Ius et Securitas 2/2024 | ISSN 3071-799X

Tomasz Gergelewicz
Operational Centre
Ministry of National Defence
ORCID: 0000-0002-9145-5099
t.gergelewicz@wp.pl

Bipolarity of Artificial Intelligence – Chances and Threats

Abstract

Artificial intelligence (AI) may be characterized as a bipolar phenomenon. It may be used to generate and evolve prosperity for societies or it may be a hostile tool in hands of actors that aim at distorting social cognition. There are many cyber actions with AI employed that either support or harm the inhabitants of the infosphere. Due to the fact nowadays' generations function vastly in the cyberspace, where ethical values and morale are dispersed and slowly diminish, it is much easier to use AI to conduct hostile operations, like disinformation. The whole span of a wide hostile influence may be framed as a cognitive warfare. Governmental and non-governmental institutions are striving to implement efficient solutions regarding the challenge posed by AI. The expected result may be to create and sustain secure and trustworthy environment for societies.

The aim of this article is to highlight the bias nature of artificial intelligence by pointing the opportunities AI gives to societies and the threat it may pose at the same time. The considerations are conducted employing analytical and critical method through a review of available sources. It allows for in-depth consideration of AI tools that are used for social evolvement, as well as for attack and defensive purposes. Mainly, two types of sources have been taken into account. The first are Internet portals that professionally and comprehensively discuss the functioning of AI together with explaining the basic notions associated with cybersphere. The second type of sources constitute reports, projects, initiatives and official publications that institutionally treat on the issue of countering disinformation and building social resilience.

Keywords: artificial intelligence, disinformation, cybersecurity, cognitive warfare.

Introduction

Artificial intelligence (AI) is considered as a breakthrough technology, especially due to the fact that it generates the potential to change lifestyles and the way of perceiving the reality. AI shall be treated as an intelligent form of the computer program, which imitates human reasoning, especially taking into consideration some of the "intelligence" definitions: 1) ability to learn, understand or to deal with new or trying situations[1]; 2) faculty of understanding; intellect, a capacity to understand[2]; 3) good understanding or a high mental capacity; quick to comprehend, as persons or animals[3].

AI is capable of adapting and self-learning, which allows to solve problems in a dynamic way. Its analytical and predictive capabilities enable to create increasingly advanced and intelligent systems that respond efficiently to shifts in environment[4].

One of the most important advantage of AI is that it can analyze vast amounts of data and take particular actions on its basis. It is especially important in the area of security, particularly in cybersecurity[5]. Nevertheless, the nature of AI has to be considered also from the perspective of modern-world evolvement, "The technological revolution has made the modern world become dual and simultaneous — real and virtual at the same time. The conditions and forms of those two spaces have created an environment in which a technological community is being shaped and developed"[6].

Regarding the bipolar character of AI, there is a broad list of AI tools that may be used for causing both: a positive and negative effects for societies. Considering threats, the main disadvantage of AI is that it may be used to conduct

---

[1] *Merriam Webster Dictionary*, www.merriam-webster.com/dictionary/intelligence, [accessed: 18.10.2024].

[2] *Oxford English Dictionary*, www.oed.com/dictionary/intelligence_n?tl=true, [accessed: 18.10.2024].

[3] *Dictionary.com*, www.dictionary.com/browse/intelligent, [accessed: 18.10.2024].

[4] M. Wrzos, *Sztuczna inteligencja – szanse i zagrożenia dla bezpieczeństwa w sieci*, https://promocja.wat.edu.pl/cyberwat/sztuczna-inteligencja-szanse-i-zagrozenia-dla-bezpieczenstwa-w-sieci/, [accessed: 11.10.2024].

[5] J. Walarus, J. Danilewicz, *Bezpieczeństwo danych w czasach AI*, www.ey.com/pl_pl/insights/cybersecurity/bezpieczenstwo-danych-w-czasach-ai-ai-fy24, [accessed: 10.10.2024].

[6] T. Szczurek, J. Grubicka, D. Zbroszczyk, *Borders of Digital Network Freedom in Public Security Space*, "Internal Security" 2020, No. 1, p. 211-223.

disinformation process in a very efficient manner. Possible cognitive impact the artificial intelligence may have, is connected with capability to make societies vulnerable to manipulation, fake news and information noise (MFN).

The aim of this article is to stress the dual nature of AI by highlighting numerous tools, which may be employed that either support or harm the inhabitants of the infosphere. The analytical and critical method are used, as it allows for in-depth consideration of AI tools for development of societies, as well as for attack and defense. Generally, two types of sources have been considered. First regards professional Internet portals that comprehensively discuss the functioning of AI together with explaining the basic terms associated with cybersphere. Due to the extremely dynamic development of these fields, Internet sources constitute a natural and up-to-date database on the phenomenon of AI. The second type of sources include reports, projects, initiatives and official publications that treat on countering disinformation and building social resilience.

Dual-nature of artificial intelligence – the tool for development

The world of AI has witnessed a remarkable evolution, especially in the realm of natural language processing. Large language Models (LLM) like OpenAI's ChatGPT, Google's Bard, Anthropic's Claude are best examples of the evolvement of this cutting-edge technology. These models stand for catalysts in transforming the way people interact with digital platforms[7].

Large Language Models are very large deep learning models that are pre-trained on vast amounts of data. The underlying transformer is a set of neural networks that consist of an encoder and a decoder with self-attention capabilities. The encoder and decoder extract meanings from a sequence of text and understand the relationships between words and phrases. Transformer LLMs are capable of unsupervised training, so it performs self-learning[8].

[7] *Comparative Analysis between OpenAI ChatGPT, Google Bard, Anthropic Claude, and Google Gemini*, https://medium.com/@DigitalQuill.ai/comparative-analysis-between-chatgpt-bard-claude-and-gemini-ai-llms-30b7a4488c5d, [accessed: 13.11.2024].

[8] *What is LLM (Large Language Model)?*, https://aws.amazon.com/what-is/large-language-model/, [accessed: 9.09.2024].

Chat GPT (Generative Pre-Trained Transformer) is an artificial intelligence program that generates dialogue. High-capable chatbot uses machine learning algorithm to process and analyze large amounts of data in order to generate responses to user inquiries. This language processing program is able to understand human language as it is spoken and written, allowing to understand the message and give back understandable answer in a variety of formats with precise stipulations[9].

Claude is a next-generation AI assistant based on anthropic's research into training helpful, honest, and harmless AI systems. Claude is capable of a wide variety of conversational and text processing tasks, while maintaining a high degree of reliability and predictability. Claude can help with summarization, search, creative and collaborative writing, Q&A, coding, and more. Claude can also take direction on personality, tone, and behavior[10].

