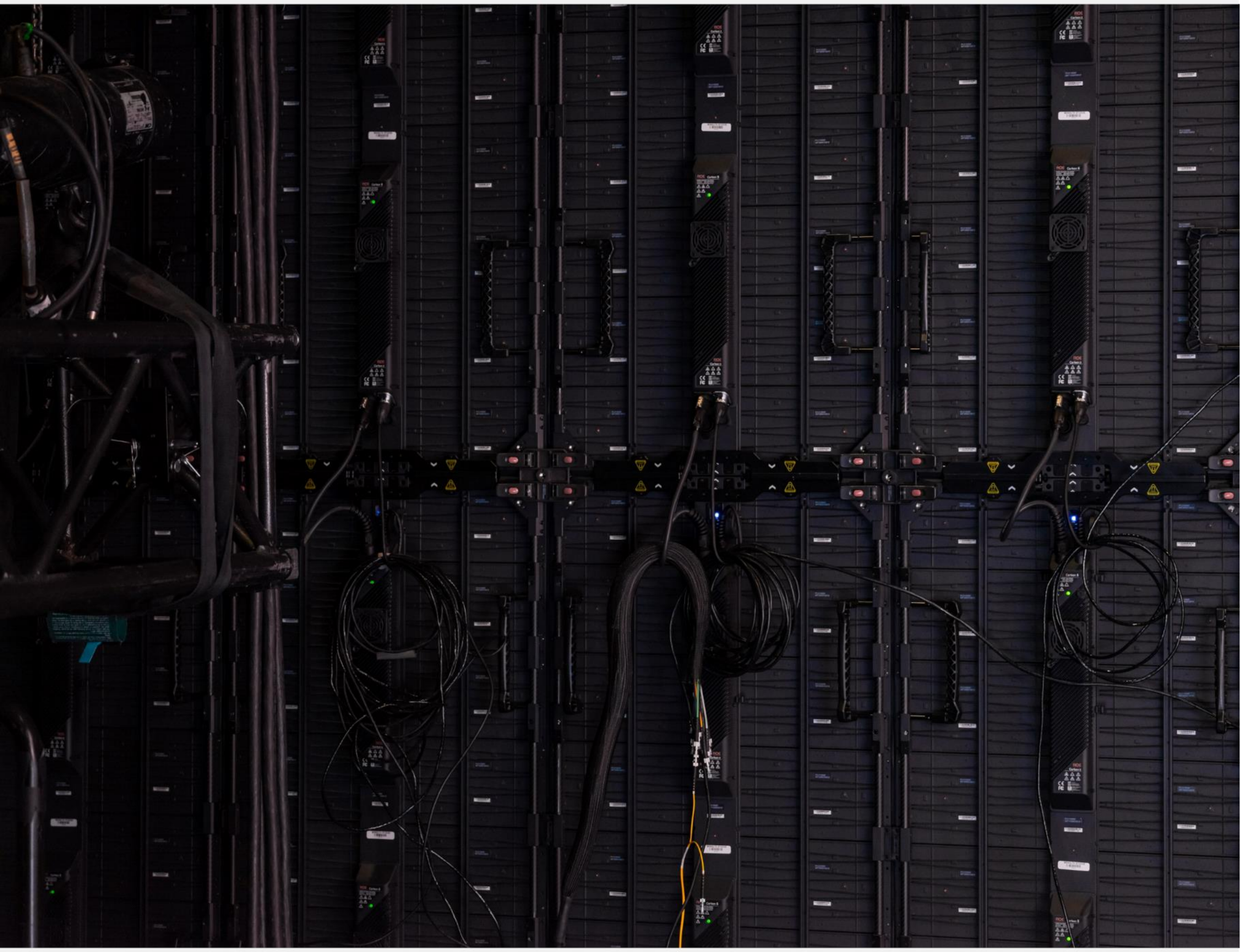
IDR
Romanian Diplomatic Institute



Beyond utopias and dystopias AI's impact on cybersecurity

Claudiu Codreanu

Policy Paper no. 35/2023

GUVERNUL ROMÂNIEI
Ministerul Afacerilor Externe

SUMMARY

- **Latest developments in artificial intelligence (AI) generated both enthusiasm and concerns.** However, the situation is far from setting the foundation for fundamental changes the way society works.
- **Major international actors are working on regulating the development and usage of artificial intelligence**. The European Parliament adopted the AI act, China already implemented a law on generative AI and algorithms, whilst the United States and United Kingdom are continuing their efforts in this regard.
- **Artificial intelligence does not yet have the potential of causing fundamental or apocalyptic changes.** Nevertheless, certain groups of persons and the environment are beginning to be affected.
- **Latest developments in the AI sector have an impact on cybersecurity and cyber operations**. There are three dimensions to this: protecting AI systems from cyberattacks, using AI for optimising cyberattacks, and using AI for strengthening cybersecurity.
- **Cybersecurity policies adopted by states should consider the latest developments and get an update in order to protect AI systems from attacks.** Key factors for enabling a sustainable development of AI and a low-risk implementation are: transparency, the rule of law, and ensuring the freedom of the press and of civil society. Moreover, an important role will also be played by digital and cyber diplomacy.

