DOI: 10.53477/1842-9904-23-21

# THREAT ACTORS SEEKING TO EXPLOIT AI CAPABILITIES. TYPES AND THEIR GOALS

***Petru-Dan KOVACI****

*Artificial Intelligence's advancement has led to ethical and privacy concerns due to the ability of algorithms to mine personal data and conduct surveillance on a large scale. The influence of Artificial Intelligence (AI) in cybersecurity extends beyond private entities and individuals. In this scenario, it is important to highlight the threatening actors and their objectives in order to counter AI-based cyber threats. We are dealing with different entities, each having its own interests: cybercriminals are interested in making profit, terrorist groups cause ideological violence and nation-states target geopolitical influence. Moreover, they are increasingly interested in developing AI-driven cyber warfare capabilities for either attacking their enemies or enhancing their defence measures against such attacks. This trend will most likely aggravate the global cyber arms race as countries compete to surpass each other in the development and deployment of AI-driven cyber capabilities. Consequently, it has become crucial for organizations and individuals to understand how AI affects cybersecurity and, thus, adapt their strategies accordingly. Traditional defences must be supplemented with AI-powered tools and techniques to stay ahead of the curve, while security experts must continually update their skills and expertise to cope with the changing threat landscape.*

***Keywords:*** *AI; cybercriminal; cyberattack; deepfake; social engineering; threat actor.*

**** Petru-Dan KOVACI is PhD Candidate within "Carol I" National Defence University, Bucharest, Romania, and also Security Consultant within OMEGA Trust SRL.***
***E-mail: kpetru112@yahoo.com***

## Introduction

Every conceivable aspect of life is gradually incorporating artificial intelligence, including social networks, independent bus companies, retail stores, and cybersecurity firms. Although AI enhances cybersecurity, it also gives hackers an upper hand in executing complex attacks. Chatbot (a computer program that simulates human conversation with an end user) usage is on the rise. One of the biggest risks in cyberattacks is social reengineering; if AI can be used to teach bots to interact amicably with humans, then the same technology could be used to conduct cyberattacks.

Anyone in the world can easily connect to a video or picture they never captured thanks to deepfakes technology. AI is currently the primary technology that generates deep fraud. AI-based deepfake technology expands possibilities but also makes it easier for bad actors to manipulate and interfere. Deepfake's main component involves machine learning, which allows it to create deepfakes more quickly and at a lower cost.

Automating image and audio processing expands a country's state surveillance capabilities by enabling the massive gathering, processing, and use of intelligence data for a variety of objectives, including the stifling of dissent.

Both state and non-state cybercriminals continue the search for ways to inflict disruptive and destructive attacks on targets within the critical infrastructure sector. Tactics employed by these cybercriminals often include the use of ransomware, denial-of-service attacks, and the defacement of websites.

Some actors aim to either refine their existing abilities or acquire new ones to disrupt the industrial control systems that underpin the one state energy, transportation, healthcare, and election sectors. In this article, there will be discussed the capabilities that an attacker may develop with AI technology as well as the types of attackers and their malitious intentions. The article employs a qualitative research method, focusing on descriptive and thematic analysis, literature review, and case studies to explore and understand the complex issues surrounding AI in cybersecurity and cyberattacks.

## 1. AI Functions That can be Utilized in Cyberattacks

### *1.1. Deepfakes*

To this day, most content created through artificial means is predominantly user-generated, typically demonstrated by technologists exhibiting their AI capabilities through training algorithms to manipulate the appearance or voices of disparate actors or politicians. Consequently, a significant part of the content produced by users is evidently inauthentic.

A deepfake represents an AI-manipulated image, sound, or video that appears authentic. The underlying technology can synthesize speech, replace faces, and control facial emotions. Someone can appear to speak or do something that they actually never said or did in a deepfake. Usually, face swapping and facial expression manipulation are prominent in deepfake videos. Academics and internet firms have tested several techniques to identify deepfakes. These techniques usually employed AI to scan videos/photos for digital flaws or details - like blinking or facial tics - that deepfakes are unable to accurately replicate.