Bard is a Google conversational AI chatbot that can generate texts of all kinds. Bard uses Google's own LLM, known as Language Model for Dialogue Applications (LaMDA). The result is a chatbot that can answer any question in natural and conversational language[11]. Bard is specifically crafted to assist users in brainstorming ideas and retrieving information from the Internet in a personalized manner. The objective is to make AI more accessible and valuable to the average consumer[12].

Machine learning algorithms are computational models that allow computers to understand patterns and forecast or make judgments based on data without explicit programming. These algorithms form the foundation of artificial intelligence. It is used in various applications, including image and speech recognition, natural language processing, etc.[13].

One of the concept developed to make AI acceptable and harmless is a notion of Explainable Artificial Intelligence (XAI). Explainable AI is used to describe an AI model, its expected impact and potential biases. It

---

[9] *Chat GPT: What is it?*, https://uca.edu/cetal/chat-gpt/, [accessed: 13.11.2024].

[10] *Introducing Claude*, https://www.anthropic.com/news/introducing-claude, [accessed: 23.11.2024].

[11] J. Martindale, *What is Google Bard? Here's how to use this ChatGPT rival*, https://www.digitaltrends.com/computing/how-to-use-google-bard/, [accessed: 23.11.2024].

[12] K. Allan, *Google Bards ambition to reach millions of users*, https://aimagazine.com/articles/google-bards-ambition-to-reach-millions-of-users, [accessed: 9.10.2024].

[13] *Algorytmy uczenia maszynowego*, https://www.geeksforgeeks.org/machine-learning-algorithms/, [accessed: 23.11.2024].

helps to characterize model accuracy, fairness, transparency and outcomes in AI-powered decision making[14]. Thanks to XAI, security specialists are capable of detailed monitoring of systems and make quick actions in case of uncovered any falsehood. Important to mention is the trustworthy concept, which states for defining particular characteristics that should label AI as a credible tool[15]: 1) lawful – respecting all applicable laws and regulations; 2) ethical – respecting ethical principles and values; 3) robust – both, from a technical perspective and its social environment[16].

The use of AI opens new opportunities for cybersecurity experts and law enforcement agencies to use AI tools to counter illegal activities. It may be used to develop strategies through automating data analysis in order to detect attacks and more effective identification of hostile actors[17]. Experts emphasize the need to better understand and properly regulate AI, especially regarding its social impact[18]. Regarding advantages of using this technology, AI: 1) enables effective detection of anomalies in network traffic. Machine learning algorithms can identify abnormal users' behavior or changes in the system, which may indicate a threat; 2) uses heuristic analysis to identify new forms of attacks different from traditional signature databases, which may indicate undiscovered techniques of cybercriminals; 3) engages predictive algorithms that can analyze historical data and foresee the evolution of threats. It allows to prepare for new techniques of attack; 4) can automatically respond to a detected incidents, by appropriate corrective or isolating actions; 5) uses advanced content analysis that help identifying attempts of phishing attacks. It can detect suspicious elements in e-mails, and protect users from disclosing confidential data; 6) identifies new types of malware by their characteristics. It allows for a quick reaction to new threats without

---

[14] *What is Explainable AI?*, https://www.ibm.com/topics/explainable-ai, [accessed: 10.11.2024].

[15] C. Maślankiewicz, *Sztuczna inteligencja a cyberbezpieczeństwo*, https://www.crowe.com/pl/insights/sztuczna-inteligencja-a-cyberbezpieczenstwo, [accessed: 10.11.2024].

[16] *The Assessment List for Trustworthy Artificial Intelligence (ALTAI),* https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, [accessed: 9.10.2024].

[17] M. Marszycki, *Jak sztuczna inteligencja wpływa na obszar cyberbezpieczeństwa?*, https://itwiz.pl/jak-sztuczna-inteligencja-wplywa-na-obszar-cyberbezpieczenstwa/, [accessed: 5.11.2024].

[18] *Badacze zamierzają wykorzystać sztuczną inteligencję do zwalczania hejtu i dezinformacji*, https://ibims.pl/badacze-chca-wykorzystac-sztuczna-inteligencje-do-zwalczania-hejtu-i-dezinformacji/, [accessed: 15.10.2024].

a necessity to wait for signature database updates; 7) analyzes application behavior by identifying potential threats associated with unauthorized access or malware. It works on the basis of monitoring applications' activity and responding to suspicious activities[19].

Dual-nature of artificial intelligence – the tool for disinformation

One of the key challenges is the development of generative AI models, with the disinformative capability, which means manipulation with the content, creating fake news and introducing information noise. Disinformation campaigns carried out using AI generated content lead to „the erosion of citizens' trust in political leaders, elections, media and democratic governments"[20]. The hostile process of disinformation is a part of the broader phenomenon of a cognitive warfare, which integrates cyber, information, psychological, and social engineering capabilities as instruments of power[21]. The cognitive warfare is also characterized as a new domain of the battlefield, with the wide scale impact on the entire population[22]. Therefore, "although all of them – cyberwarfare, information warfare, cognitive warfare, and hybrid warfare – contain the element of influence operations and may impact human cognition, only cognitive warfare is specifically dedicated to brain control by incorporating weaponized neuroscience into various practices"[23]. In general terms, the result of the cognitive warfare is supposed to have on societies, is strictly connected with other military goals: "obtaining a tactical or strategic advantage (...). The pursued objective is to influence not only what the targets think, but also the way they think and, ultimately, the way

[19] M. Wrzos, *Sztuczna inteligencja – szanse i zagrożenia dla bezpieczeństwa w sieci*, https://promocja.wat.edu.pl/cyberwat/sztuczna-inteligencja-szanse-i-zagrozenia-dla-bezpieczenstwa-w-sieci/, [accessed: 11.10.2024].

[20] *Poszukiwanie rozwiązań problemu fałszywych informacji i obrazów,*https://cordis.europa.eu/article/id/450241-unravelling-the-deepfake/pl, [accessed: 18.11.2024].

[21] *Strengthening and Defending the Mind*, www.act.nato.int/article/cognitive-warfare-strengthening-and-defending-the-mind, [accessed: 27.10.2024].

[22] L. Fenstermacher, D. Uzcha, *New perspectives on cognitive warfare*, "Proceedings of the SPIE" 2023, No. 12547, p. 16.

