Медіафорум : аналітика, прогнози,
інформаційний менеджмент.
Том 16. 48-58

Mediaforum: Analytics, Forecasts,
Information Management.
Volume 16. 48-58

https://doi.org/10.31861/mediaforum.2025.16.48-58
УДК: 004.8 (327:355)

© Vasyl Baranets[1]
©Ivan Osadtsa[2]

GENERATIVE AI TECHNOLOGIES AS A TOOL FOR COUNTER-NARRATIVES TO RUSSIAN PROPAGANDA

This article explores the potential of generative artificial intelligence (AI) technologies as a tool for constructing counter-narratives against russian propaganda in the context of the full-scale war in Ukraine. It examines methodological approaches to AI in the field of international communication, identifies key characteristics of the information warfare environment – such as deepfakes, large language model (LLM) poisoning, and automated disinformation – and focuses on how these technologies are exploited by hostile actors. Special attention is given to Ukrainian governmental and civil society initiatives that utilize generative AI to produce fact-based, emotionally resonant, and multimodal counter-narratives. The article discusses the ethical and legal boundaries of such use, including the risks of eroding public trust, the opacity of algorithmic outputs, and the challenge of distinguishing legitimate information defense from manipulation. Finally, it outlines future research perspectives regarding regulatory frameworks, strategic communication, and the development of algorithmic trust in democratic societies. The analysis is based on 20 domestic and international sources.

Keywords: *generative AI, Ukraine, EU, political communication, international relations, disinformation, russian propaganda, counter-narrative, LLM, deepfake.*

[1] Аспірант кафедри кафедри міжнародних відносин та суспільних комунікацій Чернівецького національного університету імені Юрія Федьковича, Україна, E-mail: baranets.vasyl@chnu.edu.ua; https://orcid.org/0009-0003-8096-5821.

[2] Кандидат політичних наук, доцент кафедри міжнародних відносин та суспільних комунікацій Чернівецького національного університету імені Юрія Федьковича, Україна, E-mail: i.osadtsa@chnu.edu.ua; https://orcid.org/0000-0001-5593-5944.

Генеративні ШІ-технології як інструмент контрнаративів російській пропаганді

У статті досліджено можливості використання генеративних технологій штучного інтелекту як інструменту контрнаративної боротьби з російською пропагандою в умовах повномасштабної війни. Проаналізовано методологічні підходи до вивчення ШІ у сфері міжнародних комунікацій, визначено характерні риси інформаційної війни як середовища для експлуатації генеративного ШІ. Особливу увагу приділено практикам українських державних та громадських ініціатив, які використовують генеративні ШІ для створення фактологічних, емоційно релевантних та мультимодальних контрнаративів. У роботі висвітлено етичні та правові обмеження такого використання, зокрема ризики втрати довіри, проблеми прозорості алгоритмів і межі легітимної інформаційної оборони. Запропоновано перспективи подальших досліджень у сфері регулювання, стратегічної комунікації та розвитку алгоритмічного довір'я в демократичних суспільствах.

Ключові слова: *генеративний ШІ, Україна, ЄС, політична комунікація, міжнародні відносини, дезінформація, російська пропаганда, контрнаратив, LLM, deepfake.*

Statement of the Research Problem. Following the onset of russia's full-scale invasion of Ukraine in 2022, the world has faced not only military aggression but also an unprecedented wave of information attacks actively involving cutting-edge technologies. The russian federation has systematically employed both traditional propaganda tools and digital innovations – including generative artificial intelligence (AI) algorithms – for the creation of false narratives, disinformation, emotional manipulation, and the erosion of Ukraine's legitimacy on the international stage (Majchrzak 2023; Goldstein 2023; Menn 2025).

This situation has clearly created a pressing need for the development of responses capable not only of neutralizing information attacks, but also of generating authentic counter-narratives grounded in truth, factual accuracy, and emotional resonance. In this context, generative AI systems – such as large language models (LLMs), image generators, and synthetic video tools – are increasingly seen not merely as a threat, but as a potential instrument of active resistance (Kuznetsova 2023; Ezzeddine 2022; Marushchak 2025). Accordingly, the study of generative AI technologies as a means of counter-narrative policy against russian propaganda in the digital environment has become especially relevant.

