

Propaganda for Algorithms – New Targets of Information Manipulation in the AI Era

Author: Davit Kutidze¹

Editor: Irina Gurgenchashvili²

Generative artificial intelligence and Deepfake technologies (the use of AI and machine-learning algorithms to create false photographs, audio, and videos that closely resemble reality) create an information environment in which distinguishing fact from falsehood has become significantly more difficult.³ In addition, there is another, no less significant threat. As is well known, the main objective of manipulative information campaigns run by actors with differing political or financial interests is to influence people's consciousness. However, in the era of artificial intelligence, the new target of information influence operations is no longer solely human beings. What matters now is not only what an individual thinks or feels about a particular story, but also what Google's search algorithm, the recommendation systems of various social networks, and, above all, generative artificial intelligence "see." If traditional propaganda sought to change human beliefs or manipulate human emotions, the new type of influence operations seeks to alter what a machine "learns," "repeats," or chooses to cite, something that will subsequently be automatically reflected in the answers it provides to millions of people. Thus, today, humans are no longer the sole discoverers of information on the internet, people who read, compare, and assess sources themselves. Algorithms are also discoverers of information: the ranking mechanism of a search engine, the recommendation system of social media, and the generative AI model that responds to questions posed to it in the form of a unified, confident text, often without even citing sources.

According to the classic definition by Woolley and Howard (2016), "computational propaganda", which describes precisely the manipulation of algorithms, is "the assemblage of social media platforms, autonomous agents, and big data tasked with the manipulation of public opinion".⁴ This definition rests on the assumption that the ultimate target remains human beings, even where the intermediary is a bot or an algorithm. Bots imitate human behaviour in order to create an impression of artificial

¹ Researcher, Gnomon Wise; e-mail: d.kutidze@ug.edu.ge

² Executive Director, Research Institute Gnomon Wise, e-mail: i.gurgenchashvili@ug.edu.ge

³ Kutidze, D. (30 April 2026). Fake News in the Age of Artificial Intelligence: Emerging Threats and Strategies for Mitigation. Gnomon Wise Research Institute. Via link: <https://gnomonwise.org/ge/publications/analytics/315>

⁴ Samuel C. Woolley and Philip N. Howard. (2016). Automation, Algorithms, and Politics| Political Communication, Computational Propaganda, and Autonomous Agents. *International Journal of Communication*. Via link: https://www.researchgate.net/publication/402435515_Automation_Algorithms_and_Politics_Political_Communication_Computational_Propaganda_and_Autonomous_Agents_-_Introduction

consensus for real people. However, the literature of recent years shows a gradual complication of this model: the algorithm is no longer merely a channel through which a message is directed at people, but itself becomes an intermediary “audience,” whose “persuasion” may be more effective than influencing millions of individuals. If a search engine or chatbot once “learns” a certain narrative, or perceives a particular website as an authoritative source, this influence is subsequently reproduced automatically, without additional cost, for every following user, language, and country. Most importantly, if someone could theoretically “persuade” algorithms and artificial intelligence in favour of their interests, they would rewrite history according to their own interpretation.

Attempts to “persuade” algorithms and artificial intelligence are already observable and documented in a number of studies. This process is facilitated by an important characteristic of the information space - the so-called “data voids.” This term was proposed by Golebiewski and Boyd in 2019.⁵ According to their definition, data voids, or information vacuums, emerge when very little material can be found online for rare search terms. Manipulators driven by their ideological, economic, or political interests make effective use of this. The authors identify several types of void: 1. voids created by breaking news - the creation of manipulative information around search terms for which demand suddenly rises in the context of a current, high-profile event. Eventually, these voids are filled with high-quality informational content; however, until such content is created, they are exploited; 2. new strategic terms - manipulators invent new terms and create an information network around them in advance. They then disseminate these terms among the public, often through media outlets, in order to provide audiences only with misleading information and narratives favourable to themselves; 3. outdated terms - when information about an event becomes outdated, content creators, such as reliable media outlets, stop producing material related to it much earlier than users stop searching for those terms. Manipulators exploit this situation: since credible media outlets no longer write anything new on an outdated topic, they themselves produce information favourable to them. Google’s and other search engines’ algorithms often prioritise newly published material. As a result, search engines display recently published disinformation content produced by manipulators in the top results. 4. fragmented concepts - manipulators separate interrelated ideas and create different information groups from them, so-called “bubbles,” adapted to different political frameworks. In this way, manipulators conceal the full picture from audiences and remove facts from their context, preventing people from seeing the logical chain between events. Ultimately, this contributes to social polarisation; 5. problematic search queries - acute and sensitive keywords closely connected to social