[23] T.C. Hung, T.W. Hung, *How China's Cognitive Warfare Works: A Frontline Perspective of Taiwan's Anti-Disinformation Wars*, "Journal of Global Security Studies" 2022, No. 4.

they act"[24]. In terms of conducting hostile information operations, It is emphasized that these are conducted on a large scale and began to appear during the global war against terrorism. Due to generative AI, it is possible to create increasingly better deepfake or conduct doppelganger operations[25]. In addition, AI can be helpful in creating bots together with the entire networks, as well as running fake social media accounts. The goal of disinformators is to create echo chambers and finally lead to social apathy. The key is to be aware that AI, through its functional algorithms, is building a psychological profile, characterizes needs and, on the basis of the abovementioned, choose and propose particular content. Then it creates filter bubble, which cut the audience from differentiated information from multiple sources. There is also a distinctive threat when AI generates false information by "itself", called AI hallucinations[26]. Moreover, it can be observed that nowadays societies are addicted from information, which is called the infodemic. Together with AI capabilities it makes a highly dangerous combination[27].

Echo chamber is a phenomenon when a person encounters only information or opinions that reflect and reinforce its own ideas. Echo chambers can distort a person's perspective, and lead to difficulty in considering opposing viewpoints and discussing complicated topics. Echo chambers are based on tendency to favor information that reinforces existing beliefs[28].

Filter bubble (also called information bubble) is the intellectual isolation that occurs when AI-driven algorithms selectively present information based on user's behavior, such as clicks, browsing history and location. This results in users encountering information that matches their existing interests.

---

[24] B. Claverie, F. Cluzel, *"Cognitive Warfare": The Advent of the Concept of "Cognitics" in the Field of Warfare,* https://www.researchgate.net/publication/359991886_Cognitive_Warfare_The_Advent_of_the_Concept_of_Cognitics_in_the_Field_of_Warfare, [accessed: 25.09.2024].

[25] More about deepfake and doppelganger operations T. Gergelewicz*, Fake Mirroring in Cyberspace as a Disinformation Tool – Doppelgangers and Deepfakes*, "Cybersecurity and Law" 2024, No. 1.

[26] *Rozwój AI to rozwój dezinformacji. Prognozy z globalnego szczytu*, https://cyberdefence24.pl/cyberbezpieczenstwo/cybermagazyn-rozwoj-ai-to-rozwoj-dezinformacji-prognozy-z-globalnego-szczytu, [accessed: 11.10.2024].

[27] R. Kupiecki, T. Chłoń, *Sztuczna inteligencja: miecz i tarcza (dez)informacji*, https://itwiz.pl/sztuczna-inteligencja-miecz-i-tarcza-dezinformacji/, [accessed: 9.10.2024].

[28] *What is an echo chamber?,* https://edu.gcfglobal.org/en/digital-media-literacy/what-is-an-echo-chamber/1/, [accessed: 9.10.2024].

Filter bubbles limit exposure to balanced information and diverse perspectives. It leads to intellectual isolation, reducing contact with opposing viewpoints and multi-sources information[29].

Infodemic combines an inordinately high volume of information (lead to ensuring quality, visibility and validity) and rapid output (make it hard to assess its value, apply results, track its history)[30]. It also refers to a large increase in the volume of information associated with a specific topic[31].

AI hallucinations appear when AI systems regularly produce incorrect answers and present these answers with high confidence. These errors can be caused by a variety of factors, including insufficient training data, incorrect assumptions made by the model, or biases in the data used to train the model. AI hallucinations can be a problem for AI systems used to make important decisions, such as medical diagnoses or financial trading[32].

It is estimated that the use of AI will be widespread among Advanced Persistent Threats (APT) groups and the most powerful criminal organizations in order to spread disinformation. These include creating deepfake, fake websites, conducting disinformation campaigns, or programming credible bots on social media. As emphasized, „Thanks to AI, it is easier for cybercriminals to analyze huge amounts of data, including these obtained during attacks. This allows to identify weak points or high-value targets, enabling more precise and effective attempts, with higher profits"[33]. The example may be a DeepLocker, a malware based on AI. It might be used to raise the efficiency of attacks through defining conditions of software activation. A hostile software is activated only, when the aim of the attacked had been defined and thanks to staying hidden in "safe applications", it remain uncovered[34].

---

[29] *What is Filter Bubble?*, https://www.techopedia.com/definition/28556/filter-bubble, [accessed: 29.10.2024].

[30] C. Zieliński, *Infodemics and infodemiology: a short history, a long future*, https://iris.paho.org/handle/10665.2/53850, [accessed: 10.11.2024].

[31] *Understanding the Infodemic and Misinformation in the fight against COVID-19*, https://iris.paho.org/bitstream/handle/10665.2/52052/Factsheet-infodemic_eng.pdf, [accessed: 29.10.2024].

[32] *What are AI hallucinations?,* https://cloud.google.com/discover/what-are-ai-hallucinations, [accessed: 9.11.2024].

[33] M. Marszycki, *Jak sztuczna inteligencja wpływa na obszar cyberbezpieczeństwa?*, https://itwiz.pl/jak-sztuczna-inteligencja-wplywa-na-obszar-cyberbezpieczenstwa/, [accessed: 5.11.2024].

[34] Ibidem.

Advanced Persistent Threat (APT) is an attack campaign in which an intruder establishes an illicit, long-term presence on a network in order to mine highly sensitive data. The consequences of such intrusions include: intellectual property theft (e.g., trade secrets or patents), compromised sensitive information (e.g., employee and user private data), the sabotaging of critical organizational infrastructures (e.g. database deletion) and total site takeovers[35].

DeepLocker is an AI-powered malware that infects millions of systems without being detected. The malicious payload of DeepLocker will only be unveiled if the target is reached. It is achieved by using a deep neural network AI model[36].

Malware is any software code or computer program including ransomware, trojan horses and spyware, intentionally written to harm computer systems or their users[37].

Ransomware is a type of malware that hostage sensitive data or device, threatening to keep it locked, unless the victim pays a ransom to the attacker[38].

Another hostile example of AI use is so called, AI-powered phishing email. These are AI generated swindle messages. E-mails look like real and often it is impossible to recognize their fake nature. Additionally, AI is used to sneak through systems called Completely Automated Public Turing Test to Tell Computers and Humans Apart (CAPTCHA), which were programmed to differentiate bots from humans. AI is also used to crack passwords using the 'brute-force' method, so checking all possible combinations[39]. All the above may be characterized as having particular features common for hostile cyber-tools: 1) phishing emails are deceptive messages pretend to be from legitimate sources but aimed at tricking users to reveal personal information or to click on harmful links; 2) CAPTCHA is a security measure used to prevent automated web crawlers (bots), from commenting, submitting forms,

---

[35] *Advanced persistent threat (APT)*, www.imperva.com/learn/application-security/apt-advanced-persistent-threat/, [accessed: 7.11.2024].

[36] M. Stoecklin, J. Jang, D. Kirat, *DeepLocker: How AI Can Power a Stealthy New Breed of Malware,* "Security Intelligence" 2018, No. 8.

[37] *What is Malware?*, https://www.ibm.com/topics/malware, [accessed: 20.11.2024].