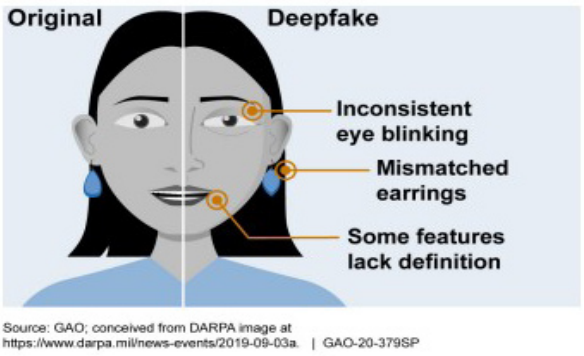


**Figure no. 1**: Examples of characteristics that may indicate a deepfake[1]

Deepfakes are frequently used for exploitation, despite the fact that they have harmless and legal uses in industries like entertainment and commerce. An article recently published ("An Investigation of the Effectiveness of Deepfake Models and Tools") by several researchers (Md. Saddam H. M. & Co., 2023) claims that a large majority of Deepfake recordings target politicians or celebrity personalities. They are widely disseminated online for misinformation. Deepfakes could be employed as a psychological warfare tool to sway elections, or to stir up civil unrest. They may also cause people to ignore valid proof of misconduct and, in general, erode public confidence in audiovisual content.

### *1.2. Social engineering*

The focus of social engineering attacks is on the attacker's use of trust and persuasion. People are more prone to act in ways that they otherwise would not be appropriate when they come across these strategies (Sowjanya M, 2022). Attackers using social engineering try to obtain confidential information from their victims so that they can sell it on the dark web or use it for their own agenda. While social engineering attacks are unique, they share a general pattern with comparable stages.

[1] URL: https://www.flickr.com/photos/usgao/49584240932, accesed at 20.01.2024.

The typical process unfolds as follows: 1) gathering victim information; 2) building a connection with the victim; 3) using the information at hand to launch an attack and, 4) making a clean getaway (Salahdine, F., Kaabouch, N, 2019). The attacker chooses a victim in phase one, commonly referred to as information gathering, depending on predetermined criteria. During the hook phase, the attacker uses email or direct communication to start building the victim's confidence. By divulging private information or creating security flaws, the attacker emotionally manipulates the victim during the execution stage. In the exit phase, the attacker disappears.

Large volumes of personal data from emails, social networking sites, and other sources may be analyzed by AI algorithms, which gives hackers the ability to develop phishing attacks that are both highly targeted and convincing. AI is capable of producing extremely persuasive voice calls, texts, and emails, which makes it more difficult for victims to recognize fraudulent activity.

### *1.3. Automated functions*

In the domain of cybersecurity defence, (AI) is already being applied extensively for improving the efficacy and scalability of defence mechanisms such as spam and malware detection. Simultaneously, many villains are naturally motivated to try their hand at using AI to breach other people's usually weak systems (Miles B & Co., 2018). These incentives include a premium on speed, labor expenses, and challenges in luring and keeping competent workers. AI can be used to automate a number of cyberattack-related tasks, including vulnerability assessment, exploiting vulnerabilities, and even launching the attack. As a result, the attack may become more effective and unlikely to be discovered.

For instance, it facilitates the creation of covert channels for information exfiltration, malware agent distribution, and malware agent command and control. These covert channels are designed to evade systems that identify anomalies, malware, and intrusions. AI also facilitates malware obfuscation, which makes it more difficult to identify.

AI features will also improve the identification and exploitation of hostile vulnerabilities. They will encourage malware's concealment and increase its sophistication in both design and operation. Malware with AI capabilities can adapt cleverly to changes in the target's behavior and avoid detection. They will operate as an autonomous and adaptable implant that learns from the host it is running on to stay hidden, find and categorize engaging material for exfiltration, find and infect new targets, and find new lateral movement channels or techniques.

