulate social media algorithms to amplify their reach and influence. AI algorithms can exploit recommendation systems and targeting features on platforms like Facebook, Twitter, and YouTube. By strategically crafting and disseminating disinformation that aligns with user preferences, these algorithms can facilitate the viral spread of false information (Bovet and Makse 2019). Additionally, these algorithms can take advantage of network effects and echo chambers within social media platforms to target specific communities or demographics, enhancing the effectiveness of disinformation campaigns. This manipulation of social media algorithms raises concerns about filter bubbles, information silos, and the potential polarisation of public discourse (Guess et al. 2020).

MANIPULATION AND INFLUENCE

AI-generated disinformation holds significant implications for political manipulation and influence. Malicious actors can exploit AI tools to propagate false narratives, shape public opinion, and sway electoral proceedings. Through the use of automated content generation, deepfakes, and targeted dissemination methods, disinformation campaigns can manipulate perceptions of candidates, political parties, and critical topics. This manipulation undermines the democratic process, distorts public discourse, and corrodes trust in political institutions (Whyte 2002). A notable concern here is the vast potential of content generation, a capability that surpasses the imaginative and creative confines of human beings, who are influenced by factors like gender, age, education, and social background. Additionally, the speed of content creation by machines is unparalleled.

The proliferation of AI-generated false information also contributes to the breakdown of social cohesion and trust. These disinformation campaigns foster polarisation within communities, exploit societal divisions, and fuel discord. Purposeful dissemination of fabricated information erodes confidence in authorities, institutions, and traditional media outlets. Exposure to AI-generated disinformation can lead individuals to become more sceptical and dubious of encountered information, which may result in a broader erosion of trust in the information ecosystem (Pennycook et al. 2020).

Furthermore, it's imperative to acknowledge the gravity of the threat posed by AI-produced disinformation to the economy and security. Firstly, its impact on businesses, stock markets, and consumer behaviour can be profound. False information about businesses, products, or financial markets can severely disrupt operations, leading to instability and financial losses. Disinformation campaigns targeting specific industries or sectors can also cultivate uncertainty, damaging investor confidence and hindering economic growth (Funso 2023). From a security perspective, AI-generated disinformation can be weaponised to advance hostile agendas. Nation-states or non-state entities can employ such campaigns to sow discord, incite violence, or undermine trust in security institutions. Disinformation can also serve as a tool for psychological warfare, influencing military manoeuvres or destabilising regions. The potential security implications of AI-generated disinformation present challenges in maintaining stability, peace, and international relations (Sedova et al. 2021).

Comprehending the impact of AI-generated disinformation on political manipulation, erosion of trust, and economic and security considerations is vital for formulating effective strategies to tackle this issue. By acknowledging the multifaceted repercussions, policymakers, researchers, and technology developers can collaborate to mitigate the harm caused by this form of disinformation, safeguarding the integrity of public discourse, democratic processes, and overall societal well-being.

COUNTERMEASURES

Legal and Regulatory Approaches

Effective combatting of AI-generated disinformation hinges on robust legal and regulatory frameworks. Governments and policymakers must evaluate and revise existing laws to tackle the challenges presented by this form of disinformation. This entails regulations that ensure transparency in AI technologies, safeguard data privacy, and establish accountability for platform providers. The development of ethical guidelines and standards for AI applications can also promote the responsible and ethical utilisation of AI technologies in content creation and dissemination. Recent initiatives, such as the European Commission's call for platforms like Google, Facebook, YouTube, and TikTok to label AI-generated content, seek to combat mis-/disinformation. However, it's noteworthy that this is a voluntary code, lacking mandatory compliance and penalties (Birchard 2023). The impending Artificial Intelligence Act, proposed in 2021, should serve as the official regulatory instrument, with lawmakers now hastening the legislative process due to evolving circumstances. However, concerns arise over the potential misuse of the legal framework for purposes that curtail personal freedoms, possibly being masked as protective measures (Chan 2023).

China, too, has employed regulatory measures requiring the registration of AI technologies that serve as a public service. Beijing supports and encourages AI technology's progress, especially in the industrial sector, positioning itself as a global leader in this realm, in direct competition with the US (He 2023).

Furthermore, international cooperation is crucial to address cross-border disinformation campaigns exploiting AI technologies. Legal and regulatory frameworks provide essential guidelines for deterring and penalising malicious actors involved in AI-generated disinformation while upholding freedom of expression and democratic values (Kertysova 2018).

Detection and Collaboration

Advanced AI detection techniques, combined with collaborative efforts among researchers, platforms, and policymakers, are imperative to counter this phenomenon effectively. These strategies work in tandem to establish a comprehensive response, preserving the integrity of information ecosystems and fostering informed public discourse.