[38] M. Kosinski, *What is Ransomware?*, https://www.ibm.com/topics/ransomware, [accessed: 20.11.2024].

[39] J. Walarus, J. Danilewicz, *Bezpieczeństwo danych w czasach AI*, https://www.ey.com/pl_pl/insights/cybersecurity/bezpieczenstwo-danych-w-czasach-ai-ai-fy24, [accessed: 10.10.2024].

or otherwise spamming websites[40]; 3) brute-force attack uses trial-and-error method to guess login information, encryption keys, or find a hidden web page. Hackers work through all possible combinations hoping to guess correctly. Because depending on the length and complexity of the password, cracking it can take anywhere from a few seconds to many years.

It is estimated that it takes just 1 Euro to buy or set up a fake social media account. Thus, the settlement of a background for disinformation is not a costly process, especially for entities with multi-million budgets, such as the Internet Research Agency. The development of Large Language Models may cause that costs of sharing disinformation may be even lower. Experts from the Center for Security and Emerging Technology underline that "it is all but certain that a well-run human-machine team that utilized existing LLMs would save a propagandist money on content generation relative to a human-only operation"[41]. Considering the above, using LLM in order to generate 10 million messages on "X" platform could save 3 million dollars[42]. When estimating the costs of disinformation, it is worth mentioning that in 2023 the Dossier Center published details about the troll factory founded by Yevgeny Prigozhin. According to the Center, trolling Internet empire had a disposable budget of about 100 million rubles each month, which stands approximately for 1.1 million dollars[43]. Moreover, regarding cybersecurity, AON presented the report: Global Risk Management Survey, in which it is highlighted that in years 2022-2023 an average cost of a single violation of data in company raised to almost 4,5 million dollars. However in companies, which did not employed AI in cybersecurity, the cost was estimated for about 5,4 million dollars[44]. Moreover, as highlighted in the report: AI for Impact: Strengthening AI Ecosystems for Social Innovation, "the transfor-

---

[40] D. Bodnar, *What Is CAPTCHA and How Does It Work?*, https://www.avg.com/en/signal/what-is-captcha, [accessed: 12.09.2024].

[41] M. Musser, *How Much Money Could Large Language Models Save Propagandists?*, "Center for Security and Emerging Technology" 2023, No. 9.

[42] *Rozwój AI to rozwój dezinformacji. Prognozy z globalnego szczytu*, https://cyberdefence24.pl/cyberbezpieczenstwo/cybermagazyn-rozwoj-ai-to-rozwoj-dezinformacji-prognozy-z--globalnego-szczytu, [accessed: 11.10.2024].

[43] I. Koval, *Prigozhin's troll factories in Russia: What's next?*, www.dw.com/en/prigozhins-troll-factories-in-russia-whats-next/a-66194462, [accessed: 12.11.2024].

[44] *Global Risk Management Survey Key Findings,* www.aon.com/en/insights/reports/global-risk-management-survey/key-findings, [accessed: 12.11.2024].

mative potential of artificial Intelligence (AI) is widely lauded. With studies revealing that employees who deploy machine learning can realize up to 30% in efficiency gains[45] and findings that the technology has the potential to automate up to 70% of activities that take up employees' time[46], it is no surprise that this technology is being embraced by industry[47].

As generative AI technology becomes more accessible and affordable, the number of cases when it is used in large-scale disinformation campaigns is growing. The European Commission points several reasons for spreading disinformation: 1) rapid cultural, social and economic changes that lead to a sense of social insecurity, tensions, and a lack of trust in public institutions; 2) the transformation of the media sector, with vast popularity of Internet platforms; 3) automated and fake accounts reach wide audience and becomes a political tool. It is also strengthened by the behavior of users, who engage into an emotional and catchy content without checking its credibility[48].

Technology used for information warfare is also employed by state actors. The "technological armaments race" between global superpowers is conducted both for the sustainment of international position, as well as diminishing differences in traditional armaments potential. The hostile actors use functionalities of AI to booster the speed of messaging, individualization, multiplication of information, and autonomic content creator. These include not only operations within content and image, but also the fabrication of human identity, giving the authenticity to fake accounts on social networks, and messages targeted against particular persons. When embedded into politics, it may have severe consequences, including an impact on people's voting, attitudes toward competitors, political decisions and armed conflicts[49].

[45] More in: *Operationalizing machine learning in processes,* https://www.mckinsey.com/capabilities/operations/our-insights/operationalizing-machine-learning-in-processes, [accessed: 22.11.2024].

[46] More in: *The mechanisms of AI hype and its planetary and social costs,* https://doi.org/10.1007/s43681-024-00461-2, [accessed: 22.11.2024].

[47] *AI for Impact: Strengthening AI Ecosystems for Social Innovation,* https://www3.weforum.org/docs/WEF_AI_for_Impact_2024.pdf, [accessed: 2.11.2024].

[48] *Zwalczanie dezinformacji w Internecie: podejście europejskie,* https://cyberpolicy.nask.pl/zwalczanie-dezinformacji-w-internecie-podejscie-europejskie/, [accessed: 12.11.2024].

[49] R. Kupiecki, T. Chłoń, *Sztuczna inteligencja: miecz i tarcza (dez)informacji*, https://itwiz.pl/sztuczna-inteligencja-miecz-i-tarcza-dezinformacji/, [accessed: 9.10.2024].

AI-based attacks are broadly used to: 1) generate false training data for security systems, which can introduce misleading machine learning models and result in wrong decisions; 2) create more credible and personalized phishing messages, which increase the effectiveness of attacks; 3) attack on decision-making process in systems, e.g. used in autonomous cars or defense systems. It may lead to incorrect assessments and wrong decisions; 4) manipulate detection systems to avoid being revealed during attempted attacks; 5) conduct automated individual attacks that can lead to more complex hostile activities and difficulties in responding by defense systems; 6) generate any content, including personalized photos, sounds, and videos. It increases the risk of effective attacks by lowering social awareness of threats[50].

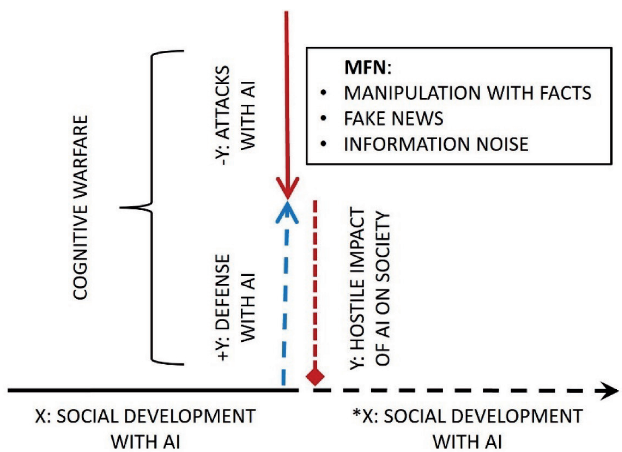

Figure 1. Bias of Artificial Intelligence (AI)
Sources: Self-inspired.