INTRODUCTION

Over the last year, technological advancements in the field of Artificial Intelligence have captured the attention of journalists, academics, politicians, and the public. The primary reason is that progress in Large Language Models (LLMs) and generative AI has exceeded expectations, outpacing the ability of governments and organisations to implement robust regulations. However, **the currently flawed idea that AI will transform the entire functioning of society and every aspect of human life, akin to electricity, hampers efforts to impose limits on AI usage** (Michel 2023, 13).

Moreover, **2024 will witness several significant elections all over the world, but most notably in the United States and the European Union**. In Europe, elections are slated for Austria, Romania, and the United Kingdom. In the Eurasian and Asian regions, elections will unfold in Georgia, Turkey, South Korea, India, and Taiwan. In Africa, elections will be organised in South Africa, Algeria, Egypt, Ghana, and Tunisia, whilst in South and Central America, they are scheduled in Mexico and Uruguay. These states are not the only ones where people will vote for their representatives and leaders, but it represents a list of the most notable. Latest elections in the US and other major international countries were targeted by various types of information and cyber campaigns, increasing the risk that AI's role will become central next year.

This study is centred on exploring the role and influence of recent advancements in artificial intelligence over the realm of cybersecurity and operations. The research is divided in two sections. The first on delves into an exploration of the latest developments within the field of AI, and the second one will analyse and highlight their impact on cybersecurity. The final section includes a list of recommendations for managing and leveraging AI instruments for cybersecurity.

The research revolves around the following question: How do the recent developments in artificial intelligence exert their influence on the realm of cybersecurity? To delve into this question, the study **evaluates the validity of the following assertion**: Developments in AI confer substantial advantages for both offensive cyber operations and cyber defenders. **These new instruments, malware, and software harnessing AI advancements aid both the attackers and the defenders**. On one hand, government and non-state hackers will be aided by the increased scalability of attacks (making it easier to disseminate across a myriad of devices, e.g., botnets), and by reducing the difficulty of certain septs of a cyberattack (e.g., automatising vulnerability scanning. Moreover, AI tools enhance phishing and disinformation campaigns by generating automated messages and scanning extensive databases. Conversely, the endeavours in cyber defence undertaken by governments, organisations (public institutions, businesses etc.), cybersecurity firms, and individual actors will also be aided. In this context, AI's most noteworthy contributions are automating vulnerability scans, deploying tools capable of real-time network activity monitoring and autonomous pattern recognition, and bolstering the capacity to counter cyber threats posed by malicious actors.

BACKGROUND AND CONCEPTS

The term "artificial intelligence" was introduced by American researcher John McCarthy in 1955 (Bjola 2022, 79; Bonfanti & Kohler 2020, 1). **Artificial intelligence refers to the action through which computers process large amounts of data using sophisticated algorithms that simulate human behaviour and reasoning** (Bjola 2022, 79). Thus, AI represents a general-purpose technology aimed at improving the speed, accuracy, and magnitude of automated decision-making processes. Artificial intelligence will contribute to enhancing human performance in activities such as prediction, optimisation, recognition, and decision-making, including in strategic or military context (Maas 2019, 285-286). Thus, **AI is described as a technology that enables the accomplishment of various tasks or the improvement of other technologies and systems, given its applicability across different domains, including objectives related to cybersecurity** (Bonfanti & Kohler 2020, 1).

However, technologies referred to as 'artificial intelligence' have only limited capacity to reproduce human intelligence, mimicking only certain aspects of it (Michel 2023, 6). **So far, no machine (e.g., system, robot, software etc.) has passed the Turing Test.** The test, proposed by British researcher Alan Turing in 1950, requires a machine to meet two conditions: to respond appropriately to variations in human dialogue and to exhibit intentions and a personality as close to human as possible (Bjola 2022, 79). Thus, there have been calls from the academic and technological community to replace the umbrella term "artificial intelligence" with precise and individualised terminology that best describes the specific technical capabilities of the respective system (Michel 2023, 15).

In addition to this, **machine learning consists of large sets of data, learning algorithms, and computational power for training these algorithms**. A significant portion of recent AI advancements comes from **deep learning, which utilises deep neural networks.** Even though the networks comprise numerous layers of data, they still cannot achieve one hundred percent accuracy, often making incorrect predictions. Furthermore, **large language models (LLMs)** still produce unpredictable errors and false texts despite being trained on vast amounts of data (Bonfanti & Kohler 2020, 1; Bezombes, Brunessauax & Cadzow 2023, 10; Michel 2023, 18).

Moreover, **a deepfake represents a video, photograph, or audio material generated by complex algorithmic systems (machine learning or deep learning mechanisms) with limited or no human oversight**. Deepfakes amplify cybercrime threats, particularly concerning identity theft and bypassing biometric authentication systems. Since they create something new rather than just imitating reality, deepfakes have the potential to deceive people, leading them to believe that the material in question represents reality (Bray, Johnson & Kleinberg 2023).