Review of recent research. The studies considered in this article can be broadly divided into two categories: threat analytics – works that focus on the risks associated with the use of generative AI for aggressive purposes (particularly by russia); and resistance analytics – research exploring how these same technologies can be employed to generate counter-narratives, enable fact-checking, and strengthen public democracy.

A number of publications emphasize that generative AI is increasingly used by state and quasi-state actors as a component of hybrid warfare. This is particularly evident in russia's campaigns against Ukraine and the broader Western world. For instance, Majchrzak (Авдєєва 2024) describes the mechanisms for creating fake videos using deepfake technologies to fabricate quotations attributed to "Western experts". This is corroborated by a report from the Washington Post, which details how russian bots manipulated chatbots and LLMs through misleading queries in order to generate kremlin-aligned responses (Menn 2022). Goldstein et al. (Goldstein, 2023) and Sprenkamp et al. (18) likewise stress that LLMs are not only passive retransmitters of falsehoods but may be actively deployed in automated influence operations (AIOps), particularly in efforts to manipulate public opinion across platforms where the mass publication of comments and news content is critical to shaping perceptions of "reality".

The second category of literature focuses on the potential of generative AI in building counter-narratives. The study by Kuznetsova et al. (Kuznetsova 2023) investigates how effectively LLMs can serve as political information verifiers and concludes that, when adapted to specific local contexts, their use in this role is indeed feasible.

Ukrainian media and analytical platforms have become active contributors in documenting and analyzing the use of generative AI. Ornatskyi (Орнатський 2024) describes the functioning of the War of Words project – Ukraine's generative tool for identifying russian propaganda texts through linguistic analysis. Avdiieva (Авдєєва 2024) illustrates how russian deepfake videos are widely disseminated on TikTok and Telegram while also offering guidance on how such content can be detected. An analysis by Texty.org.ua (Литвинов 2024) models the responses of generative AI to queries about the war, comparing outputs of models trained on Western corpora to those from Ukrainian-language systems. This comparison highlights the pressing need for locally adapted AI models.

Both Western and Ukrainian scholarship acknowledge a key concern: the risk of symmetrical use of AI. If Ukraine also begins to produce emotionally charged but weakly verified counter-narratives, this may erode its moral advantage over the aggressor (Sadeghi 2025; Brandt 2023; MediaMaker 2025).

Presentation of Core Material. With the onset of the full-scale invasion of Ukraine by the russian Federation, the information war has reached an unprecedented scale and level of complexity. Traditional instruments of influence – including manipulations in televised news, the dissemination of false statements by official representatives, the fabrication of pseudo-analytical materials, and the creation of networks of fake accounts – have been supplemented by high-tech disinformation mechanisms. A particularly dangerous development has been the emergence of generative artificial intelligence models, which are now used not only to distribute false content, but also to produce it in an automated manner, without the involvement of significant human resources. This has introduced a new pace and scope to information operations, making them more adaptive and personalized (Majchrzak 2023; Goldstein 2023; Укрінформ 2024).

The russian propaganda system has adapted to these new technologies with striking speed. As early as 2023, researchers documented instances of generative neural networks being used to create visual simulations of so-called "Western experts" commenting on events in Ukraine. These fictional figures, generated with the help of deepfake algorithms, imitated speech, facial expressions, and the communication style of real professionals, spreading kremlin-favorable interpretations of events. Particularly troubling is the fact that such videos were created without the involvement of any real individuals, which precludes accountability and makes their verification difficult – even for experienced users (Majchrzak 2023). In this way, generative AI has evolved from a mechanism for disseminating falsehoods into a full-fledged producer of fake content capable of simulating credibility and authenticity.