The field of AI for disinformation detection is ever-evolving, with future research aimed at enhancing detection techniques. Exploring deep learning architectures, like transformer models, can enhance the accuracy and efficiency of identifying AI-generated disinformation. Furthermore, integrating multimodal approaches that encompass text, image, and audio analysis can bolster the ability to spot sophisticated disinformation campaigns that utilise diverse media types (Nguyen et al. 2022). Real-time detection systems

and AI tool integration within social media platforms can also bolster proactive identification and mitigation of AI-generated disinformation (Ojha 2023). Sustained research and innovation in this field are crucial to outpace emerging disinformation techniques and ensure the efficacy of countermeasures.

Psychological Effects and Platform Accountability

Future research should delve into the psychological ramifications of AI-generated disinformation on individuals and society. Understanding how people perceive and respond to such disinformation offers insights for effective interventions. Research can explore cognitive biases, emotional manipulation's impact, and factors influencing susceptibility to disinformation (Zhou et al. 2023). Examining AI-generated disinformation's role in shaping public opinion, attitudes, and behaviours can guide strategies to promote media literacy, critical thinking, and resilience to disinformation. Scrutinising the psychological dimensions will contribute to comprehensive approaches addressing both technical and human aspects of the issue.

Research should also focus on the responsibilities and accountability of AI platforms and providers in combating such disinformation. This encompasses ethical considerations in developing AI tools for potential disinformation use.

Frameworks for responsible AI development, algorithmic transparency, and mechanisms for auditing and regulating AI platforms warrant exploration. Additionally, strategies to enhance platform policies, empower users, and ensure effective content moderation can curb disinformation spread. Understanding AI platforms' role in the disinformation ecosystem informs guidelines and best practices for responsible AI use and governance (Jobin et al., 2019; Juršenas et al. 2022).

CONCLUSION

This text has delved into the intricate realm of AI-generated disinformation, encompassing its context, significance, definitions, types, causes, effects, and mitigation methods. The analysis underscores that AI-generated disinformation poses substantial threats to economic stability, social unity, and democratic processes. Addressing this challenge mandates advanced AI detection techniques, collaborative endeavours among stakeholders, and robust legal and regulatory frameworks.

The impact of AI-generated disinformation spans policymakers, researchers, and the wider public. To uphold democratic processes, safeguard citizens, and promote responsible AI use, policymakers must acknowledge the urgency and devise comprehensive strategies. Legislative and regulatory structures that endorse transparency, accountability, and ethical application of AI technologies must be established. Researchers hold a pivotal role in shaping detection methods, comprehending psychological implications, and investigating ethical considerations. Their findings contribute to the development of effective interventions and defences. In fostering resilience against disinformation, society at large requires initiatives like media literacy programs, education in critical thinking, and public awareness campaigns.

Collaboration serves as a catalyst for interdisciplinary research, global partnerships, and the exchange of knowledge, information, and best practices. By amalgamating our expertise, resources, and perspectives, we can construct a formidable defence against AI-generated disinformation, safeguard the integrity of information ecosystems, and fortify our societal resilience.

REFERENCES

Birchard, Rosie. 2023. “AI content: EU asks Big Tech to tackle disinformation.” DW.com, May 6. Accessed July 17, 2023. https://www.dw.com/en/ai-content-eu-asks-big-tech-to-tackle-disinformation/a-65830533.

Bond, Shannon. 2023. “AI-generated deepfakes are moving fast. Policymakers can´t keep up.” *NPR.org*, April 27. Accessed May 31, 2023. https://www.npr.org/2023/04/27/1172387911/how-can-people-spot-fake-images-created-by-artificial-intelligence.

Bovet, Alexander, Herán A. Makse. 2019. “Influence of fake news in Twitter during the 2016 US presidential election.” *Nature Communications*, no. 7 (January). Accessed June 3, 2023. https://www.nature.com/articles/s41467-018-07761-2.

Chan, Kelvin. 2023. “How Europe is leading the world in the push to regulate AI.” Apnews.com, June 14. Accessed July 17, 2023. https://apnews.com/article/ai-act-artificial-intelligence-europe-regulation-94e2b38703b38fdbfabc9580f845ef9a

DemDigest. 2018. “Generative adversarial networks: how fake news fuels authoritarians.” March 30. Accessed May 31, 2023. https://www.demdigest.org/fake-news-fuels-authoritarian-resurgence/.

Funso, Richard. 2023. “The Emerging Threat of AI-Powered Disinformation and Business Impact.” March 31. Accessed June 4, 2023. https://www.linkedin.com/pulse/emerging-threat-ai-powered-disinformation-business-funso/.