Overt the last year, **the prevalence of using AI text generation platforms such as ChatGPT or Google Bard, and image generation systems such as Stable Diffusion, Midjourney, or Dall-E, has brought the rapid development of AI technologies into the public and political spotlight**. One of the first researchers and industry representatives to publicly voice concerns about AI was **Geoffrey Hinton**, a pioneer of artificial intelligence systems. Hinton stepped down from his position at Google and publicly warned about the dangers of recent advancements in AI (Knight 2023a). During the same period, **Eliezer Yudkowsky**, one of the most important researchers in the field of AI, emphasized in a 2023 article for *TIME* magazine that, under current conditions, if an overly powerful AI systems were to be developed, it is expected that "every single member of the human species and all biological life on Earth dies shortly thereafter". In addition to this, researchers in biological and nuclear security have pointed out that information provided by generative AI could assist terrorist groups in the creation of biological weapons, even though the risks are currently low (Service 2023).

In this context, several representatives from the tech and AI industry signed an **open letter in March 2023, calling for a 6-month suspension of AI lab activities that train systems more powerful than GPT-4**, advocating for industry regulation and monitoring of powerful AI systems (Yudkowsky 2023; Milmo & Stacey 2023). During the same period, British Prime Minister Rishi Sunak announced that the UK would host a **global summit on safety in artificial intelligence**, aimed at "like-minded states" (Milmo & Stacey 2023). More than this, G7 agreed to establish the "**Hiroshima AI process**", a forum intended to discuss current issues related to recent advancements (Milmo & Stacey 2023).

In July 2023, the United Nations organised the AI for Good conference in Geneva, where nine of the most advanced humanoid robots were brought to speak about AI, but it is unclear whether their responses were pre-programmed (Ferguson 2023). Later, UN Secretary-

General António Guterres warned during the first session of the Security Council on AI in July 2023 that the use of AI systems could lead to a "horrific" amount of death and destruction, calling for the formation of a new intergovernmental panel (Milmo & Farah 2023).

During the same period, Google, Microsoft, Anthropic, and OpenAI established a **body to discuss the regulation of AI tech development**. Its objectives include discussing security risks with the academic and political communities, developing standards for evaluation and releasing advanced AI models, as well as promoting the use of AI to combat climate change or for medical progress. Forum members, long with Amazon, Meta, and other relevant actors, reached an agreement with the White House to adopt safety assurances in the development and promotion of AI systems (e.g., introducing labels to mark content created by AI). However, there are also critics of the initiative, as the tech industry has a rich history of failing to uphold commitments to self-regulation. (Milmo 2023a; Benson 2023)

THE EFFECTS OF THE LATEST DEVELOPMENTS OF AI ARE ALREADY VISIBLE

The current impact on human life and the planet

Until the destruction of the human species, **the usage of AI systems has already produced disastrous effects on certain individuals and groups, especially those from marginalised or discriminated communities.** For instance, in November 2022, a young African American man was arrested in Georgia (US state), being detained for six days after a facial recognition software erroneously identified him as the perpetrator of a series of robberies in Louisiana (Malik 2023). The incident highlighted that facial recognition software is not calibrated and trained with sufficient data to accurately identify people of colour, and that the algorithms used replicate human biases. Thus, the usage of AI in activities for which it was not technically designed without responsible human oversight or in contexts lacking adequate regulations can pose significant risks for members of vulnerable groups (Michel 2023, 11).

Another issue that has been relatively overlooked due to apocalyptic warnings is the **impact of AI systems on the environment**. Just as cryptocurrency production consumes a massive amount of electricity, the cloud servers on which AI relies and all the systems that maintain them will consume an increasingly larger amount of electricity, as well as the manufacturing of chips, and so on. For instance, mining Bitcoin alone consumes more electricity than Norway and Ukraine combined (Stokel-Walker 2023).

Moreover, there is also a risk of an impending arms race in AI weapon systems (Maas 2019, 286). Likewise, the same area includes the issue of implementing **fully autonomous weapon systems**, which operate without external commands throughout the entire process of target identification, tracking, selections, and attack (Rosert & Sauer 2021). Nonetheless, there is still no clear international regulation regarding these systems.

Using AI systems for disinformation

The usage of manipulated or counterfeit images and videos within electoral campaigns is not a recent development. For instance, Ron DeSantis, the Governor of Florida and candidate in the primaries of the Republic Party, disseminated AI generated images depicting Donald Trump embracing Anthony Fauci, the former White House Chief Medical Advisor. Fauci has faced disdain from American conservatives due to his policies aimed at restricting the spread of SARS-CoV-2 and advocating for the COVID-19 vaccine. Similarly, several deepfakes featuring President Joe Biden have surfaced, wherein he 'announces' that American citizens will be drafter to fight in the Russo-Ukrainian war. (Milmo 2023b)

Thus far, artificially generated videos of this nature have not constituted turning points and have not fully harnessed the potential offered by AI generators. However, this landscape might transform in 2024, precisely during a critical juncture – amidst electoral campaigns unfolding across various nations. **In 2024, over 70 states are slated to conduct national and/or regional elections, involving a population of more than 2 billion people, predominantly situated in the Global South** (Madung 2023). A major risk arises from generative AI, which could produce credible texts, audios, and videos for disinformation campaigns during electoral processes. (Milmo & Stacey 2023). For instance, investigative efforts pursued by several African journalists have brought light to instances where the Twitter algorithm (now known as X) has been systematically and seamlessly manipulated to disseminate propaganda and misinformation in Keyna and Nigeria during recent electoral campaigns (Madung 2023).