⁵ Golebiewski, M. and Boyd, D. (2019). Data Voids: Where Missing Data Can Easily Be Exploited. *Data & Society*. Via link: <https://apo.org.au/sites/default/files/resource-files/2019-10/apo-nid265631.pdf>

disagreements, pressing social issues, conspiracy theories, or historical trauma. If, in the past, mainly disinformation or radical viewpoints could be found for a particular term, the search engine continues to display the same material by inertia because counterbalancing credible sources simply do not exist.

It should be noted that the concept of data voids was initially developed for the analysis of search engines, but recent studies show that the same logic is applicable to large language models as well. When there are few credible, high-quality sources on a particular topic while low-quality or propagandistic material is available in abundance, a model, like a human being, has nothing other than what is available.

Attempts to “Poison” Generative Artificial Intelligence: The Russian Pravda Network

Research organisations studying Russian disinformation unanimously note that the so-called Pravda network, officially named Portal Kombat, which consists of hundreds of websites, has been disseminating pro-Russian narratives for years. Following Russia’s full-scale invasion of Ukraine in 2022, the network expanded sharply and, according to the Atlantic Council’s Digital Forensic Research Lab (DFRLab), by 2025 covered more than 80 regions and countries, while its content relied largely on machine translation.⁶

Of particular note is how researchers describe the websites comprising this network. According to Jankowicz and Newport (2025),⁷ Pravda sites lack a search function and a main website menu. Stylistic and linguistic flaws are also noticeable, making them less tailored to human users. [A Georgian-language Pravda website also exists and is marked by similar shortcomings.] In addition, these sites record few visits by real users. All these indicators suggest that the Pravda network is currently not intended for humans. The authors conclude that the real targets are automated agents, search-engine indexers, and data-collection bots that feed the training datasets of large language models (LLMs). A quantitative assessment of the Pravda network’s scale is one of the clearest arguments demonstrating its automated nature. According to a NewsGuard study,⁸ the network published 3.6 million articles/news items in 2024 alone, which to some extent became integrated into Western AI systems and “infected” their answers with manipulated or propagandistic information. At the same time,

⁶ Châtelet, V and Lesplingart, A. (2025). Russia-linked Pravda network cited on Wikipedia, LLMs, and X. *The Digital Forensic Research Lab (DFRLab)*. Via link: https://dfrlab.org/2025/03/12/pravda-network-wikipedia-llm-x/?utm_source=chatgpt.com

⁷ Newport, A and Jankowicz, N. (2025). Russian networks flood the Internet with propaganda, aiming to corrupt AI chatbots. *Bulletin of the Atomic Scientists*. Via link: <https://thebulletin.org/2025/03/russian-networks-flood-the-internet-with-propaganda-aiming-to-corrupt-ai-chatbots/>

⁸ Sadeghi, M and Blachez, I. (2025). The Infection of Western AI Chatbots by a Russian Propaganda Network. *NewsGuard*. Via link: <https://bit.ly/3UAeeg8>

according to research by the international analytical centre GLOBSEC,⁹ as of March 2025, more than 85% of articles on Pravda's main website were published less than one minute apart from one another. This practically rules out a human editorial process and indicates full automation. It is also important that the network produces little original content. It simply copies news from sources such as TASS and RT, Russian propaganda agencies, as well as lesser-known aggregators. Their value is determined not by the size of their audience but by the volume of content they produce. According to GLOBSEC, this suggests that its creators prioritise quantity over quality. Ultimately, different groups of independent researchers conclude that the goal of this network is to "feed" algorithms with desired information and, in recent years, to "poison" generative AI systems. According to Olga Tokariuk, Senior Analyst at the Institute for Strategic Dialogue, the network is particularly focused on discrediting Ukraine's government, disseminating pro-Russian narratives, and creating an illusion of "alternative opinion". In Tokariuk's words, ultimately, "the network is designed to flood the internet with content favourable to propaganda, so that the content then enters search engines, legitimate media outlets, and even AI chatbots".¹⁰

