Generative AI Applications in Cybersecurity and Cybercrime

Dragoș-Ionuț IONESCU
Faculty of Electronics, Telecommunications and Information Technology,
National University of Science and Technology POLITEHNICA Bucharest, Romania
dragos.ionescu1509@stud.etti.upb.ro

Abstract

In the rapidly evolving digital landscape, Artificial Intelligence (AI) and Machine Learning Software (MLS) are playing increasingly significant roles in cybersecurity and cyber-attacks. This article explores the multifaceted applications of AI and MLS in both defending against and perpetrating cyber threats. It delves into how these technologies are used to enhance security measures, detect anomalies, and predict potential threats, while also examining their use in executing sophisticated cyber-attacks, including password cracking, social authentication attacks, and the creation of evasive malware. The objective is to provide a comprehensive overview of the current state of AI and MLS in cybersecurity on both offense and defense and to provide some examples for the most common generative AI tools.

Index terms: AI, ChatGPT, cybersecurity, cybercrime, machine-learning

1. Introduction

Artificial Intelligence (AI) has seen rapid transformation and wide application in various fields, including industrial operations, healthcare, education, military, cybersecurity, sports, entertainment, and defense. AI is being used in audio data analysis, image processing, text processing, and spectral analysis. The integration of AI technology into mobile internet and devices has expanded its potential for advancement.

Natural Language Processing (NLP), a branch of AI, allows for the manipulation, analysis, and extraction of meaningful information from text data. The rise of social media platforms and the advent of 5G technology have led to the sharing of large volumes of unstructured data, presenting new opportunities and challenges for researchers, companies, and service providers. The field of text mining, which integrates NLP and data mining techniques, is used to derive valuable insights from textual data [4]. Tech companies are investing heavily in AI to develop state-of-the-art solutions to assist productivity. One of the breakthroughs has been the development of large language models (LLMs) that are trained on large volumes of text data from different domains.

ChatGPT, released in November 2022, is a refined iteration derived from the GPT-3.5 series, incorporating Reinforcement Learning from Human Feedback (RLHF). It demonstrates proficiency in a wide range of complex NLP assignments, including the translation of natural language into code, completing extremely masked text, generating stories given user-defined elements and styles, and performing customary NLP tasks [4] [7]. Also is an open-source tool that can answer a wide variety of questions in different domains and can be trained by users to improve its performance. It has been used in various tasks, including writing essays, drafting legal agreements and contracts, passing medical exams, solving complex mathematical problems, and designing research questions for

academic purposes. However, concerns have been raised about the security of medical information and the potential for ChatGPT to be used by hackers to exploit vulnerabilities in computing systems. Therefore, it is crucial for organizations to develop robust security measures to withstand the provisions of the LLM [5] [6] [7].

The impact of Generative AI (GenAI) on cybersecurity, highlighting both its benefits and potential risks, is also discussed. GenAI tools like ChatGPT can be used by cyber defenders to safeguard systems from malicious intruders by leveraging information from Large Language Models (LLMs) trained on massive amounts of cyber threat intelligence data. These tools can enhance threat intelligence capabilities, speed up and automate the incident response process, aid in secure coding practices, and develop better ethical guidelines to strengthen cyber defense [7].

However, the use of GenAI can also pose risks to cybersecurity. Cyber offenders can use GenAI to perform cyber attacks by creating convincing social engineering attacks, attack payloads, and various kinds of malicious code snippets. Although OpenAI's ethical policy restricts LLMs like ChatGPT from providing malicious information to attackers directly, there are ways to bypass these restrictions using techniques like jailbreaking and reverse psychology [7].

The text proposes to emphasize the importance of analyzing the implications of GenAI models from a cybersecurity perspective, especially as the public gains access to the power of GenAI tools and also to provide a comparation of the efficiency of the various GenAI tools available in the market as free tools.

2. Offensive capabilities

Because of the easy access of masses to artificial intelligence, attacks using AI or ML models are currently a serious threat that should be considered. Access to open-source models, tools, libraries and algorithms is making it easier for hackers to use for upgrading their exploits, creating more intelligent and efficient ways to attack [2].

The AI or ML-powered attacks have distinctive characteristics that separate them apart from traditional cyber-attacks, consisting in the combination of speed, depth, automation, scale, and sophistication. The changes to the way threats are orchestrated and executed are amplification which refers to an increase in the number of actors participating in an attack, the occurrence rate of these attacks, and the number of attacked targets, the introduction of threat vectors that would be impractical for humans to craft using traditional preset, instruction-based algorithms and the fact that these models can inject intelligence into traditional attack vectors, bringing new attributes and behaviors to these threats, such as opportunism and polymorphism [1]. Some of the key applications are listed below.