AI will facilitate both the processes of producing disinformation and disseminating it over large groups of people. A 2023 study published in the *Science Advances* journal highlighted that Tweeter (X) content generated by ChatGPT-3 (the predecessor of GPT-4) demonstrated superior capabilities in both informing and misleading people, compared to human-crafter tweets (Spitale, Biller-Andorno & Germani 2023). Thus, AI systems possess the potential to substantially enhance the effectiveness of disinformation campaigns, improving the ability of external actors to interfere in the domestic affairs of other states (Kenny 2023, 224).

As AI-based tech advance and become widely accessible, their usage in disinformation campaigns will become increasingly pervasive (Kenny 2023, 224). Moreover, **disinformation will become much more precisely targeted at specific groups of people or even individuals using data collected from the online sphere** (Benson 2023). AI systems enable actors to create a significant number of fake accounts on social media, employing machine learning mechanisms to construct elaborate false profiles, including details, posts and even photographs. These accounts can be managed entirely through bot networks giving the impression that a substantial and credible number of citizens share the same viewpoint (Kenny 2023, 235).

These new AI-based tools could enable a much larger number of individuals to generate much more credible images, messages, and videos, compared to the technology available 2-3 years ago. This advancement facilitates the generation of false or propagandistic materials. Nevertheless, the precise impact of these emerging platforms on the conduct of the 2024 electoral campaigns remains ambiguous. Most probably, the elections will be affected, but this does not suggest that AI tools will generate fundamental changes in their processes or in the preferences of the majority of voters, especially compared to the elections organised after 2015. What has become clear is that the production of false texts and images has become nearly cost-free, effortless, accessible, and does not require advanced technological skills. Furthermore, the materials produced appear credible, and distinguishing between artificially generated and genuine content has become increasingly challenging for both humans and technologies. More than this, the scale of spreading disinformation increases, as bot networks that automatically generate artificial and credible content can now be created at a much larger scale compared to what was possible just a few years ago.

Efforts for national and international regulations on AI

The majority of discussion regarding the recent advancements in AI have focused on the distant potential of technology bringing humanity’s demise, rather than addressing the issues that are already unfolding: the proliferation of mass surveillance, discrimination based on various criteria, or the spread of disinformation campaigns (Bhuiyan & Robins-Early 2023). Nevertheless, several international actors have begun attempts to regulate AI-based systems.

In June 2023, **The European Parliament adopted the AI act**. The legislation proposes limits and restrictions concerning the collection of biometric data and imposes bans on high-risk AI applications, such as facial recognition technology and predictive policing systems (Bhuiyan & Robins-Early 2023).

The new legislation could be implemented by the end of the year, with high hopes that it will generate a 'Brussels effect', setting a global standard. The AI act classifies artificial intelligence systems into five categories. Firstly, **systems exhibiting unacceptable risks will be prohibited. They include applications manipulating individuals, those used by governments for population classification based on personal or socioeconomic criteria. This category also encompasses real-time facial recognition and predictive policing systems**. Secondly, high-risk systems, such as those employed in critical infrastructures, education, and border control, will be subjected to rigorous monitoring. Systems posing limited risks must comply with a set of minimum transparency requirements, and users interacting with artificial intelligence platforms (e.g., ChatGPT or Midjourney) must receive explicit warning that they are engaging with an AI. Lastly, systems with minimal or no risks (the most common ones), such as those used in email spam filters or in video games, will be exempt from additional obligations. Furthermore, developers integrating artificial intelligence systems in a domain or activity must ensure human oversight. (Milmo 2023c)

Likewise, **the United States, the United Kingdom, and China have enacted or are planning to formulate their own legislation on AI** (Milmo 2023c). Efforts are still ongoing in a incipient stage in the US. The lack of comprehensive regulations in the US has eld to a controversial process wherein major AI industry players are proposing their own regulatory frameworks, tailored to minimise disruption to their current operations (Bhuiyan & Robins-Early 2023).

In parallel, **China has embarked on the implementation of one of the world's first regulatory frameworks concerning algorithms and artificial intelligence**. The legislation encompasses the establishment of a registry for algorithms, serving as a central database where authorities can find information regarding algorithms, the source for data used for training, and potential security risks. Moreover, these regulations mandate that vendors of algorithms obtain explicit consent from individuals in instances where their images and voices are manipulated for the creation of videos, such as deepfakes. (O'Shaughnessy & Sheehan 2023).

In April 2023, China enacted a legislation regulating content-generating AI systems, including prohibitions to prevent discrimination against specific groups and imposing limits or legal liability of developers. However, the act also aims to reinforce China's authoritarian regime. The regulatory measures outline conditions for preserving individuals' privacy and avoiding profiling based on online activity, along with requirements for transparency and accountability. Moreover, **content generated by AI must "reflect the fundamental Socialist values" of China, prohibiting material that "undermines the authority of the state"**. More than this, these regulations only apply to the private sector. (O'Shaughnessy 2023).