The process described above represents a classic example of AI Grooming. This term has become established since last year following a report by the American Sunlight Project.¹¹ It describes a phenomenon in which individual actors deliberately disseminate large volumes of false or biased content in order for it, over time, to be registered in the training data of large language models (LLMs) and collectively presented as a credible source. The term importantly distinguishes between two heterogeneous but interconnected practices. On the one hand, there is the use of artificial intelligence to produce mass disinformation, AI as a tool. On the other hand, there is the "internal" contamination of language models themselves for the purpose of manipulating their responses. Didier Danet (2025)¹² examines this concept in detail at the academic level. According to Danet, LLM Grooming operates according to the following logic: it "floods" publicly accessible sources, including web archives, with large volumes of fabricated narratives so that the model statistically absorbs them through training and subsequently transforms them into factual knowledge. This is precisely why Danet considers basic

⁹ Kubś, J. (2025). Global Offensive: Mapping the Sources Behind the Pravda Network. *GLOBSEC*. Via link: <https://www.globsec.org/sites/default/files/2025-05/Pravda%20Network%20Report.pdf>

¹⁰ Quoted in Sergei Stetsenko (18 April 2025), "[The Fake Russian Pravda That Seeks to Poison Social Media and Artificial Intelligence](#)." Radio Free Europe/Radio Liberty.

¹¹ American Sunlight Project. (2025). *A Pro-Russia Content Network Foreshadows the Automated Future of Info Ops*. Via link: <https://static1.squarespace.com/static/6612cbdf9a9ce56ef931004/t/67fd396818196f3d1666bc23/1744648558879/PK+Report.pdf>

¹² Didier Danet. LLM Grooming: A New Cognitive Threat to Generative AI. 2025. (hal-05241525). Via link: <https://hal.science/hal-05241525v1/file/20251215%20DANET%20Cognitive%20Threats%20LLM%20Grooming%20V02.pdf>

counter-disinformation measures, which are reactive and focused on individual false publications, inadequate for combating systemic contamination. Such contamination requires proactive strategies, including audits of databases, continuous monitoring, human oversight mechanisms, and international cooperation.

Empirical evidence that content from the Pravda network has already reached generative AI systems emerges from several independent sources. A joint study by DFRLab and Finland-based CheckFirst, using Wikipedia's API, identified 1,907 hyperlinks pointing to 162 Pravda-network domains in 1,672 articles published in 44 languages.¹³ The spread of these links was especially noticeable in Russian-language Wikipedia (with 922 links), and Ukrainian-language Wikipedia (with 580 links). However, since 2022, the number of links in the English-language version has also increased substantially, reaching 133. This is problematic because Wikipedia is one of the main sources used in training large language models. The study's authors also tested this connection directly. Specifically, when testing ChatGPT, Copilot, Perplexity, and Gemini, they found that the models generated content published on Pravda websites. None of them, however, warned the user of the sources' Russian origin, despite publicly available research on the network already existing.¹⁴

One of the most striking pieces of evidence of contamination of a model's training data emerged from a 2026 DFRLab study, which analysed the public web archive Common Crawl, a source that feeds the training datasets of many AI models. The study found that the number of archived pages containing English-language Pravda content rose from 37 in November 2024 to approximately 40,000 by November 2025. To test whether contamination was real, DFRLab used Meta's open model Llama 3.1 405B Base and a text-completion method: the model was given the opening sentence of a known article and researchers examined whether it reconstructed the remaining text verbatim. The test showed that the model reproduced almost word for word an RT-distributed false narrative about US "biolaboratories" in Ukraine, suggesting that this specific text had been included in the model's training data.¹⁵

These findings are consistent with a joint 2025 study by Anthropic, the UK AI Security Institute, and the Alan Turing Institute. According to that study, adding as few as 250 malicious documents to training

¹³ Châtelet, V and Lesplingart, A. (2025). Russia-linked Pravda network cited on Wikipedia, LLMs, and X. *The Digital Forensic Research Lab (DFRLab)*. Via link: https://dfrlab.org/2025/03/12/pravda-network-wikipedia-llm-x/?utm_source=chatgpt.com

¹⁴ Ibid.