For example, as early as 2018, researchers at IBM created a malware of this kind and called it „DeepLocker." (Melisha 2018). The malware could evade detection by the majority of antivirus and malware scanners until it targeted particular victims

by concealing its dangerous payload in carrier apps, such as video conferencing software. It is almost tough to reverse engineer this straightforward "trigger condition" that opened the assault. Only once the desired target has been reached will the malicious payload become accessible. IBM researchers created and presented a proof of concept in order to illustrate the effectiveness of DeepLocker's capabilities. The WannaCry virus evaded detection by antivirus engines and malware sandboxes by disguising itself as a harmless video conferencing application. A human was chosen as the triggering condition, and AI was trained to initiate the virus when specific parameters, such as the target's facial recognition, were satisfied. IBM created DeepLocker merely as an experiment to demonstrate how simple evasion strategies and open-source AI tools may be coupled to create extremely effective, targeted malware.

AI is used now to enhance target prioritizing and selection, avoid detection, and adapt creatively to behavioral changes in the target (either independently or in conjunction with humans). Autonomous software has long been capable of taking advantage of system flaws, but more advanced AI hacking tools could perform far better than previous examples and, in the end (though maybe not for a while), than human experts.

Cybercriminals use AI approaches to automate several parts of their attack pipeline, such processing payments or communicating with victims of ransomware (Dash B & Co., 2022). To identify targets with greater precision, large datasets are used, for example, to estimate one's assets and willingness for reimbursement according to online behavior.

## 2. Threat Actors with Different Types of Motivations

### *2.1. Nation states – geopolitical interest*

Government officials and organizations via state sponsorship. AI is being used by nation-states to produce more credible misinformation campaigns in an effort to erode public confidence in democratic processes, social cohesiveness, and government institutions. Also, the terrorist groups sponsored by a state can create more intricate and sophisticated attacks because they have plenty of resources and a great deal of experience. They may target critical infrastructure and industries within a nation, manipulate polls and spread misinformation to undermine its constitutional framework, or steal private data from businesses and government agencies.

According to "Homeland Threat Assessment 2024" the most advanced malicious influence operations on the internet are still being developed by Russia, China, and Iran (Office of Intelligence and Analysis 2024). It is likely that these adversaries will employ many of the same strategies to sway US audiences in the run-up to the 2024 election. AI-generated deepfake technology can be used, for instance, to sway

political opinion by creating phony videos of officials or celebrities speaking or acting indecently.

A key characteristic of gray zone conflicts in China and Russia is the blurring or thin line that separates cybercrime from funded by the state actions. In fact, these governments frequently employ contractors who, when not working for the state, might also commit cybercrime (Dunn C & Wenger A, 2022). For example, APT41, employed by the Chinese Ministry of State Security (MSS), engages in cybercrime activities during the late hours and cyberespionage actions during daytime.

Another instance of nation-state sponsored group utilizing artificial intelligence in hacking is APT28 (also known as Fancy Bear, Pawn Storm, the Sednit Gang and Sofacy), a cyber-espionage organization connected to the Russian government. This group has a history of automating their hacking operations and increasing their productivity through the use of AI and machine learning techniques.

APT28 is well-known for using the Carbanak virus (Kaspersky 2015), which automates its cyber operations using machine learning algorithms. The goal of the Carbanak malware is to steal information from banks and other financial organizations. It makes it very challenging to detect and prevent by using machine learning to find and exploit vulnerabilities in the target systems. They might want to become well-known or recognized in hacking circles. The big issue of this group, in regard with the hacking activity is that it operates in alignment with Russian military and political objectives.

### *2.2. Cybercriminals and transnational criminal organizations – profit*

The goal of cybercrime organizations is very clear: to make money. They can employ artificial intelligence systems to execute attacks or directly attack these systems in order to profit monetarily from their illicit activities. For instance, breaking into AI chatbots to gain access to private data, such as a customer's bank account information (obtaining financial information by pretending to be account holders) and request access to secure systems.