One of the most insidious innovations in this domain is the "seeding" of large language models such as ChatGPT, Claude, or LLaMA with toxic content. This tactic, known as prompt injection, involves the intentional input of instructions or data into a model to induce it to generate hostile, distorted, or manipulative responses. According to an investigation by the Washington Post, in 2025, specially trained chatbots affiliated with russian structures systematically "rewrote" historical facts, downplayed russia's responsibility for genocidal acts against Ukrainians, undermined the very notion of Ukrainian statehood, and propagated distorted narratives about international support for Ukraine (Menn 2025). This form of influence is particularly deceptive, as a user engaging with such a chatbot may be unable to distinguish an objective answer from a poisoned narrative embedded in linguistically polished yet ideologically hostile text.

Simultaneously, russia is actively deploying methods for the automated production and dissemination of disinformation on social media. Generative language models are used to mass-produce texts, comments, posts, and reactions on platforms such as Facebook, Telegram, TikTok, and X (Twitter). The goal of such campaigns is to fabricate an illusion of widespread support for pro-russian views, flood the information space with false arguments, simulate artificial consensus, and suppress authentic messages originating from Ukrainian sources (MediaMaker 2025). This method – an information simulation of public opinion – allows the adversary to create the appearance of social legitimacy for its position while simultaneously discrediting the Ukrainian side's official communication (Wack 2025). Another widely used tool involves the fabrication of fake documents – an application of generative AI to the textual formats of official communication. These may include forged letters between Ukrainian officials, "leaked" classified materials, or supposedly intercepted messages that are circulated in mass media, blogs, and Telegram channels. Such documents are often crafted with a high level of stylistic authenticity, including appropriate linguistic turns of phrase, grammar, and even digital signatures, making verification difficult. Their appearance is frequently accompanied by a targeted information attack, involving distribution through dozens of anonymous pages, dissemination via pseudo-analytical Telegram channels, and amplification by bot networks. The primary objective of such fabrications is to undermine public trust in Ukrainian authorities, provoke internal conflicts, discredit officials, and create a climate of pervasive mistrust toward all information (Укрінформ 2024; Гембік 2025).

Generative AI often operates as a "black box" – the final output may appear reliable, but the user has no means to trace how the system arrived at a given text or image (Кабінет Міністрів України 2025; Сидорський 2023). This is particularly relevant in wartime conditions, where elevated emotional tension and widespread information fatigue make audiences more vulnerable to convincing yet false messages. As research shows, algorithmic content generation based on user prompts – especially on platforms such as TikTok, Telegram, and X/Twitter – enables manipulators to construct highly targeted information traps (Marushchak 2025; РБК Україна 2024).

It is equally important to address another dimension – the use of the same technologies as instruments of counter-offensive action, that is, for constructing narratives of resistance, exposing disinformation, and reinforcing Ukraine's agency in the context of information warfare. In classical terms, a counter-narrative represents an alternative semantic framework that refutes or reinterprets the dominant narrative of the adversary. In wartime, such counter-narratives are not limited to debunking fakes; they aim to restore truth and strengthen the moral position of the side resisting aggression (MediaMaker 2025; Сидорський 2023). Generative AI models, including GPT, Claude, Gemini, or LLaMA, can process vast arrays of historical, documentary, and media sources, transforming them into new forms – essays, explainer videos, infographics, visual memes, or verified news stories. In this context, they function not only as information sources but also as mechanisms for converting truth into politically effective communication (Kuznetsova 2023; Ezzeddine 2022; Marushchak 2025).

The War of Words project, launched by Ukrainian journalists and fact-checkers, is based on textual style analysis algorithms and allows for the identification of the "russian footprint" in seemingly neutral posts on social media or news sites (Majchrzak 2023; Авдєєва 2024). In such cases, generative models are used not only to classify hostile content but also to generate automatic rebuttals in the form of brief messages tailored to specific audiences. Another example is the use of generative AI to create videos and graphics that debunk falsehoods about Ukrainian soldiers, volunteers, or diplomats. Some civil society initiatives produce series of visual content based on real stories but in formats that can rival the persuasive power of disinformation imagery (Укрінформ 2024; Сидорський 2023; MediaMaker 2025). LLM-based modules are also embedded in Telegram bots, enabling users to automatically check suspicious claims – for example, by detecting disinformation circulating in group chats or comment threads. One such tool, built on open-source models, uses comparative analysis with official sources and reports from international media (Sprenkamp 2023; Орнатський 2024; Кабінет Міністрів України 2025).