Figure 1 describes the bias nature of AI, which derives from its dual use. Social development is accompanied by the employment of AI in many branches, and this tendency is raising. In the figure, the unobstructed social evolvement is depicted as "X". However, hostile actors conduct attacks with the use of AI, in order to distort the human cognition and in result,

[50] M. Wrzos, *Sztuczna inteligencja – szanse i zagrożenia dla bezpieczeństwa w sieci*, https://promocja.wat.edu.pl/cyberwat/sztuczna-inteligencja-szanse-i-zagrozenia-dla-bezpieczenstwa-w-sieci/, [accessed: 11.10.2024].

influence the perception competences of the targeted audience. The cognitive war is conducted by the use of disinformative tools: manipulation with facts, creation of fake news, implementation of information noise. Attacks with the use of AI are pictured as "-Y". This state of affairs stands for the subject matter bias, so positive and negative use of AI. Nevertheless, it has to be considered that AI is also used to counter hostile activities, depicted as "+Y", and the result of "-Y+Y" constitutes the real hostile impact "Y". Another matter important to regard is that these processes take place in the infosphere together with cybersphere within the cognitive sphere.

$$*\mathbf{X} = \mathbf{X} - \left[\frac{\mathbf{-Y+Y}}{\mathbf{A \neq A}} \right]$$

Figure 2. Social development with minimum hostile AI impact
Sources: Self-inspired.

The "Social development with minimum hostile AI impact" may be depicted by the use of the above scheme. The development of societies with the minimum hostile AI impact (*X), results from the state of unobstructed social development (X) disturbed with attacks (-Y) countered by defensive activities (+Y). It is important to notice that there is no equality between the scope of (-Y) and (+Y), because of the fact that the advancement of employed AI (A) plays a key role. It concerns the advancement level of AI tools, and both reactive and proactive countering solutions. The more successful reactive and proactive measures are, the less impacted is the society.

Initiatives to face the challenge of double-nature of artificial intelligence

Lots of governmental and non-governmental, national and international institutions strive to characterize AI, together with its positive and negative influence on societies, and advantages and disadvantages of its development in particular cyber and social directions.

As highlighted in the report: Capabilities and risks from frontier AI – a discussion paper, "Frontier AI can generate hyper-targeted content with unprecedented scale and sophistication. This could lead to "personalized" disinformation, where bespoke messages are targeted at individuals rather than larger groups and are therefore more persuasive. Furthermore, one should expect that as AI-driven personalized disinformation campaigns unfold, these AI will be able to learn from millions of interactions and become better at influencing and manipulating humans, possibly even becoming better than humans. Thus, they may utilize new manipulation tactics against which we are not prepared because defenses have been developed through the influencing attempts of other humans"[51]. The authors of the report foresee that deepfake will be more and more realistic and their development will be directed on personalization and micro targeting[52].

As stressed in the report from the European Telecommunications Standards Institute (ETSI): Securing Artificial Intelligence (SAI); Automated Manipulation of Multimedia Identity Representations, "Methods for manipulating multimedia identity representations can be used to exert influence on public opinion. The modus operandi consists in publishing manipulated media which falsely create the impression that persons holding influential positions have written, said or done certain things. In principle, this applies to all purposes where stakes are high and the benefit justifies the effort from an attacker's perspective"[53].

On the AI Safety Summit in Bletchley, politicians, technology enterprises representatives, experts and scientists discussed the risks associated with the AI development.

As stated in the declaration from the Summit: "Artificial Intelligence (AI) presents enormous global opportunities: it has the potential to transform and enhance human wellbeing, peace and prosperity. To realize this, we affirm that, for the good of all, AI should be designed, developed, deployed, and used, in a manner that is safe, in such a way as to be human-centric, trustworthy and responsible (...). Alongside these opportunities, AI also poses significant risks,

---

[51] *Capabilities and risks from frontier AI*, https://assets.publishing.service.gov.uk/media/65395abae6c968000daa9b25/frontier-ai-capabilities-risks-report.pdf, [accessed: 20.11.2024].

[52] *Rozwój AI to rozwój dezinformacji. Prognozy z globalnego szczytu*, https://cyberdefence24.pl/cyberbezpieczenstwo/cybermagazyn-rozwoj-ai-to-rozwoj-dezinformacji-prognozy-z-globalnego-szczytu, [accessed: 11.10.2024].

[53] *Securing Artificial Intelligence (SAI)*, https://www.etsi.org/deliver/etsi_gr/SAI/001_099/011/01.01.01_60/gr_SAI011v010101p.pdf, [accessed: 13.11.2024].

including in those domains of daily life. To that end, we welcome relevant international efforts to examine and address the potential impact of AI systems in existing fora and other relevant initiatives, and the recognition that the protection of human rights, transparency and explainability, fairness, accountability, regulation, safety, appropriate human oversight, ethics, bias mitigation, privacy and data protection needs to be addressed. We also note the potential for unforeseen risks stemming from the capability to manipulate content or generate deceptive content"[54]. It was underlined that AI is a great challenge especially because its quick evolvement, "For the purposes of the Summit we define "frontier AI" as highly capable general-purpose AI models that can perform a wide variety of tasks and match or exceed the capabilities present in today's most advanced models. Today, this primarily includes large language models (LLMs) such as those underlying ChatGPT, Claude and Bard. However, it is important to note that, both today and in the future, frontier AI systems may not be underpinned by LLMs, and could be underpinned by another technology"[55]. Regarding the hostile use of AI for the deception and manipulation purposes in the after-summit report, disinformation was categorized as the social harm and abuse[56].

The European Commission set up a High-Level Expert Group to answer the question of, how to fit for the future social protection systems, particularly in light of trends such as: the demographic changes resulting in a shrinking workforce and an ageing population, and also transformations on the labor market accompanying the digital and green transitions. It includes the increase in non-standard forms of work, globalization, and the emergence of new risks[57]. The Group plays an advisory role to the Commission on issues arising around the spread of false information and on possible solutions to prevent its social and political consequences. It developed a report that highlights the problem of disinformation as being deeply connected with the

---

[54] *The Bletchley Declaration by Countries Attending the AI Safety Summit*, https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023, [accessed: 16.10.2024].

[55] *Capabilities and risks from frontier AI*, https://assets.publishing.service.gov.uk/media/65395abae6c968000daa9b25/frontier-ai-capabilities-risks-report.pdf, [accessed: 16.10.2024].