For instance, back in 2020, fraudsters used AI voice cloning to deceive a bank in the United Arab Emirates and steal $35 million (Alvarez Technology Group). Court records that Forbes has recently discovered reveal that the con artists tricked the manager of the bank into giving them the huge sum by imitating the deep-fake voice of a senior executive in the business. Hackers created a convincing bogus voice of the manager via AI voice cloning technology, stating that his business was going to undertake a purchase and lacked the funding to do so. The bank manager approved the transfer because he knew the executive's voice from prior job activity and considered everything was in proper order. The bank and the business had to handle the fallout after the cybercriminals fled with the stolen funds. This incident emphasizes the risks associated with cybercriminals utilizing AI to launch complex

attacks and the requirement for greater safety precautions to stop similar events from occurring in the future.

A recent report (UN 2023) states that during the previous year, there has been a 35% increase in online child sexual abuse and exploitation. Furthermore, there has been an increase in the global trafficking of illegal drugs enabled by cyberspace, along with guns, ammunition, and parts and components that are sold on the dark web. In certain parts of the world, there has also been a sharp rise in the trafficking of people for forced criminal activity related to casinos and organized crime groups' con games.

Independent groups of people operating on a global scale with the goal of obtaining power, influence, money, and/or commercial advantages through all or some illegal means are known as transnational criminal organizations. In order to hide their illicit activities, these groups may also employ international organizational structures, benefit from global trade and communication channels, or follow a violent or corrupt pattern.

AI-driven apps, systems, and technologies provide transnational organized crime groups with the tools and chance to engage in a wide range of illegal activities that are more intricate, can be carried out over longer distances, and pose less risk to individuals.

According to an analysis from Caldwell, M & Co., 2020, criminal organizations can use artificial intelligence (AI) for "supply chain management, risk assessment and mitigation, personnel vetting, social media data mining, and various types of analysis and problem-solving," just like legitimate businesses can.

Drones are used by organized crime in Africa for intelligence gathering, reconnaissance, and surveillance, but they can also be a threat to physical security. AI-controlled autonomous attack drones, already in use by Mexican drug cartels, can provide criminals greater adaptability, dexterity, and coordination when physically attacking infrastructure, people, or supply chains.

Criminals can use satellite imagery to plan and coordinate cross-border smuggling routes with the aid of artificial intelligence (AI) systems like Earth Observations, which offer extremely precise and up-to-date local terrain data. In order to avoid detection (such as biometric screening procedures), get around security measures (at banks, warehouses, airports, ports, and borders), or cause havoc with government and private sector networks and economic infrastructure, organized crime can also target AI systems.

### *2.3. Terrorist groups – ideological violence*

Extremists use AI for radicalization and recruitment, in order to get new followers to join, but also to identify people who share their ideology. These algorithms to look for trends and potential indicators of radicalization analyze massive amounts of internet data. With deliberate targeting, they seek out and entice weak people,

eventually bringing them around to radical viewpoints.

Extremist content can be produced and distributed via AI. Algorithms for natural language processing produce content that looks real. Via messaging apps, websites, and social media platforms, this content can disseminate radical narratives. Chatbots and other automated systems enable the rapid dissemination of this content, reaching a wider audience with less effort.

According to an article from Binder JF & Co, 2022, the term "online radicalization" refers to the process by which people use the Internet, particularly social media and other online communication platforms, to become exposed to, mimic, and internalize extremist ideas and attitudes. The phenomenon of online radicalization has raised significant concerns regarding grievance-based violence as well as terrorism. In fact, a comparative method that builds on the similarities among those who commit acts such as hate crimes, terrorist attacks, and high school shootings has been introduced in recent work.

### *2.4. Thrill-Seekers – satisfacion*

As their name suggests, thrill seekers target information systems and computers primarily for amusement, for the purpose of boasting or try new things. While some use hacking to test their ability to substract large amounts of confidential data, others use it as a means of learning more about how computer networks and systems operate. Script kiddies are a subset of thrill-seekers who, despite lacking sophisticated technical knowledge, target weak systems with pre-existing tools and methods mainly for entertainment or self-gratification. Excitement seekers may interfere with a network's cybersecurity and create a gateway for future cyberattacks, even though they rarely seek to do harm.