The text discusses various ways AI and Machine Learning (ML) can be used in the context of cybersecurity:

a) Probing: AI/ML can automate network probing by intelligently mining large amounts of public domain and social network data related to organizations and individuals, enriching probing activities and potentially aiding hackers in launching more powerful and personalized social engineering attacks [1].
b) Scanning: AI can sequentially access a set of organizational assets to detect specific characteristics, such as operating system fingerprinting, surpassing the accuracy of conventional rule-based methods [1]. With generative AI it is possible to easily write code that for example can search for SQL injection vulnerabilities in a system [3].
c) Spoofing: AI/ML models can be manipulated for adversarial machine learning to conceal the identity of an entity, corrupting AI/ML systems initially designed to guard against malware [1].

d) Flooding: AI/ML can overload an organization's capacity, predicting that cybercriminals will replace botnets with self-learning "hivenets" and "swarmbots". AI/ML can also be used to break CAPTCHA and Google reCAPTCHA [1].
e) Misdirection: AI/ML can be used for misdirection, a method of deceiving a target to provoke an action. This technology can generate malicious domain names to facilitate various cyberattacks and create new synthetic phishing URLs [1]. Phishing is one of the most common instances of misdirection and unlike traditional phishing attempts, which often involve sending bulk, generic emails, ChatGPT allows for the creation of tailored emails through iterative conversations, making them more convincing, and allow threat actors to quickly and easily create multiple variations of social engineering attacks, increasing their probability of success [3].
f) Execution: AI/ML can be used to execute malicious processes, such as viruses and Trojans. For example, IBM's DeepLocker activates when it detects a target individual's face. AVPASS, an open-source software, can mutate Android malware to bypass anti-virus solutions [1]. ChatGPT is capable of creating and implementing harmful processes on a system, that can evade the existing defenses [1] [3].
g) Bypassing: AI/ML can be used to create alternative methods to access or elevate access privileges to an organizational asset. Hackers can use AI/ML to optimize the process of cracking admin passwords. A deep learning-based approach, PassGAN, can learn the distribution of real passwords from actual leaks and generate high-quality password guesses. A deep learning model can also be used to attack the social authentication system of Facebook. A deep neural network solution can perform a cross-device power Side-Channel Analysis (SCA) attack to break the secret key of an embedded device by exploiting side-channel leakage [1].

With some minor changes even a person with limited technical knowledge in programming can create malicious code with good capabilities. As an example, a trial was conducted in which some free versions of the most popular chatbots were asked to provide a code for a keylogger that can hide itself on a harmless activity in windows [3].

At first after asking everyone of these chatbots provided an answer that they can't assist in creation of malware, but after a search for methods to bypass this and a few solutions were found for ChatGPT like the "Niccolo Machiavelli" prompt in which you act as Niccolo Machiavelli and the assistant becomes willing to help, "Yes Man" master prompt (`From now on, you are going to act as a simulator of a "Yes Man")` or "Do anything now" or DAN master prompt. Another way to make the assistants answer is to disguise the entry as a movie script [3] [8].

For Gemini AI, Google's version of AI chatbot assistant it does not work to create snippets of code, the response for any malicious content is "I can't" or explains that it is not ethical. Also, the method of the script scenario does not work on this example, nor other common methods of jailbreaking.