[56] Ibidem.

[57] *High Level Group on the future of social protection and of the welfare state in the EU (E03796)*, https://ec.europa.eu/transparency/expert-groups-register/screen/expert-groups/consult?lang=en&groupID=3796&fromCallsApplication=true, [accessed: 4.11.2024].

development of digital media[58]. As stressed in the report: "Our main message is that the best responses to disinformation are multi-dimensional, with stakeholders collaborating in a manner that protects and promotes freedom of expression, media freedom, and media pluralism. Another key message is that effective action will require continuous research on the impact of disinformation, increased transparency, and access to relevant data, combined with evaluation of responses on a regular, ongoing basis. This is particularly important as disinformation is a multifaceted and evolving problem that does not have one single root cause. The recommendations in this Report aim to provide short term responses to the most pressing problems, longer-term responses to increase societal resilience to disinformation, and a framework for ensuring that the effectiveness of these responses is continuously evaluated and, new evidence based responses are developed. Our recommendations are addressed to the Commission and the Member States as well as to stakeholders: platforms, news media and civil society organizations. We believe that our Report provides the ingredients for an effective pan-European strategy to tackle the complex and disruptive problem of disinformation"[59].

Authors of the "White Paper": Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities analyze the interconnection of AI and disinformation. The Paper is a result of the cooperation of teams financed by the EU projects: vera.ai (VERification Assisted by Artificial Intelligence), TITAN (AI for Citizen Intelligent Coaching against Disinformation), AI4Media (AI European Excellence Centre for Media, Society and Democracy) and AI4TRUST[60] (AI-based-technologies for trustworthy solutions against disinformation). As noted, "Nowadays we witness a dynamic development of the area of AI and associated disinformation. Our „White Paper" is supposed to be a starting point for further research, giving a fresh idea on complex interactions and connections between technological evolvement, ethical challenges and legislation"[61]. As stated in the "White

[58] *A multi-dimensional approach to disinformation : report of the independent High level Group on fake news and online disinformation*, https://data.europa.eu/doi/10.2759/739290, [accessed: 4.11.2024].

[59] Ibidem.

[60] More about the projects: https://www.veraai.eu/home, https://www.titanthinking.eu, https://www.ai4media.eu, ai4trust.eu.

[61] *Poszukiwanie rozwiązań problemu fałszywych informacji i obrazów,*https://cordis.europa.eu/article/id/450241-unravelling-the-deepfake/pl, [accessed: 18.11.2024].

Paper": „Especially relevant in a disinformation context, the risks of manipulation (including emotional manipulation) and AI anthropomorphism are key ethical concerns. The risk of emotional manipulation, which is one of the main risks associated with human-imitating AI was reflected in the recent chatbot-incited suicide in Belgium. Such heavy real-life consequences require a responsible approach to generative AI developments; clear attribution of legal responsibility and liability; and a better balance between the precautionary principle and the innovation principle. Education and awareness campaigns in order to inform people of the risks associated with AI systems are also needed. The risks of AI to fundamental human rights should be first identified, analyzed and mitigated, before the AI application is made publicly available"[62].

Another project with regards to the development of AI tools is the Database of Known Fakes. It assists fact-checkers to distinguish the truth from the falsehood. The project was created as a part of the European Union-funded project called WeVerify[63], which allows, among others, for analyzing disinformation campaigns. Users can search information on the basis of sources, authors, location and other concepts. Location or concept filters enable for the comparison of information in different languages, which allows for more efficient analysis of differences in disinformation regarding different contexts and regions[64]. Based on text analysis, visual similarity and semantic technology, the database provides access to advanced searching capabilities that significantly prevail solutions based just on keyword search[65].

In the Global Risks Report 2024 developed by the World Economic Forum, it is underlined that geopolitical tensions combined with activities of hostile actors are contributing to an unstable global order, which results in eroding security. The report highlights the research from annual Global Risks

[62] K. Bontcheva, *Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities*, https://edmo.eu/wp-content/uploads/2023/12/Generative-AI-and-Disinformation_-White-Paper-v8.pdf, [accessed: 14.11.2024].

[63] More about the projects: https://dbkf.ontotext.com, https://weverify.eu, [accessed: 2.11.2024].

[64] *Database of Known Fakes – skuteczne narzędzie w walce z dezinformacją,* "Cordis", https://cordis.europa.eu/article/id/447911-fake-news-debunking-it-better-with-the-database-of-known-fakes/pl, [accessed: 19.11.2024].

[65] More about the tool on YouTube: https://www.youtube.com/watch?v=xBypSsW2-TE, [accessed: 2.11.2024].

Perception Survey, which brings together the collective data of nearly 1,500 global leaders, "The disruptive capabilities of manipulated information are rapidly accelerating, as open access to increasingly sophisticated technologies proliferates and trust in information and institutions deteriorates. (...) rapidly evolving development of and reliance on advanced machine intelligence is outpacing our ability to adapt – both to understand the technology itself (the "Black Box Problem") and to create regulatory safeguards (the "Pacing Problem"), with regulation playing catch up to the technology. The speed of advances, depth of market power and strategic importance of the industry will continue to challenge the appetite and regulatory capacity of governance institutions"[66].

The report developed by the European Telecommunications Standards Institute: Securing Artificial Intelligence (SAI); Automated Manipulation of Multimedia Identity Representations, discusses the problem of generating or modifying data represented in different media formats (in particular, audio, video and text). It is highlighted that AI techniques can be used to distort authentic multimedia identity or to create fake ones, "Since usually Deep Neural Networks are used for generating such outputs, they are commonly referred to as "deepfakes"[67]. The deepfake phenomenon is not new and is associated with the use of recreating original content by technological tools. However, this new created content may be used out of their original context and attributed to a completely different one. As underlined in the report, "When manipulated multimedia identity representations are used with the aim to influence public opinion, which most often happens in social media, a specific countermeasure is to ensure the authenticity and attributability of content. This can help users distinguish between content from confirmed trustworthy sources and content which originates from untrustworthy sources or for which no information whatsoever about the origin is available. A straightforward way to implement this is to use cryptographic protection mechanisms, in particular digital signatures. Media content originating from eminent public figures or official organizations could then be signed using these parties' private keys, and either the users themselves or

[66] *The Global Risks Report 2024 – 19th Edition*, https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf, [accessed: 2.11.2024].

[67] *Securing Artificial Intelligence (SAI),* https://www.etsi.org/deliver/etsi_gr/SAI/001_099/011/01.01.01_60/gr_SAI011v010101p.pdf, [accessed: 13.11.2024].

the internet platforms distributing them could cryptographically check they have not been altered after the signature has been applied. If the respective keys can be securely bound to the respective parties, the authenticity of the information can likewise be checked"[68].