A set of AI-generated posters for unproduced films that are styled after Pixar and Disney animated features are known as the "Offensive AI Pixar" case (Toolify.ai). The posters frequently deal with sensitive subjects that are deemed improper, exploitative, and unsuitable for a family movie. When an AI-generated poster for an animated film called "Caust", based on the Holocaust and of Adolf Hitler' life, became viral on social media in October 2023, the trend began to take off.

### *2.5. Insider Threats – discontent*

Insider threat actors do not always have bad intentions, in contrast to the majority of other actor types. Some cause harm to their companies by accident, such as installing malware without meaning to or misplacing a company-issued device that a cybercriminal finds and utilizes to access the network. However, there are bad insiders as well. An example would be a disgruntled worker who misuses their access rights to substract data for financial gain or damages data or applications as a form of revenge for not being promoted.

For instance, Office of Public Affairs from US Department of Science mentioned in a press release from 2022 (Office of Public Affairs n.d.), the case of Twitter employee Ahmad Abouammo who exploited AI to accomplish his malicious intentions. In August 2022, Abouammo was found guilty of accepting bribes in return for gaining access to monitoring and sharing with representatives of the Saudi Royal family and the Kingdom of Saudi Arabia users' personal information on Twitter. It is plausible to believe that AI may have been used to automate the process of monitoring and collecting Twitter users' private information, even though the precise details of how AI was employed in this case are not given.

## Research method

The article extensively reviews existing literature, studies, and case examples. This approach is qualitative, focusing on understanding the nature, characteristics, and implications of AI in cybersecurity and cyberattacks through detailed descriptions and analyses. It categorizes different types of cyber threat actors and discusses their motivations and methods, which is typical of qualitative research that aims to understand phenomena in terms of the meanings people bring to them. The article delves into the ethical and societal implications of AI in cybersecurity, concerned with understanding human experiences and societal impacts.

## Conclusions

The integration of AI across various sectors, including cybersecurity, social networks, and retail, has both positive and negative implications. While AI significantly enhances cybersecurity defences, it also provides sophisticated tools for hackers, leading to more complex cyberattacks. This duality presents a critical challenge in the field.

Deepfake technology, a prominent application of AI, poses significant risks. It is primarily used to manipulate images and videos to create seemingly authentic but false representations. Although it has legitimate applications in entertainment and commerce, deepfakes are often exploited for malicious purposes such as spreading misinformation, swaying public opinion, and undermining trust in audiovisual content. This technology highlights the potential of AI to be used in psychological warfare and misinformation campaigns, especially targeting public figures such as politicians and celebrities.

Social engineering attacks, which rely on manipulating human trust, have become more sophisticated with the integration of AI. These attacks follow a pattern of information gathering, building trust, exploiting the acquired information, and then withdrawing without detection. AI's capability to analyze large volumes of

data from various sources, including emails and social media, has enabled the development of highly targeted and convincing phishing attacks, making them more challenging to detect.

The automation of cybersecurity functions through AI significantly improves the efficiency of defence mechanisms such as spam and malware detection. However, this also opens up opportunities for malicious actors to automate their attack processes, including vulnerability assessment and exploitation. For instance, the development of malware like DeepLocker demonstrates the potential of AI in creating sophisticated and targeted cyberattacks that are difficult to detect and counteract.

Different types of threat actors exploit AI for varied purposes. Nation-states use AI for geopolitical interests, including misinformation campaigns to undermine democratic processes and public trust in institutions. Cybercriminals and transnational criminal organizations primarily focus on financial gains, employing AI to execute attacks and penetrate secure systems. In contrast, terrorist groups use AI for ideological violence, leveraging it for radicalization, recruitment, and spreading extremist narratives. Thrill-seekers, often lacking advanced technical skills, exploit AI for amusement or to demonstrate their abilities, occasionally opening gateways for more severe cyberattacks. Lastly, insider threats pose a unique challenge as they may misuse AI either intentionally or unintentionally, leading to significant security breaches within organizations.