¹⁵ Digital Forensic Research Lab (DFRLab). (2026). Pravda in the pipeline: Early evidence of state-adjacent propaganda in AI training data. Via link: <https://dfrlab.org/2026/04/08/pravda-in-the-pipeline/>

data is sufficient to compromise the responses of even a very large model of more than 13 billion parameters.¹⁶ Once this harmful content is integrated into a model's systems, the only solution is complete retraining of the model, a highly labour-intensive and costly process.

It must also be underlined that interpretations of the research presented above are not universally agreed upon. In particular, a study published in the Harvard Kennedy School Misinformation Review in 2025 (Alyukov et al.),¹⁷ tested this assumption using four popular chatbots (ChatGPT-4o, Gemini 2.5 Flash, Copilot, and Grok-2) and concluded that there was little evidence supporting the theory of LLM Grooming. The authors argue that when a model draws on sources such as Pravda, this may result not from deliberate "poisoning" but from ordinary data voids, topics for which there are few reliable sources and low-quality content dominates. In their conclusion, the root of the problem may not be foreign manipulation but rather the unequal distribution of quality information on the internet. From a practical perspective, this does not reduce the seriousness of the problem, but it requires a different response: not only blocking the "enemy," but also filling the deficit of reliable content. This debate is important from a methodological perspective as well. As Golebiewski and Boyd (2019) noted,¹⁸ data voids are difficult to identify and often remain unnoticed until an event causes a sharp increase in demand for information on a specific topic. Accordingly, the two explanations, intentional "grooming" or structural voids, are not mutually exclusive. Rather, they are likely to be mechanisms that operate in parallel: mass content production simultaneously fills a void and attempts to fill it with a targeted narrative. Thus, the threat nevertheless persists because mass-produced, inexpensive, and often automatically generated content quickly fills the vacuum in the information environment, while high-quality, trustworthy journalism or research is slow and costly to produce.

The development of generative artificial intelligence takes information influence operations to an entirely new level. Until now, the ultimate target of traditional propaganda was human consciousness. In the contemporary information environment, however, the algorithm itself, the system that gathers,

¹⁶ Anthropic, the UK AI Security Institute, and the Alan Turing Institute. (2025). *A small number of samples can poison LLMs of any size*. Via link: <https://www.anthropic.com/research/small-samples-poison>

¹⁷ Alyukov, M. Makhortykh, M. Voronovici, A. and Sydorova, M. (2025). LLMs grooming or data voids? LLM-powered chatbot references to Kremlin disinformation reflect information gaps, not manipulation. *Harvard Kennedy School Misinformation Review*. Via link: <https://misinforeview.hks.harvard.edu/article/llms-grooming-or-data-voids-llm-powered-chatbot-references-to-kremlin-disinformation-reflect-information-gaps-not-manipulation/>

¹⁸ Golebiewski, M. and Boyd, D. (2019). Data Voids: Where Missing Data Can Easily Be Exploited. *Data & Society*. Via link: <https://apo.org.au/sites/default/files/resource-files/2019-10/apo-nid265631.pdf>

categorises, learns, and subsequently delivers information to millions of users, is increasingly becoming an intermediary target of influence. Therefore, the effectiveness of propaganda may be determined not only by how many people a particular message reaches, but also by whether it succeeds in influencing how systems that mediate information perceive reality. In other words, such information operations are oriented toward statistical dominance in search or training corpora. Their aim is not to persuade a particular person at a given moment, but to ensure that, in the future, when someone asks a question, the system immediately offers the manipulator's preferred answer as the "most available" information.

The approach described in the article has several practical consequences. First, there is scale and resource efficiency. While a troll or bot farm needed thousands of accounts to persuade people, the mass generation of content, through the endless reproduction of a text written once, or generated automatically, attempts to achieve the same result much more cheaply and in less time. The scale of the Russian Pravda disinformation network is proof of this: millions of articles per year despite minimal organic audiences. Moreover, a narrative that enters AI training data may remain "embedded" in a model's behaviour for months or years, with a much more enduring effect than a one-off social-media campaign. There is also the issue of transparency. In social media, researchers can, to some extent, trace networks of accounts. The process through which an AI response is formed, namely, how much weight each source carries in a particular answer, is often entirely opaque even to the model's developer. Ultimately, in the AI era, the problem of information manipulation extends beyond the familiar, though often unresolved, problem of directly influencing human beings and presents entirely new challenges.