THE IMPACT OF AI SYSTEMS ON CYBERSECURITY

In a report issued by the European Union Agency for Cybersecurity (ENISA) (Bezombes, Brunessauax & Cadzow 2023, 10), three dimensions characterise the relationship between artificial intelligence and cybersecurity. The first dimension refers to the **cybersecurity of AI** – the safety and vulnerabilities of AI models and algorithms. The second dimension encompasses the **implementation of AI in cybersecurity** – AI tools and means for enhancing cybersecurity measures and activities. The third dimensions centres on the **malicious use of AI** – maliciously or adversarial usage of AI to deploy sophisticated forms of cyberattacks.

Cyberattacks mediated by novel AI technologies

According to a 2023 study conducted by the cybersecurity firm *Imperva*, 47.4% of all web traffic recorded in 2022 was attributed to automated traffic (bots). **The fact that the majority of internet traffic comes from bots generated new concerns, culminating in the emergency of the 'dead internet theory'**. No one can be entirely sure of what they see online is real, or that the people they interact with online are real. Despite efforts by social media companies to address the issue of using bots for comments, reactions, advertisements etc., the concern that AI-based online systems might eventually misclassify bot behaviour as genuine while questioning the authenticity of human actions (Tiffany 2021).

Autonomous systems hold the potential to yield significant advantages in cyber operations, particularly in the rapid and extensive execution of repetitive and less intricate tasks (Perez 2023, 187). Therefore, **AI has the capacity to amplify the volume of cyber threats and alter their characteristics, introducing novel and unexplored threats** (Bonfanti & Kohler 2020, 2). These autonomous cyber capabilities can be deployed across various phases of a cyberattack, encompassing reconnaissance, infiltration, and command and control (Perez 2023, 188). Cyberattacks do not unfold entirely under a direct human oversight throughout all operational stages. The Stuxnet malware, employed by the United States and Israel to disrupt Iran's nuclear program, operated autonomously, especially when it needed to infiltrate an air-gapped network while targeting Iranian nuclear facilities (Perez 2023, 204). Furthermore, diminishing financial barriers and technological expertise needed for cyberattacks is expected to render cyber tools accessible to a broader array of actors than in previous years (Bonfanti & Kohler 2020, 2).

Artificial intelligence holds the potential to elevate the level of sophistication involved in the writing and operating of malware. Malware augmented by AI capabilities possesses the ability to adapt and respond autonomously to real-time changes in the behaviour of the target, thereby eluding cybersecurity measures. Autonomous malware is capable of learning from its surroundings to avoid detection, identify and infiltrate new targets, locate valuable data, and facilitate new cyberattacks. Notably, researchers at IBM developed AI-driven malware as early as 2018, underscoring that this kind of technology is not entirely new (Bonfanti & Kohler 2020, 3).

Concurrently, as anticipated, **cybercriminals developed their own clones of ChatGPT**, asserting that these platforms can aid hackers to write codes for malware or better phishing emails. These clones appear to involve two large language models (LLMs) that emulate the functionalities of ChatGPT and Google Bard, generating texts in response to user requests. However, these systems themselves may be deceptive, as they may not perform as claimed. Nevertheless, their role is currently limited. Since the onset of summer in 2023, both the Federal Bureau of Investigation (FBI) and Europol have issued cautions, announcing that cybercriminals are exploring methods to integrate generative AI into their illicit activities. **Large language models could expedite fraud endeavours, identity theft, and social engineering, while also potentially refining the quality of phishing texts composed in English**. (Burgess 2023)

Attacks against AI systems

In addition to the usage of AI to enhance cyber operations, AI systems themselves will become significant targets of cyberattacks, as they will be implemented in important societal, governmental, economic, and security areas (Whyte 2023, 313). Thus, **ensuring the security of cyber systems relying on AI technologies becomes paramount** (Bonfanti & Kohler 2020, 2).

ChatGPT and its counterparts have undergone multiple modifications to prevent their exploitation for generating personal information, hate speech, or instructions related to illicit activities such as explosions. However, a study conducted by researchers from Carnegie Mellon University in July 2023 revealed the specific sets of text inputs can circumvent existing defensive measures. These attacks are referred to as **adversarial attacks**, which involve the introduction of carefully crafted texts aiming to gradually persuade the chatbot to evade the limitations imposed by its developers. **Adversarial attacks may encompass techniques such as data poisoning, involving the injection of misleading data into the AI's dataset, leading the learning algorithm to make errors, or the creation of adversarial examples – materials designed to devise and mislead, causing misclassification** (Knight 2023b; Bonfanti & Kohler 2020, 2).

There is no consensus regarding the modes of interaction between autonomy and cyber, as well as the application of international law in this context. For instance, while **autonomous weapons** are classified as physical entities, cyber systems are immaterial, despite being generated by physical entities. Likewise, another challenge will emerge, and maybe already is relevant, in the realm of **autonomous vehicles**. Alongside artificial intelligence, the Internet of Things, and other systems intertwining the physical and cyber domains, autonomous vehicles introduce novel security risks, as they are vulnerable to cyberattacks that have the potential to compromise their optimal functionality. Manipulating road signs or deploying **adversarial attacks** (modifications that confound machine learning systems) could pose substantial cybersecurity threats and jeopardise public safety (Perez 2023, 186; Blumenthal & Csernatoni 2022, 2-4).

