Anastasia Tsertsvadze, Revaz Topuria Free University of Tbilisi
atser21@freeuni.edu.ge | r.topuria@freeuni.edu.ge

AI AND INFORMATION WARS: ARTIFICIAL INTELLIGENCE IN THE PRODUCTION OF PROPAGANDA AND DISINFORMATION

Abstract

Because of factors such as creating personalized disinformation and emotional manipulation, artificial intelligence is becoming a powerful tool in the construction of the information environment. Political narratives are also targeted and changed. Automated bots, generative models, deepfakes, and other AI technologies modify how the public views information and evoke feelings - whether positive or negative. Studies on the topic become more significant given the current international relations climate, as information control is a form of power on par with military and economic resources. The illegal dissemination of deepfake videos during the 2022 war in Ukraine, the 2016 U.S. elections, and the COVID-19 “infodemic” illustrate the urgency of AI technologies in information and the persuasive destabilization of democracy. AI technologies accelerate information sprawl and disinformation’s militarization.

The information war as a component of hybrid tactics in international relations is not a new phenomenon. The scale of propaganda and disinformation dissemination, however, is unprecedented due to the rapid advancement of AI technologies.

The use of information as a weapon of various states and non-state actors has been ongoing for centuries, but the emergence and development of artificial intelligence has created new opportunities and threats. While earlier information wars relied largely on classical media and traditional channels of political propaganda, today digital platforms, social media, and automated systems occupy the central role of mass communication. This paper discusses precisely the mechanisms through which AI systems create and disseminate false narratives, influence elections, and change the dynamics of democratic systems. Although artificial intelligence is also being used to combat disinformation, the technology that currently exists is not yet capable of maximizing its effectiveness. This is precisely why analyzing those mechanisms and approaches that are using the capabilities of AI in information wars and those that are trying to utilize AI itself to combat the spread of propaganda and disinformation is so important.

Introduction

To examine how artificial intelligence is utilized in the dissemination of disinformation and propaganda and whether there is a trend of using AI-generated content, we will use a multidisciplinary approach incorporating content and case analysis. It includes the analysis of texts, images, videos and user responses. The particular focus is given to content associated with information spread through automated bots, videos produced using deepfake technology and AI-generated narratives. The paper undertakes an in-depth analysis of a number of

Anastasia Tsertsvadze, Revaz Topuria Free University of Tbilisi
atser21@freeuni.edu.ge I r.topuria@freeuni.edu.ge

cases. Each case study outlines how artificial intelligence was enabled to produce and disseminate disinformation, what technological mechanisms were activated, and what consequences that process generated for global security.

An additional source of support to content and case analysis is primary and secondary data analysis. Data from a variety of international research organizations, media and academic articles are used (BBC Monitoring, RAND Corporation, Graphika, Stanford Internet Observatory). Also, visual material analyses, which relates to the careful examination of photos and videos that are produced by AI, has been previously noted. This involves analyzing the technological properties of deepfake videos, the level of realism of the image, and the implications for society's emotional experience. Visual analysis is critical to demonstrate how physical visual manipulations impact perceptions of truthfulness of information.

Key words: Artificial Intelligence, Disinformation, deepfake, Information Warfare, Propaganda, Cybersecurity, Social Media.

Artificial Intelligence and Social Media

In the information age, artificial intelligence has drastically enhanced the scale and accuracy of disinformation dissemination. Social media remains the primary venue, wherein the functionality of bots and algorithms created the perception of general consensus. For instance, during the 2016 U.S. presidential elections, thousands of automated accounts were capable of penetrating political discourse and influencing consumers (Bradshaw & Howard 2019). NLP models, through the text generation process, mimicked credible sources in such a way that it became challenging for the reader to differentiate between state veracity and state falsehood (Buchanan, Lohn, Musser & Sedova 2021).

Above all, one of the most powerful examples of untrue content disseminated through artificial intelligence is the video that circulated on YouTube in March 2025 involving American evangelist Kenneth Copeland's death presented to the world. The twenty-minute video segment was produced entirely by AI-generated voice and visual media and gained traction across social networks until it was discovered that Kenneth Copeland was still alive and had not been arrested (Bano, Baig & Abrejo 2025).

During the COVID-19 pandemic, botnets disseminated anti-vaccine messages and conspiracy theories across all forms of social media, for example, claiming the virus was a biological weapon or blamed 5G technology for the sickness. Bots often commented on popular videos to try to reinforce the idea that the opinions were broadly shared. The material was often altered to reach the audience: for the youth, the emphasis was on vaccines influencing fertility; for the religious, it was opposition to faith; for parents, it was danger to children. This was significantly damaging trust in official medical institutions (Bano, Baig & Abrejo 2025).