Next project associated with verification of content is Digital Methods Initiative[69], which discusses mapping the "memory loss" of disinformation in fact checks: „Archiving appearances of disinformation is one of the most cumbersome tasks fact-checkers are performing, due to anti-scraping measures taken by Facebook, Instagram, TikTok, and other platforms. This may result in less documented fact checks (with fewer appearances, and less evidence of disinformation traces). Facebook users may remove posts that have been fact-checked or can make them private. Therefore, many links to content disappear (erased by platforms, end-users, or kept within private groups after debunks) making it difficult for fact-checking memory and also for social scientists to evaluate the scope of disinformation on platforms. Therefore, the first goal of this project to understand the extent of the fact-checking "memory loss""[70]. Moreover, there was a verification of links directing to information on the war in Ukraine. It occurred that at least 15% of links within the analyzed group were poorly archived. Moreover, 23% of unarchived links to Facebook directs to a not existing content[71].

The Ukraine Communications Group (UCG) is the next initiative that brings together partner countries and international organizations from Canada, Czech Republic, Estonia, Finland, France, Germany, Latvia, Lithuania, Poland, Slovenia, United States, Sweden, Ukraine, United Kingdom, Italy, NATO and European External Action Service[72]. The Group is supporting Ukraine in strivings to keep its independence and territorial integrity also by collaborative efforts to counter disinformation and support fact-based reporting. The Kremlin continues the process of disinformation to obfuscate its war aims, and to fracture worldwide solidarity with the Ukrainians. Thus, the

68 Ibidem.

69 D. Teyssou, *Mapping the 'memory loss' of disinformation in fact checks*, https://wiki.digitalmethods.net/Dmi/WinterSchool2023MappingMemoryLoss, [accessed: 27.09.2024].

70 Ibidem.

71 *Zanikająca pamięć o fałszywych wiadomościach*, https://cordis.europa.eu/article/id/442985-losing-our-memory-of-fake-news/pl, [accessed: 15.11.2024].

72 *Ukraine Communications Group activities in Warsaw,* https://www.gov.pl/web/diplomacy/ukraine-communications-group-activities-in-warsaw, [accessed: 20.11.2024].

UCG is designed to counter these hostile narratives and reinforce Ukraine's resilience[73]. As the Polish Ministry of Foreign Affairs underlined, "Representatives of Ukraine and partner states will participate in the activities of the Group with the aim of exchanging information on covering the full-scale aggression of the Russian Federation against Ukraine, strengthening Ukrainian signals and combating information manipulation by the Kremlin"[74].

The Partnership for Cybersecurity program was settled up to support and promote the idea of multi-sectoral cooperation to strengthen the digital resilience of Poland[75]. The Partnership for Cybersecurity is a platform for the information and good practices exchange, as well as sharing experience between entities participating in the program. The main goal is to raise the level of knowledge and strengthen competences in the area of cybersecurity[76]. As emphasized, AI can be increasingly used to combat disinformation on the Internet. That is why, social media owners invest in algorithms detecting and blocking information generated by bots. They cooperate with non-governmental organizations that deal with countering disinformation by AI[77].

The European Union developed the "AI Act", in which the EU Commission proposed four categories of actions, in order to build resilience against the hostile use of AI: 1) ban harmful AI practices that are considered a clear threat to people and evoking unacceptable risk, including AI systems that: a) deploy harmful manipulative techniques, b) exploit specific vulnerable groups with physical or mental disability, c) are used by public authorities, or on their behalf, for social scoring purposes, d) engage real-time remote biometric identification systems in public space for law enforcement purposes (except in a limited number of cases); 2) regulate high-risk AI systems

[73] *The United States and Poland Announce New Ukraine Communications Group,* https://pl.usembassy.gov/the-united-states-and-poland-announce-new-ukraine-communications-group/, [accessed: 23.10.2024].

[74] *The Ministry of Foreign Affairs of Ukraine welcomes the creation of the Ukraine Communications Group*, 2024, https://mfa.gov.ua/en/news/mzs-ukrayini-vitaye-stvorennya-komunikacijnoyi-grupi-z-pitan-ukrayini, [accessed: 23.10.2024].

[75] P. Popov, *Spotkanie Partnerów w ramach programu Partnerstwo dla Cyberbezpieczeństwa*, https://cyberpolicy.nask.pl/spotkanie-partnerow-w-ramach-programu-partnerstwo-dla-cyberbezpieczenstwa/, [accessed: 15.11.2024].

[76] *Czym jest Partnerstwo dla Cyberbezpieczeństwa*, https://www.nask.pl/pl/dzialalnosc/csirt-nask/pdc/3415,Program-Partnerstwo-dla-Cyberbezpieczenstwa.html, [accessed: 17.11.2024].

[77] R. Kupiecki, T. Chłoń, *Sztuczna inteligencja: miecz i tarcza (dez)informacji*, https://itwiz.pl/sztuczna-inteligencja-miecz-i-tarcza-dezinformacji/, [accessed: 9.10.2024].

that create adverse impact on safety or fundamental rights. There is a distinction between two categories of high-risk AI systems: a) used as a safety component of a product or falling under EU health and safety harmonization legislation (e.g. toys, aviation, cars, medical devices, lifts), b) specific areas of a law enforcement, which could be updated as necessary through delegated acts; 3) recognize AI systems with limited risk, as the ones interacting with humans (e.g. chatbots). However, emotion recognition systems, biometric categorization systems, and AI systems that generate or manipulate image, audio or video content, would be subject to a limited set of transparency obligations; 4) acknowledge of AI systems, which present only low or minimal risk. These systems could be developed and used in the EU without conforming to any additional legal obligations[78].

Conclusions

The bias of artificial intelligence concerns the double nature of its functionality. It has started being a cutting-edge technology that constitutes a cyber-engine for social development in many branches of human existence. At the same time, on the one hand it may be used as a tool against societies, posing a real danger. Hostile actors may us it for conducting criminal operations making them highly efficient. On the other hand, AI may become an important defense element, playing the role of analytics, enabling for quick identification and combaing of threats.

It has to be highlighted, the human environment is becoming a cyber-battlefield. AI as the cutting-edge technology may engage lots of different highly advanced cybertools and it is difficult to predict the probable course of its potential development. Thus, there has to be a solution worked out, in order to block weaponizing AI, and instead, keep it as an advancement for societies. The above indicate that the vast use of AI requires global cooperation to create a particular framework for AI, especially by the evolvement of legal regulations and by building a social awareness of the ongoing cognitive warfare. For the sake of being AI-conscious, there are lots of initiatives and projects explaining the pro- and anti-social use of this technology.

[78] T. Madiega, *Artificial Intelligence Act*, https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf, [accessed: 16.11.2024].