The impact on cybersecurity

AI tools have the potential to enhance existing software for malware detection and identification, but they should serve as a complement to traditional mechanisms rather than completely replacing them (Bonfanti & Kohler 2020, 3). **Using AI to streamline the processes of analysis and intrusion in cyber operations provides significant opportounist for attackers, akin to tools used for analysing large volumes of data** (Whyte 2023, 309). However, **cyber defence can also benefit from AI based tools**. Recent advancements in AI not only provide advantages to cyber attackers but also to cyber defenders, representing a double-edged sword for cybersecurity. Therefore, the validity of the assertion stated at the beginning of this study is conform after this research, as latest developments in AI generate serios benefits for both offensive and defensive cyber operations. Artificial intelligence represents *only a tool*, and the impact of it depends on how the systems is designed and how it is utilised. Nonetheless, AI platforms may primarily favour cybercrime. Regarding the actions of state actors, artificial intelligence may not necessarily play a fundamental role, but it will expedite and optimise certain stages of cyber operations. Concerning cyber defence, AI tools can reinforce existing programs and human efforts in identifying intrusions and vulnerabilities, but they cannot entirely replace them.

AI has the potential to enhance **capabilities in the detection, analysis, and prevention** of cyber threats, particularly in areas such as spam, phishing, and malware detection (Bonfanti & Kohler 2020, 3). Thus, this could lead to an increase in the speed of identifying malicious cyber activities and improve the ability to scan a wider range of data. Moreover, AI could be harnessed for **automated vulnerability scanning and testing within a system or network** (Bonfanti & Kohler 2020, 3). Concurrently, AI tools may play a pivotal role in learning and discerning patterns and anomalies within a system or network, although the precision of these applications remains a subject of scrutiny.

AI holds the potential to be deployed for **monitoring online environments and social media platforms**, discerning intricate patterns and indicators of nefarious campaigns and activities orchestrated by bot networks (Bonfanti & Kohler 2020, 3). AI systems could be harnessed to identify mass-created false accounts and misinformation messages propagated through coordinated campaigns, with machine learning and pattern recognition gaining a substantial role in these initiatives. Nevertheless, the application of AI systems in this domain is not without its constraints, necessitating a rigorous human supervision to prevent the erroneous classification of authentic content as disinformation.

Artificial intelligence enables the execution of significantly more complex and extensive cyber operations compared to previous periods. Equally important is the issue of cyberattacks against AI systems, organisations, and platforms utilising AI. Nevertheless, **the impact of artificial intelligence on cybersecurity processes will not fundamentally alter the dynamics**. Certain variants of AI tools are already integrated into the majority of offensive and defensive cyber operations, including antivirus and firewall systems. However, these involve significantly more intricate, intelligent systems, capable of autonomous decision-making without human intervention or monitoring.

CONCLUSIONS AND RECOMMENDATIONS

There can be no prospect for institutions or states with deeply entrenched structural inequalities to reconfigure their entire value system and responsibly implement AI-based systems. In the absence of broader measures safeguarding the right of the entire society within these AI systems operate, the ethically sound implementation of artificial intelligence is highly improbable. States that fail to ensure equal rights for all citizens, (e.g. restrictions on freedom of speech, limiting the rights of women, the LGBTQ+ groups or ethnic minorities) have few chances to implement public AI initiatives that could be ethical, impartial or fair (Michel 2023, 33).

Key factors facilitating sustainable development of artificial intelligence and its low-risk deployment involve institutional transparency, respecting the rule of law, and safeguarding civil society and media freedom. These components serve as fundamental prerequisites for the ethical implementation of AI-based systems (Michel 2023, 25). Therefore, legislation regarding AI must encompass provisions that guarantee transparency, enabling scrutiny by civil society and media to assess the methods employed in the development and application of these systems (O'Shaughnessy 2023).

Ensuring human oversight of artificial intelligence, its precision, explainability, as well as ensuring transparency and safety, are essential pillars supporting the cybersecurity of AI and systems relying on AI. These elements represent fundamental criteria for shaping any policies and strategies in this domain (Bezombes, Brunessauax &

Cadzow 2023, 11), as they are crucial for preventing abuses, ensuring the proper functioning of systems, and avoiding the replication of biases and human errors. Therefore, **human oversight** of AI systems in specific environments is imperative, not only to ensure their smooth operation but also to identify unusual behaviours resulting from potential cyberattacks (Bezombes, Brunessauax & Cadzow 2023, 10). Monitoring AI-based systems is equally crucial, especially in the field of cybersecurity, aiming to prevent attacks that could bypass detection capabilities of AI tools or to avoid misidentifications of behaviours and software as malicious.

Databases and their corresponding infrastructures should be constructed and administered by adopting strict privacy principles. The value of datasets for AI technologies is dependent on their applicability and capabilities of exploiting the data. The aggregation and dissemination of an ever-expanding volume of data produce vulnerabilities that privacy regulations may not swiftly address. Nevertheless, these regulations are susceptible to modification, particularly in the event of an authoritarian regime assuming power, potentially resulting in the misuse of data previously safeguarded by legal provisions. Furthermore, concern persists regarding the potential exploitation of extensive datasets by private entities (Michel 2023, 17-19).