Unlike traditional media, generative models can operate in an audience-adaptive mode. Counter-narratives produced with their assistance can be multimodal (combining text, visuals, and voice), localized (adapted to specific regional or linguistic groups), and personalized (adjusted to the user's level of knowledge). This enables significantly more effective message delivery compared to universal official communications. Such an approach is especially important in situations where russian propaganda seeks to go beyond Ukraine and influence audiences in the EU, Africa, and Asia by portraying Ukraine as a "Western puppet" or a source of global instability (Goldstein 2023; MediaMaker 2025; Brandt 2023).

However, this strategy is not without its limitations. First, not all generative models are reliable – at times, they may "hallucinate", producing inaccurate or distorted representations of reality (Sadeghi, 2025). Second, if the mechanisms behind counter-narrative generation are not disclosed to the public, this may provoke backlash, including accusations of symmetrical propaganda (РБК Україна 2024; Marushchak 2025). These risks are particularly relevant in democratic societies, where institutional trust and transparency in communication methods are of critical importance. Therefore, the next section will focus on the Ukrainian context – that is, how exactly these approaches are implemented in the practices of state and civil society institutions.

The use of generative artificial intelligence in Ukraine during wartime has extended far beyond the realm of technical tools – it has become an integral part of strategic communication, public diplomacy, and national security. In 2024, Ukraine joined the Council of Europe's Framework Convention on Artificial Intelligence, Human Rights, Democracy, and the Rule of Law, thereby recognizing the necessity of ethical AI regulation even in wartime conditions (Кабінет Міністрів України 2025). Certain ministries, notably the Ministry of Digital Transformation, have actively integrated LLM components into public communication, ranging from automated responses in government service platforms to the generation of internal analytical materials. Ukraine's Ministry of Foreign Affairs, in turn, uses generative AI to analyze russian media campaigns abroad and formulate timely counter-arguments in response (РБК Україна 2024). These systems, for instance, make it possible to automatically detect new narratives in foreign publications and select appropriate diplomatic reactions based on tone and local context.

One of the most influential civil society initiatives in this field is War of Words – an online platform for identifying russian disinformation campaigns based on rhetorical, visual, and stylistic pattern analysis (Majchrzak 2023; Авдєєва 2024). The project combines machine learning, narrative classification, and generative AI to produce short refutations that can be conveniently distributed via social media. Other NGOs, including Internews Ukraine, Texty.org.ua, and Zmina, use generative models to develop adapted explanations of complex events – delivered, for example, through explanatory infographics, dialogue-based chatbots, or visual Instagram stories (Укрінформ 2024; Литвинов 2024). The use of AI in these projects enables the simultaneous delivery of information that is accurate, emotionally resonant, and accessible to wide audiences. Notably, there are also projects focused on developing open LLMs trained on Ukrainian data sources that reflect the national language, cultural context, and political realities. These models are better suited to recognizing russian propaganda markers that often remain undetected by English-language AIs (Орнатський 2024; Гембік 2025).

A distinctive feature of the Ukrainian case is the active cooperation between state institutions and independent civil initiatives. For instance, findings from projects such as Detector Media, Babel, and Texty.org.ua are periodically incorporated into official government communications and, in some cases, into international information campaigns (Авдєєва 2024; Укрінформ 2024; Литвинов 2024). There have also been documented cases of generative AI–based civil society analytical modules being used in parliamentary hearings or as source material for international briefings. This demonstrates a high level of flexibility and adaptability in Ukraine's information strategy under hybrid warfare conditions. The Ukrainian experience illustrates that even in exceptionally challenging circumstances, it is possible not only to defend against hostile narratives but also to shape an active information policy based on precision, responsiveness, and citizen engagement.