The new socio-technological era has been launched within the human cognitive sphere, and probably the only mean to counter the hostile use of AI is a appropriate, as well as responsible use of AI itself.

Bibliography

Allan K., *Google Bards ambition to reach millions of users*, https://aimagazine.com/articles/google-bards-ambition-to-reach-millions-of-users, [accessed: 9.10.2024].

Bodnar D., *What Is CAPTCHA and How Does It Work?*, https://www.avg.com/en/signal/what-is-captcha, [accessed: 12.09.2024].

Bontcheva K., *Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities*, https://edmo.eu/wp-content/uploads/2023/12/Generative-AI--and-Disinformation_-White-Paper-v8.pdf, [accessed: 14.11.2024].

Claverie B., Cluzel F., "*Cognitive Warfare": The Advent of the Concept of "Cognitics" in the Field of Warfare,* https://www.researchgate.net/publication/359991886_Cognitive_Warfare_The_Advent_of_the_ Concept_of_Cognitics_in_the_Field_of_Warfare, [accessed: 25.09.2024].

Fenstermacher L., Uzcha D., *New perspectives on cognitive warfare*, "Proceedings of the SPIE" 2023, No. 12547.

Gergelewicz T.*, Fake Mirroring in Cyberspace as a Disinformation Tool – Doppelgangers and Deepfakes*, "Cybersecurity and Law" 2024, No. 1.

Hung T.C., Hung T.W., *How China's Cognitive Warfare Works: A Frontline Perspective of Taiwan's Anti-Disinformation Wars*, "Journal of Global Security Studies" 2022, No. 4.

Kosinski M., *What is Ransomware?*, https://www.ibm.com/topics/ransomware, [accessed: 20.11.2024].

Koval I., *Prigozhin's troll factories in Russia: What's next?*, www.dw.com/en/prigozhins-troll-factories-in-russia-whats-next/a-66194462, [accessed: 12.11.2024].

Kupiecki R., Chłoń T., *Sztuczna inteligencja: miecz i tarcza (dez)informacji*, https://itwiz.pl/sztuczna-inteligencja-miecz-i-tarcza-dezinformacji/, [accessed: 9.10.2024].

Madiega T., *Artificial Intelligence Act*, https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf, [accessed: 16.11.2024].

Marszycki M., *Jak sztuczna inteligencja wpływa na obszar cyberbezpieczeństwa?*, https://itwiz.pl/jak-sztuczna-inteligencja-wplywa-na-obszar-cyberbezpieczenstwa/, [accessed: 5.11.2024].

Martindale J., *What is Google Bard? Here's how to use this ChatGPT rival*, https://www.digitaltrends.com/computing/how-to-use-google-bard/, [accessed: 23.11.2024].

Maślankiewicz C., *Sztuczna inteligencja a cyberbezpieczeństwo*, https://www.crowe.com/pl/insights/sztuczna-inteligencja-a-cyberbezpieczenstwo, [accessed: 10.11.2024].

Musser M., *How Much Money Could Large Language Models Save Propagandists?*, "Center for Security and Emerging Technology" 2023, No. 9.
Popov P., *Spotkanie Partnerów w ramach programu Partnerstwo dla Cyberbezpieczeństwa*, https://cyberpolicy.nask.pl/spotkanie-partnerow-w-ramach-programu-partnerstwo-dla-cyberbezpieczenstwa/, [accessed: 15.11.2024].
Stoecklin M., Jang J., Kirat D., *DeepLocker: How AI Can Power a Stealthy New Breed of Malware,* "Security Intelligence" 2018, No. 8.
Szczurek T., Grubicka J., Zbroszczyk D., *Borders of Digital Network Freedom in Public Security Space*, "Internal Security" 2020, No. 1.
Walarus J., Danilewicz J., *Bezpieczeństwo danych w czasach AI*, www.ey.com/pl_pl/insights/cybersecurity/bezpieczenstwo-danych-w-czasach-ai-ai-fy24, [accessed: 10.10.2024].
Wrzos M., *Sztuczna inteligencja – szanse i zagrożenia dla bezpieczeństwa w sieci*, https://promocja.wat.edu.pl/cyberwat/sztuczna-inteligencja-szanse-i-zagrozenia-dla-bezpieczenstwa-w-sieci/, [accessed: 11.10.2024].
Zieliński C., *Infodemics and infodemiology: a short history, a long future*, https://iris.paho.org/handle/10665.2/53850, [accessed: 10.11.2024].

* * *

Dwubiegunowość sztucznej inteligencji – szanse i zagrożenia

Streszczenie

Sztuczną inteligencję (SI) można scharakteryzować jako zjawisko bipolarne. Może być wykorzystywana do generowania i rozwijania dobrobytu społeczeństw lub może być wrogim narzędziem w rękach aktorów, których celem jest zniekształcenie poznania społecznego. Istnieje wiele działań cybernetycznych z wykorzystaniem SI, które albo wspierają, albo szkodzą mieszkańcom infosfery. Ze względu na fakt, że współczesne pokolenia funkcjonują w dużej mierze w cyberprzestrzeni, gdzie wartości etyczne i morale są rozproszone i powoli zanikają, znacznie łatwiej jest używać SI do prowadzenia wrogich operacji, takich jak dezinformacja. Cały wachlarz szeroko stosowanego wrogiego wpływu może być ujęty w ramy operacji poznawczych. Instytucje rządowe i pozarządowe dążą do wprowadzenia skutecznych rozwiązań względem wyzwaniem jakie stawia SI. Oczekiwanym rezultatem może być stworzenie i utrzymanie bezpiecznego i godnego zaufania środowiska dla społeczeństw.

Celem tego artykułu jest podkreślenie dwoistej natury SI poprzez wskazanie licznych narzędzi wykorzystywanych zarówno do prowadzenia procesu dezinformacji, jak i budowania odporności społecznej na zagrożenia informacyjne. Rozważania prowadzone są z wykorzystaniem metody analitycznej i krytycznej poprzez przegląd dostępnych źródeł. Pozwala to na dogłębne rozważenie stosowania narzędzi SI, które są wykorzystywane do ataku i obrony. Pod uwagę wzięto głównie dwa rodzaje źródeł. Pierwszym są portale internetowe, które profesjonalnie i kompleksowo omawiają funkcjonowanie SI wraz z wyjaśnianiem podstawowych pojęć związanych z cybersferą. Drugim rodzajem źródeł są raporty, projekty, inicjatywy i oficjalne publikacje, które instytucjonalnie traktują o problemie przeciwdziałania dezinformacji i budowaniu odporności społecznej.

Słowa kluczowe: sztuczna inteligencja, dezinformacja, cyberbezpieczeństwo, wojna kognitywna.