Human-centric models should be embraced in the utilisation, implementation, and regulation of AI. Policies regarding artificial intelligence ought to emphasise the assurance of personal privacy, the safeguarding of personal data, and the protection of minority groups rights, as well as considering the environmental ramifications of these technologies. Nonetheless, ethical policies surrounding artificial intelligence are contingent upon the sociopolitical frameworks and values upheld by the states and organisations proposing them (Michel 2023, 34).

Furthermore, **cybersecurity policies** need to be reassessed and adapted to incorporate the influence of the latest developments in artificial intelligence. Essentially, it is crucial to strengthen policies aimed at ensuring the cybersecurity of artificial intelligence systems. Moreover, it is critical to swiftly implement policies tailored to counteract the new episodes of disinformation in 2024, before the commencement of electoral campaigns. Thus, **current strategies and policies must be updated to cope with new developments, considering the increasing frequency and intensity of cyberattacks, especially those perpetrated by non-state actors and cybercriminals.** Recent advancements in artificial intelligence should be integrated into international discussions within the UN framework to regulate the activities of state actors in cyberspace.

Digital and cyber diplomacy can contribute to internationally deal with the ongoing developments in the fields of artificial intelligence and cybersecurity. Firstly, there is a pressing need to expedite the current UN discussions concerning the regulation of cyberspace, given the looming risk of AI systems amplifying global cyberattacks. Secondly, UN discussions regarding AI systems should consider cybersecurity dimensions, even though the priorities should be the systems that exert a significant influence on human life, such as autonomous weapon systems and social control systems. Thirdly, digital diplomacy becomes an important mean for countering online disinformation campaigns, wherein AI systems can play a substantial role for detecting campaigns of this nature. **However, discussions and policies regarding artificial intelligence must extend beyond dystopian and apocalyptic narratives, as the technology possesses the potential to strengthen cyber security or accelerated advancements in medicine and science, for instance.**

The ways in which the newest AI systems will influence the export of digital authoritarianism and technologies enabling digital repression remains to be seen. These developments will be followed by a study that will appear in the **next number of RDI's journal, *România Occidentală*, to be published in December 2023.** The article, written by Sînziana Dumitrescu and Claudiu Codreanu, will follow China's export of digital authoritarianism to African countries.

REFERENCES

Benson, T. 2023. This disinformation is just for you. *Wired*. https://www.wired.com/story/generative-ai-custom-disinformation/.

Bezombes, P., Brunessauax, S., & Cadzow, S. 2023. *Cybersecurity of AI and standardisation*. ENISA. https://www.enisa.europa.eu/publications/cybersecurity-of-ai-and-standardisation.

Bhuiyan, J., & Robins-Early, N. 2023. The EU is leading the way on ai laws. The US is still playing catch-up. *The Guardian* https://www.theguardian.com/technology/2023/jun/13/artificial-intelligence-us-regulation.

Bjola, C. 2022. AI for development: Implications for theory and practice. *Oxford Development Studies*, 50(1), 78-90. https://doi.org/10.1080/13600818.2021.1960960.

Blumenthal, M. S., & Csernatoni, R. 2022. *Computers on wheels: Automated vehicles and cybersecurity risks in Europe*. EU Cyber Direct. https://eucyberdirect.eu/research/computers-on-wheels-automated-vehicles-and-cybersecurity-risks-in-europe.

Bonfanti, M. E., & Kohler, K. 2020. Artificial Intelligence for cybersecurity. *CSS Analyses in Security Policy*, 265. https://doi.org/10.3929/ethz-b-000417116.

Bray, S. D., Johnson, S. D., & Kleinberg, B. 2023. Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity*, 9(1). https://doi.org/10.1093/cybsec/tyad011.

Burgess, M. 2023. Criminals have created their own ChatGPT clones. *Wired*. https://www.wired.com/story/chatgpt-scams-fraudgpt-wormgpt-crime/.

Ferguson, D. 2023. Robots say they have no plans to steal jobs or rebel against human. *The Guardian*. https://www.theguardian.com/technology/2023/jul/08/robots-say-no-plans-steal-jobs-rebel-against-humans.

Imperva. 2023. 2023 Bad bot report. https://www.imperva.com/resources/resource-library/reports/2023-imperva-bad-bot-report/

Kenny, J. 2023. Advanced artificial intelligence techniques and the principle of non-intervention in the context of electoral interference. În F. Cristiano, D. Broeders, F. Delerue, F. Douzet, & A. Gery (ed.), *Artificial Intelligence and International Conflict in Cyberspace* (pp. 223-257). Routledge. https://doi.org/10.4324/9781003284093.

Knight W. 2023b. A new attack impacts major AI chatbots—and no one knows how to stop it. *Wired*. https://www.wired.com/story/ai-adversarial-attacks/.

Knight, W. 2023a. What really made Geoffrey Hinton into an AI doomer. *Wired*. https://www.wired.com/story/geoffrey-hinton-ai-chatgpt-dangers/.

Maas, M. M. 2019. How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons. *Contemporary Security Policy*, 40(3), 285-311. https://doi.org/10.1080/13523260.2019.1576464.

Madung, O. 2023. AI hysteria is a distraction: algorithms already sow disinformation in Africa. *The Guardian*. https://www.theguardian.com/global-development/2023/aug/09/ai-chatgpt-doomerism-threat-already-here-big-tech-algorithms-sow-disinformation.