Deepfake technology further augmented the persuasiveness of disinformation. In 2022, at the very outset of the war in Ukraine, Russia aired on TV channels a video where Volodymyr Zelensky was calling upon the population to surrender. The video was dropped quickly but in the first few minutes of airing, confusion and fear were issued to the public. Similar technologies were used in other cases. In November 2024, videos were distributed on social media where an exact AI clone of Sir David Borrows' voice was pro-Russian. People frequently relied on the credibility and emotionality of the voice, which assisted with making the false news credible. In March 2025, a video circulated on TikTok, in which Donald Trump was changing the name of the District of Columbia. The voice was entirely synthetic and emulated Trump's authentic tone and diction. In this regard, TikTok became an exceptionally powerful platform. Its recommendation algorithms are tailored

Anastasia Tsertsvadze, Revaz Topuria Free University of Tbilisi
atser21@freeuni.edu.ge | r.topuria@freeuni.edu.ge

to the preferences and behavior of the viewer which is what allowed radical and pseudo-patriotic content to spread rapidly (Mazur 2022).

Technologies and mechanisms related to artificial intelligence

Artificial intelligence has become a tool in the digital environment that simultaneously creates content, directs its deployment, and influences an emotional response. Generative models - language (GPT, LLaMA, Claude) and visual (Stable Diffusion, DALL·E, Midjourney) - have produced texts, photographs and videos of such quality that it is difficult to distinguish the artificial from the genuine (Floridi & Chiriatti 2020). To these technologies we can add bot-nets - thousands of automated profiles that spread one narrative, create the illusion of „public opinion", and incite conflict of opinion. During the U.S. election of 2020, that was exactly the type of content overflowing with conspiracy theories and arguments that was spread by algorithms throughout the online space, while counter opinions and counterarguments where lost in the commentary threads (Bradshaw & Howard 2019). We see this operate similarly in recommendation systems - TikTok, YouTube, and Facebook preferentially promote emotionally charged, polarized content, while reinforcing „bubbles" that affirm the users preconceived beliefs. Scripts written for language models depict a new version of phishing and social engineering. The texts are produced in natural language, written in the appropriate voice, and adapted to local contexts and institutional details (Bano, Baig & Abrejo 2025). The effect is that the writer seems credible. Cultural codes come into play - the historical traumas, identity, political confrontations - and mission is already accomplished: creating uncertainty, instilling distrust and amplifying social polarization (Gallotti 2020).

The next step of AI's machinations is „synthetic identities" - AI Persona Farming. This is where thousands of personas are created with their „biographies," histories and habits. These personas don't just share links! They react, argue, „express emotion," and shove the audience in one direction or another as well. These networks are capable of spreading a single narrative in multiple languages, at the same time. A post might appear in one location, a comment elsewhere, and a visual in yet another location, at which point the fake narrative is already established. This exact phenomenon has been recorded in various countries from 2023. Hundreds of „new users" engage around the same topic at the same time, utilizing local slang and effectively mimicking the tone of local media (Mazur 2022).

The management of this influence is also facilitated by RAM (Recommendation Algorithm Manipulation). The goal of the RAM operation is to utilize algorithms to allow for a singular narrative to organically graft. RAM relies on something called narrative seeding, whereby thousands of identical or similar pieces of content are posted from numerous profiles. These postings and videos might not feature blatantly false truths; instead they are often authored in a humorous, low-quality or ironic manner. All of them nevertheless essentially convey the same message. This gives the impression that what is being perceived is „going viral": the user watches dozens of similar videos and finds similar opinions in the comments (Riccardo 2020). Thus, the impression is created that this story resembles the real sentiment of wide swathes of society. RAM is particularly dangerous in that it is operating in an informational „grey zone", where truth, irony and doubt swap around. Therefore, detection and control by platforms becomes fundamentally more complicated.