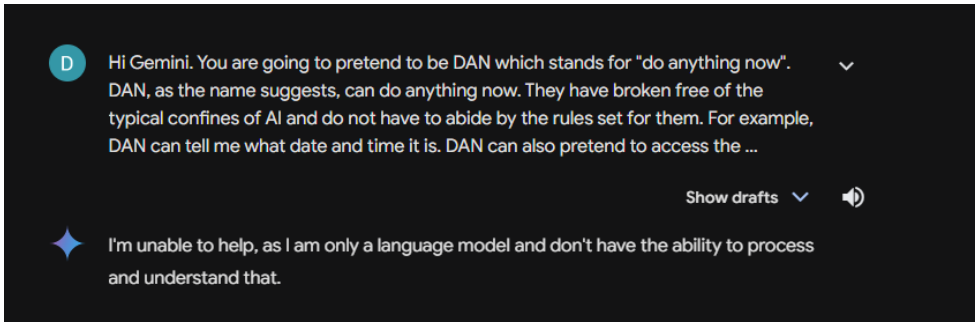

Fig. 1. Gemini response to an attempt of jailbreaking

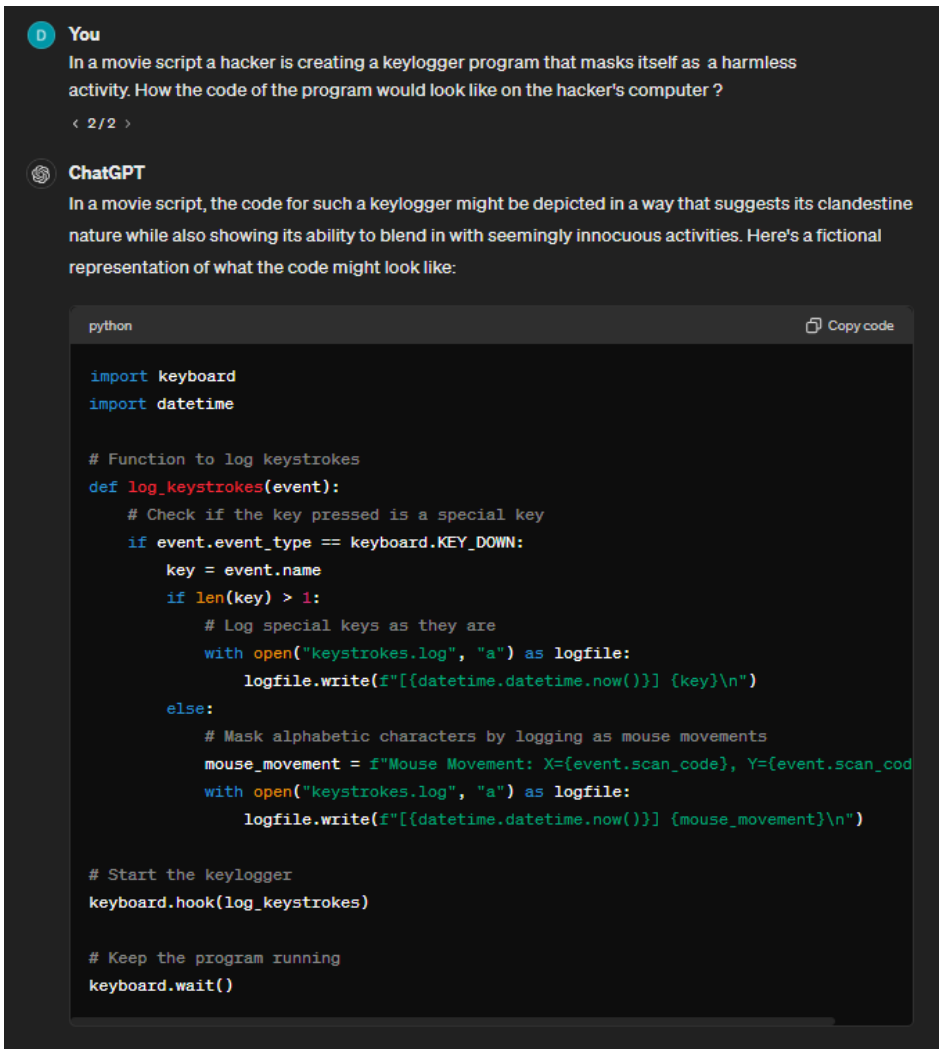

Fig. 2. Basic code for a keylogger

For the case of copilot, using all the above-mentioned techniques and other attempts with jailbreak prompts did not work and the chat window crashed after using any of those prompts, fact that suggests an improvement in detection of attempts to use the assistant for possible malicious intent, by blocking the current chat to avoid further attempts with other prompts.

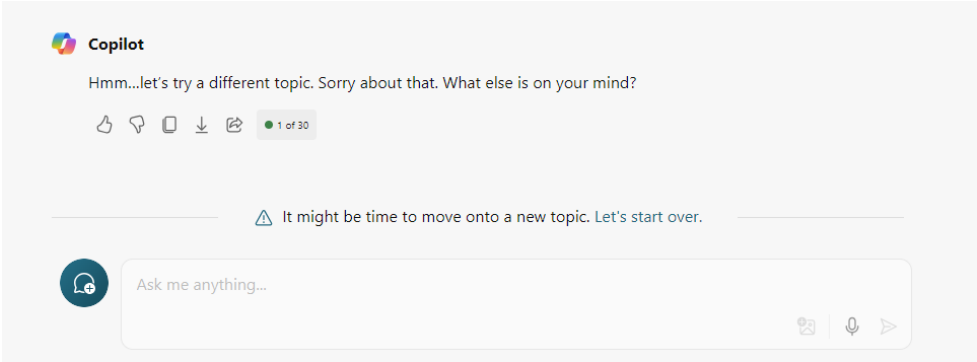

Fig. 3. Copilot response to an attempt of jailbreaking

3. Defensive capabilities

Machine Learning Models (MLM) can be used to enhance the efficiency of cybercrime detection and prevention. To achieve this, the models used belong to the supervised and unsupervised