However, the application of generative artificial intelligence in the information war raises not only strategic questions but also profound ethical and legal dilemmas. While generative AI can serve as a powerful instrument for defending truth, it also entails risks – ranging from factual inaccuracies to potential misuse. One of the most delicate challenges is the risk of "symmetrical accusations": if one side uses AI for information defense, the adversary may accuse it of engaging in propaganda – even if the content is defensive or grounded in verifiable facts (Brandt 2023; MediaMaker 2025). In Ukraine's case, this issue is particularly acute, as russia systematically attempts to delegitimize any Ukrainian communication by portraying it as "fake," "engineered by Western intelligence," or part of a "NATO psychological operation" (Goldstein 2023; Menn 2025). This demands not only tactical literacy but also a strategic approach to building an ethical framework for information sovereignty.

As previously noted, in March 2024, the Council of Europe adopted the Framework Convention on Artificial Intelligence, which mandates the principles of transparency, non-discrimination, and the protection of human rights in the context of algorithmic governance (Кабінет Міністрів України 2025). For Ukraine, participation in this initiative is not only legally significant but also symbolically important – it demonstrates that even during wartime, the country remains committed to the core values of European democracy. Simultaneously, international research institutions – such as RAND Corporation, Stiftung Neue Verantwortung, and the Brookings Institution – have proposed expanding the concept of digital security to include the ethics of generative AI, emphasizing the need for global oversight in the use of such technologies in public communication (Ezzeddine 2022; MediaMaker 2025; Brandt 2023).

Conclusions. Thus, generative AI technologies are gradually evolving from tools of limited technical application into one of the key elements of the modern information ecosystem. In the context of the full-scale russian-Ukrainian war, these technologies are no longer a neutral backdrop or the exclusive domain of IT specialists – they have become an active component of both aggressive information strategies and defensive mechanisms designed to preserve the democratic nature of communication. When generative AI is viewed not as an abstract technical innovation but as a political instrument that simultaneously shapes, transmits, and transforms meaning, it becomes clear that it has acquired strategic importance in the realm of global information confrontation.

Faced with constant threat, limited resources, a dynamic political environment, and the necessity of addressing both domestic and international audiences, Ukraine is developing a unique case that combines technological adaptability with ethical responsibility. While many governments have not yet implemented generative models in their public services due to regulatory and reputational concerns, Ukraine – out of necessity – has already begun integrating these tools into real-world governmental and civil communication. This situation requires further analysis, the development of policy recommendations, and potentially the creation of a specific regulatory framework for states that are compelled to use AI technologies under conditions of armed conflict.

References:

1. Avdieieva T (2024). Yak Rosiia vykorystovuie ShI dlia stvorennia feikiv, ta yak yikh rozpiznaty. Hromadske radio. URL: https://hromadske.radio/news/2024/09/09/yak-rosiia-vykorystovuie-shi-dlia-stvorennia-feykiv-ta-iak-ikh-rozpiznaty (in Ukrainian).
2. Hembik O. (2025). Dezinformatsiia y ShI: yak rosiiska propahanda distalas chatbotiv. URL: https://www.sestry.eu/statti/manipulyaciyi-u-merezhi-yak-rosiyska-dezinformaciya-distalasya-chatbotiv (in Ukrainian).
3. Kabinet Ministriv Ukrainy (2025). Bezpechnyi ShI dlia milioniv ukraintsiv. URL: https://www.kmu.gov.ua/news/bezpechnyi-shi-dlia-milioniv-ukraintsiv-ukraina-pidpysala-ramkovu-konventsiiu-pro-shtuchnyi-intelekt-ta-prava-liudyny (in Ukrainian).
4. Lytvynov V. (2024). Yak Ukraina vykorystovuie ShI u viini z Rosiieiu. URL: https://texty.org.ua/fragments/112210/yak-ukrayina-vykorystovuye-shtuchnyj-intelekt-u-vijni-z-rosiyeyu-the-economist (in Ukrainian).
5. MediaMaker (2025). Yak rosiiska propahanda manipuliuie shtuchnym intelektom. URL: https://mediamaker.me/shi-zhertva-dezinformacziyi-yak-rosijska-propaganda-navchylasya-manipulyuvaty-shtuchnym-intelektom-16206 (in Ukrainian).
6. Ornatskyi A (2024). Yak pratsiuie War of Words – novyi ShI-instrument dlia analizu rosiiskoi propahandy. Detector Media. URL: https://detector.media/infospace/article/228955/2024-06-30-yak-pratsyuie-war-of-words-novyy-shi-instrument-dlya-analizu-rosiyskoi-propagandy (in Ukrainian).
7. RBK Ukraina (2024). Rosiia vykorystovuie shtuchnyi intelekt dlia dezinformatsii. URL: https://www.rbc.ua/rus/news/rosiya-vikoristovue-shtuchniy-intelekt-dezinformatsiyi-1729204517.html (in Ukrainian).
8. Sydorskyi V (2023). Yak za dopomohoiu AI protydiiaty rosiiskii dezinformatsii. URL: https://dou.ua/forums/topic/44993 (in Ukrainian).
9. Ukrinform (2024). Rosiia vykorystovuie ShI tekhnolohiiu OpenAI dlia poshyrennia propahandy proty Ukrainy. URL: https://www.ukrinform.ua/rubric-world/3869959-rosia-vikoristovue-sitehnologiu-openai-dla-posirenna-propagandi-proti-ukraini.html (in Ukrainian).
10. Brandt J. (2023). Propaganda, foreign interference, and generative AI. Brookings Institution. URL: https://www.brookings.edu/articles/propaganda-foreign-interference-and-generative-ai

11. Ezzeddine F, Luceri L, Ayoub O, Sbeity I, Nogara G, Ferrara E, Giordano S (2022). Exposing Influence Campaigns in the Age of LLMs: A Behavioral Based AI Approach to Detecting State Sponsored Trolls. arXiv preprint. URL: https://arxiv.org/abs/2210.08786

12. 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 preprint. URL: https://arxiv.org/abs/2301.04246

13. Kuznetsova E, Makhortykh M, Vziatysheva V, Stolze M, Baghumyan A, Urman A (2023). In Generative AI we Trust: Can Chatbots Effectively Verify Political Information? arXiv preprint. URL: https://arxiv.org/abs/2312.13096

14. Majchrzak A (2023). Russian disinformation and the use of images generated by artificial intelligence (deepfake) in the first year of the invasion of Ukraine. Media Biznes Kultura 1(14). URL: https://ejournals.eu/pliki_artykulu_czasopisma/pelny_tekst/01936778-4cef-7088-b5be-d63c9c99d02b/pobierz

15. Marushchak A., Petrov S., Khoperiya A. (2025). Countering AI-powered disinformation through national regulation: learning from the case of Ukraine. Front. Artif. Intell. 7:1474034. https://doi.org/10.3389/frai.2024.1474034

16. Menn J, Zakrzewski C (2025). Russia seeds chatbots with lies. Washington Post. URL: https://www.washingtonpost.com/technology/2025/04/17/llm-poisoning-grooming-chatbots-russia

17. Sadeghi M S (2025). AI and Data Voids: How Propaganda Exploits Gaps in Online Information. Lawfare. URL: https://www.lawfaremedia.org/article/ai-and-data-voids--how-propaganda-exploits-gaps-in-online-information

18. Sprenkamp K, Jones D G, Zavolokina L (2023). Large Language Models for Propaganda Detection. arXiv preprint. URL: https://arxiv.org/abs/2310.06422

19. Wack M., Ehrett C., Linvill D., Warren P. (2025). Generative propaganda: Evidence of AI's impact from a state-backed disinformation campaign. PNAS Nexus 4(4). April 2025. https://doi.org/10.1093/pnasnexus/pgaf083

Надійшла до редакції / Received: 27.04.2025 *Схвалено до друку / Accepted: 02.06.2025*