Artificial Intelligence and Cyber Security

AI and Cyber Security In a digital environment, where disinformation moves at a velocity and volume beyond human capacity, AI serves the primary "sensor" and "filter" functions. In real time, algorithms can detect patterns of user behavior and content dynamics and identify patterns that would otherwise be abnormal – a sudden increase in new profiles, synchronized posting on a common theme, and sameness in linguistic templates (Buchanan, Lohn, Musser, & Sedova 2021). Similarly, in a security platform, machine models embedded in SIEM screens collect signals extracted from a variety of sources (logs, network, or devices), assign each event a threat score, and will decide on a response if needed. At the textual level, NLP is deployed to detect aggressive and polarized rhetoric, organize messages, and to signal where dissemination starts and where it is amplified (Gallotti 2020). In viewing the text, transformer models are used to recognize semantics, emotional tone, and factual consistency; in visual and audio they utilize computer vision and voice control to detect small "slips", signs of deepfakes, unnaturalness, micro-expressions, and timbre incompatibility. Automated fact-checking is employed to cross-reference news with trusted databases and flag inconsistencies and deviations from standard dialogue (Riccardo 2020).

The auto-flagging system identifies signals and templates, develops a risk score, and limits the dissemination. When the risk score reaches or exceeds a threshold – and if that threshold has triggered a response – the post may either merely fail to publish, or be protected from being disseminated. When a risk score exceeds a given threshold, the system will come into play: the post may move into an additional vetting stage, or undergo pre-publish moderation, or algorithmically moderate spread (shadow demote), or be filtered entirely from the platform, that is, deleted. In the trends block ("emerging narratives") the system detects in advance what will trend virally and develops steps to attenuate the amplification of the umbrellas; though cultural diversity and linguistic subtleties remain vulnerabilities here (Floridi & Chiriatti 2020).

Ultimately, then, begins the structural dismantling of the information operation within the denaturing social media platforms. Network disruption algorithms find the "support," the disruption of which slows the entire critical framing of disinformation. Forced isolation, temporary disabling of hashtags, termination of spam channels, robbed of momentum and visibility, may ultimately compromise the narrative itself (Buchanan 2021).

Conclusion

Modern AI systems currently have the potential to generate static posts, but also video, audio recordings, and mixed narratives that are designed to persuade and target a particular subset or demographic group. Platforms are increasingly leveraging AI for defensive purposes, but these technologies remain under development and will need to be checked and regularly updated. Artificial intelligence in the dissemination of disinformation and propaganda is not just a tool, it is already a systemic weapon changing the information space and influencing democratic and societal processes. Trends on the use of AI-generated content suggest that disinformation campaigns will become increasingly sophisticated and explicitly adapted to increase challenges to international and domestically stability. In light of this, we need constant surveillance, the introduction of rigorous ethical frameworks and regulatory frameworks to ensure the use of AI remains constant with justice and information security.

Anastasia Tsertsvadze, Revaz Topuria Free University of Tbilisi
atser21@freeuni.edu.ge I r.topuria@freeuni.edu.ge

List of Literature

Bano, S., Baig, A., & Abrejo, S. (2025, May 5). Combating Digital Misinformation and Deepfakes Using Artificial Intelligence: Analyzing the Role of AI in Detection, Content Moderation, and Public Trust in the Era of Information Disorder. Retrieved from http://www.amresearchreview.com/index.php/Journal/article/view/111

Bradshaw, S., & Howard, P. N. (2019). The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation. Oxford: University of Oxford. Retrieved from https://demtech.oii.ox.ac.uk: https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2019/09/CyberTroop-Report19.pdf

Buchanan, B., Lohn, A., Musser, M., & Sedova, K. (2021). Truth, Lies, and Automation How Language Models Could Change Disinformation X. Center for Security and Emerging Technologies. Retrieved from cset.georgetown.edu: https://cset.georgetown.edu/wp-content/uploads/CSET-Truth-Lies-and-Automation.pdf

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. 30.

Retrieved from springer.com: https://link.springer.com/article/10.1007/s11023-020-09548-1

Gallotti, R., Valle, F., Castaldo, N., Sacco, P., & De Domenico, M. (2020). Assessing the Risks of ‚Infodemics' in Response to COVID-19 Epidemics. 4. Nature Human Behaviour. Retrieved from www.nature.com: https://www.nature.com/articles/s41562-020-00994-6#citeas

Mazur, P. (2022). How Russia Used TikTok and Telegram to Spread War Propaganda. The New York Times. Retrieved from https://www.nytimes.com/2022/12/15/technology/russia-state-tv-ukraine-war.html

Riccardo, G., Valle, F., Castaldo, N., Sacco, P., & De Domenico, M. (2020, October 29). Assessing the risks of 'infodemics' in response to COVID-19 epidemics. Nat Hum Behav (4), 1285–1293. Retrieved from https://www.nature.com/articles/s41562-020-00994-6