Malik, K. 2023. Fantasy fears about AI are obscuring how we already abuse machine intelligence. *The Guardian*. https://www.theguardian.com/commentisfree/2023/jun/11/big-tech-warns-of-threat-from-ai-but-the-real-danger-is-the-people-behind-it.

Michel, A. H. 2023. *Recalibrating assumptions on AI*. Chatham House. https://www.chathamhouse.org/2023/04/recalibrating-assumptions-ai.

Milmo, D. 2023a. Google, Microsoft, OpenAI and startup form body to regulate AI development. *The Guardian*. https://www.theguardian.com/technology/2023/jul/26/google-microsoft-openai-anthropic-ai-frontier-model-forum.

Milmo, D. 2023b. Doctored Sunak picture is just latest in string of political deepfakes. *The Guardian*. https://www.theguardian.com/technology/2023/aug/03/doctored-sunak-picture-is-just-latest-in-string-of-political-deepfakes.

Milmo, D. 2023c. TechScape: Can the EU bring law and order to AI?. *The Guardian*. https://www.theguardian.com/technology/2023/jun/27/techscape-european-union-ai.

Milmo, D., & Farah, H. 2023. Malicious use of AI could cause 'unimaginable' damage, says UN boss. *The Guardian*. https://www.theguardian.com/technology/2023/jul/18/malicious-use-of-ai-could-cause-huge-damage-says-un-boss.

Milmo, D., & Stacey, K. 2023. Rishi Sunak's AI summit: what is its aim, and is it really necessary?. *The Guardian*. https://www.theguardian.com/technology/2023/jun/09/rishi-sunak-ai-summit-what-is-its-aim-and-is-it-really-necessary.

O'Shaughnessy, M. & Sheenan, M. 2023. *Lessons from the world's two experiments in AI governance*. Carnegie. https://carnegieendowment.org/2023/02/14/lessons-from-world-s-two-experiments-in-ai-governance-pub-89035.

O'Shaughnessy, M. 2023. *What a Chinese regulation proposal reveals about ai and democratic values*. Carnegie. https://carnegieendowment.org/2023/05/16/what-chinese-regulation-proposal-reveals-about-ai-and-democratic-values-pub-89766.

Perez, L. 2023. Is Stuxnet the next Skynet? Autonomous cyber capabilities as lethal autonomous weapons systems. În F. Cristiano, D. Broeders, F. Delerue, F. Douzet, & A. Gery (ed.), *Artificial Intelligence and International Conflict in Cyberspace* (pp. 186-222). Routledge. https://doi.org/10.4324/9781003284093.

Rosert, E., & Sauer, F. 2021. How (not) to stop the killer robots: A comparative analysis of humanitarian disarmament campaign strategies. *Contemporary Security Policy*, 42(1), 4-29. https://doi.org/10.1080/13523260.2020.1771508.

Service, R. 2023. „Could chatbots help devise the next pandemic virus?". *Science*, 380(6651), 1211. https://www.science.org/doi/epdf/10.1126/science.adj3377.

Spitale, G., Biller-Andorno, N. & Germani, F. 2023. „AI model GPT-3 (dis)informs us better than humans". *Science Advances*, 9(26), 1-9. https://www.science.org/doi/epdf/10.1126/sciadv.adh1850.

Stokel-Walker, C. 2023. TechScape: Turns out there's another problem with AI – its environmental toll. *The Guardian*. https://www.theguardian.com/technology/2023/aug/01/techscape-environment-cost-ai-artificial-intelligence.

Tiffany, K. 2021. Maybe you missed it, but the Internet 'died' five years ago. *The Atlantic*. https://www.theatlantic.com/technology/archive/2021/08/dead-internet-theory-wrong-but-feels-true/619937/.

Whyte, C. 2023. Learning to trust Skynet: Interfacing with artificial intelligence in cyberspace. *Contemporary Security Policy*, 44(2), 308-344. https://doi.org/10.1080/13523260.2023.2180882.

Yudkowsky, E. 2023. Pausing AI developments isn't enough. We need to shut it all down. *TIME*. https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/.

IDR

Institutul Diplomatic Român

Our mission. The Romanian Diplomatic Institute (RDI) has the mission to make a substantial contribution to increasing the quality of Romanian diplomacy through training, further education, research, the development of critical and strategic thinking and international networking. A good foreign policy serves as a beneficial domestic policy.

Guiding principles: human resource development, professionalism, respect and dialogue, and responsibility for the community.

Based on the founding legal attributions of the RDI, the further development of the Institute is carried out, according to the needs identified in the MFA, along the following four directions:

- Training and further education of diplomats and other trainees;
- Deepening the research and expertise dimension on regional and functional issues;
- Operating the RDI as a think-tank of the MFA;
- Integration of the RDI into an international network of similar relevant institutes.

Author: Claudiu Codreanu
Cover design: Claudiu Codreanu
Editors: Dragoș C. Mateescu, Mihai Constantinescu
Cover photo: https://unsplash.com/photos/a-close-up-of-a-server-room-3Nwt6w-KU3E

The Romanian Diplomatic Institute

https://www.idr.ro/en/ | secretariat@idr.ro

Bd. Primăverii nr. 17, sector 1, Bucharest, postal code 011972