Guess, Andrew et al. 2019. “Less than you think: Prevalence and predictors of fake news dissemination on Facebook.” *Science advances,* vol. 5 (January). Accessed June 4, 2023. https://pubmed.ncbi.nlm.nih.gov/30662946/.

He, Laura, 2023. “China takes major step in regulating generative AI services like ChatGPT.” CNN Business, July 14. Accessed July 17, 2023. https://edition.cnn.com/2023/07/14/tech/china-ai-regulation-intl-hnk/index.html.

Jiawei Zhou et al. 2023. “Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions.” Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). (April). Accessed June 4, 2023. https://jiaweizhou.me/assets/chi23_ai_misinfo.pdf

Jobin, Anna et al. 2019. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence*, vol. 1 (September). Accessed June 4, 2023. https://www.nature.com/articles/s42256-019-0088-2#citeas.

Ojha, Rahul. "AI Tools Combat Fake News." April 9. Accessed June 4. https://www.linkedin.com/pulse/ai-tools-combat-fake-news-rahul-ojha/.

Himelein-Wachowiak, McKenzie et al. 2021. "Bots and Misinformation Spread on Social Media: Implications for COVID-19." *Journal of Medical Internet Research*, vol. 23 (May). Accessed July 17, 2023. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8139392/.

Hurst, Luke. 2023. "Rapid growth of 'news' sites using AI tools like ChatGPT is driving the spread of misinformation." May 9. Accessed May 31, 2023. https://www.euronews.com/next/2023/05/02/rapid-growth-of-news-sites-using-ai-tools-like-chatgpt-is-driving-the-spread-of-misinforma.

Juršenas, Alfonsas. 2022. "The Role of AI in the Battle Against Disinformation." NATO Strategic Communications Centre of Excellence, February 28. Accessed July 17, 2023. https://stratcomcoe.org/publications/the-role-of-ai-in-the-battle-against-disinformation/238.

Kertysova, Katarina. 2018. "Artificial Intelligence and Disinformation." *Security and Human Rights*, vol. 29 (December). Accessed June 4, 2023. https://www.researchgate.net/publication/338042476_Artificial_Intelligence_and_Disinformation.

Nasir, Jamal Abdul. 2021. "Fake news detection: A hybrid CNN-RNN based deep learning approach." *International Journal of Information Management Data Insights*, vol. 1 (April). Accessed May 31, 2023. https://www.sciencedirect.com/science/article/pii/S2667096820300070.

Nguyen, Thanh Thi et al. 2022. "Deep Learning for Deepfakes Creation and Detection: A Survey." Computer Vision and Image Understanding, vol 223 (October). Accessed June 4, 2023. https://www.sciencedirect.com/science/article/abs/pii/S1077314222001114.

Pennycook, Gordon et al. 2020. “The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings.” Management Science. Vol. 66 (February). Accessed June 3, 2023. https://pubsonline.informs.org/doi/10.1287/mnsc.2019.3478.

Ram, Saravana et al. 2022. “Deep Fake Detection Using Computer Vision-Based Deep Neural Network with Pairwise Learning.” Intelligent Automation and Soft Computing, vol. 35 (August). Accessed June 1, 2023. https://www.researchgate.net/publication/362723448_Deep_Fake_Detection_Using_Computer_Vision-Based_Deep_Neural_Network_with_Pairwise_Learning.

Rossler, Andreas, et al. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. Cornell University. Accessed June 2, 2023. https://arxiv.org/pdf/1901.08971.pdf.

Sedova, Katerina. 2021. “AI and the Future of Disinformation Campaigns Part 1: The RICHDATA Framework.” CSET. (December). Accessed June 4, 2023. https://cset.georgetown.edu/wp-content/uploads/CSET-AI-and-the-Future-of-Disinformation-Campaigns.pdf.

Walorska, Agnieszka. 2020. Deepfakes & Disinformation. Potsdam: Friedrich Naumann Foundation for Freedom. Accessed May 31, 2023. https://shop.freiheit.org/#!/Publikation/897.

Whyte, Christopher. 2020. “Deepfake news: AI-enabled disinformation as a multi-level public policy challenge.” Journal of Cyber Policy, vol. 5 (August). Accessed June 3, 2023. https://www.tandfonline.com/doi/citedby/10.1080/23738871.2020.1797135?scroll=top&needAccess=true&role=tab&aria-labelledby=cit.

Adapt
Institute

Na vŕšku 8
811 01 Bratislava
Slovak Republic

office@adaptinstitute.org
+421 908 327 491
www.adaptinstitute.org