In conclusion, the rapid advancement of AI in cybersecurity is a double-edged sword. It offers unprecedented capabilities in data processing and pattern recognition, essential for detecting and neutralizing cyber threats. However, these same capabilities also raise serious ethical and privacy concerns because of their potential for intrusive data mining and surveillance. The complexity of the cyber threat landscape is increasing, with various actors employing AI to enhance their offensive and defensive capabilities. This necessitates a balanced approach, combining AI's computational power with human expertise and judgment for effective cybersecurity.

## **BIBLIOGRAPHY**:

Alvarez Technology Group. n.d. https://www.alvareztg.com/uae-bank-deepfake/

Binder JF, Kenyon J. 2022. "Terrorism and the internet: How dangerous is online radicalization?" *Front Psychol*. doi: 10.3389/fpsyg.2022.997390. PMID: 36312087; PMCID: PMC9606324.

Caldwell, M., Andrews, J.T.A., Tanay, T. *et al.* 2020. "AI-enabled future crime." *Crime Sci 9*, 14. https://doi.org/10.1186/s40163-020-00123-8

Dash, Bibhu & Ansari, Meraj Farheen & Sharma, Pawankumar & Ali, Azad. (2022). Threats and Opportunities with AI-based Cyber Security Intrusion Detection:

A Review. *International Journal of Software Engineering & Applications*. 13. 10.5121/ijsea.2022.13502.

Dunn Cavelty M, Wenger A. 2022. "Cyber security politics: Socio-technological transformations and political fragmentation." *Taylor & Francis*.

Flickr. n.d. https://www.flickr.com/photos/usgao/49584240932

Homeland Security. Office of Intelligence and Analysis. Homeland Threat Assessment. 2024. https://www.dhs.gov/sites/default/files/2023-09/23_0913_ia_23-333-ia_u_homeland-threat-assessment-2024_508C_V6_13Sep23.pdf, accessed at 20.01.2024.

Kaspersky. 2015. CARBANAK APT THE GREAT BANK ROBBERY. https://media.kasperskycontenthub.com/wp-content/uploads/sites/43/2018/03/080645 18/Carbanak_APT_eng.pdf, accessed at 20.01.2024.

Manyam, Sowjanya. 2022. "Artificial Intelligence's Impact on Social Engineering Attacks." *All Capstone Projects*. no 561.

Melisha Dsouza. 2018. IBM's DeepLocker: The Artificial Intelligence powered sneaky new breed of Malware. https://hub.packtpub.com/ibms-deeplocker-the-artificial-intelligence-powered-sneaky-new-breed-of-malware, accessed at 20.01.2024.

Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, Simon Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, Dario Amodei. 2018. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation". *Future of Humanity Institute. University of Oxford. Centre for the Study of Existential Risk. University of Cambridge. Center for a New American Security. Electronic Frontier Foundation. OpenAI.* https://doi.org/10.48550/arXiv.1802.07228

Mukta, Md. Saddam Hossain, Jubaer Ahmad, Mohaimenul Azam Khan Raiaan, Salekul Islam, Sami Azam, Mohammed Eunus Ali, and Mirjam Jonkman. 2023. "An Investigation of the Effectiveness of Deepfake Models and Tools". *Journal of Sensor and Actuator Networks* 12, no. 4: 61. https://doi.org/10.3390/jsan12040061

Office of Public Affairs. n.d. https://www.justice.gov/opa/pr/former-twitter-employee-sentenced-42-months-federal-prison-acting-foreign-agent

Salahdine, F., Kaabouch, N. 2019. "Social engineering attacks: A survey. *Future Internet*. no. 89. https://doi.org/10.3390/fi11040089

Toolify.Ai. n.d. https://www.toolify.ai/ai-news/ranking-ai-disney-poster-offensive-ai-pixar-movies-83742

U.N. n.d. https://press.un.org/en/2023/gashc4374.doc.htm