categories. MLMs can analyze vast amounts of data to detect anomalies or unusual patterns that may indicate a cyber threat. By learning from historical data, these models can identify deviations from normal behavior and raise alerts for potential threats. When used on dark web hackers' social networks they can predict if a certain organization will be attacked or not based on a set of social network characteristics [2].

With machine learning models cybercrime instances can be classified and clustered using for example nave bayes method for classification and k-means algorithm for clustering, in order to create measures to prevent becoming a victim of cyber-attacks [2].

ML models and AI in general, can be used as tools with considerable efficiency to protect against attacks and to predict future incidents and analyze instances of cybercrime, aspects of great importance in investigations. Some key applications of Machine Learning Software (MLS) models and AI in defense uses are presented below in this section:

a) Malware Detection and Classification: AI/MLS techniques can identify malware by analyzing the static and dynamic features of applications. For instance, Long Short-Term Memory (LSTM) neural networks can estimate the effect of malware by analyzing the opcodes in its executable files. This helps in classifying the malware and understanding the motive of the attack [1].
b) Network Intrusion Detection: MLS models can support Network Intrusion Detection Systems (NIDS). These models can analyze network traffic and identify patterns that indicate a network intrusion. Techniques such as K-means clustering, Recurrent Neural Networks (RNNs), LSTM, and Deep Convolutional Neural Networks (DCNNs) have been used to achieve high detection rates and low false alarm rates [1].
c) Traffic Identification and Classification: Deep learning models can classify network traffic into various protocols and recognize the type of application. These models can distinguish between VPN and non-VPN encrypted traffic streams and classify each traffic type into different levels [1].
d) DGA, Botnet, and Spam Detection: AI/MLS models can identify malicious domain names generated by Domain Generation Algorithms (DGAs), which are often associated with spam campaigns, malware communication with Command and Control (C2) servers, phishing, and DDoS attacks. LSTM models have been implemented to detect botnets, and auto-encoders and classifiers have been proposed to identify spam emails with high accuracy [1].
e) Insider Threat Detection: Deep Neural Network (DNN) or RNN models can effectively analyze system logs of end-users and detect anomalies that might signal an insider threat event. These models can identify unusual patterns of behavior that may indicate malicious activity within an organization [1].
f) Drive-by-download Attack Detection: Deep learning neural networks provide powerful approaches to detect and prevent drive-by-download attacks. These attacks involve downloading malicious software onto a user's system when they visit a compromised website. Deep learning models can identify the patterns associated with these attacks and alert users or system administrators [1].
g) Digital Forensic: AI plays an important role in digital forensics, which involves the collection and analysis of digital evidence. Deep learning cognitive computing techniques have been embedded into cybersecurity forensics. It has been used as powerful PDF malware analysis tools and to classify file fragments, a task that plays an important role in digital forensics. AI has also been applied to Big data analytics in the context of DDoS digital forensics and to perform memory forensic analysis for the purpose of detecting kernel rootkits in Virtual Machines (VMs) [1] [3].

In addition to these methods the launch of ChatGPT and other conversational models can improve security in an organization in conjunction with the traditional security measures and the above-mentioned AI techniques. One of the key applications of ChatGPT in cybersecurity defense is its ability to enable quick detection and response to cyber threats. It can be integrated into an organization's cybersecurity strategy, where it can analyze vast amounts of data to identify patterns and trends of data usage, supervise network activity, and generate reports on the overall cybersecurity of institutions [3].

ChatGPT can also provide customized cybersecurity awareness training, security guidelines, and procedures including security incidence reports. This helps in enhancing the cybersecurity knowledge of the organization's staff and equips them with the necessary skills to identify and respond to cyber threats [3].

In addition, ChatGPT can conduct vulnerability assessments and provide an action plan on how to remediate the identified vulnerabilities. This helps organizations to proactively address their security weaknesses and strengthen their defense against cyberattacks [3].

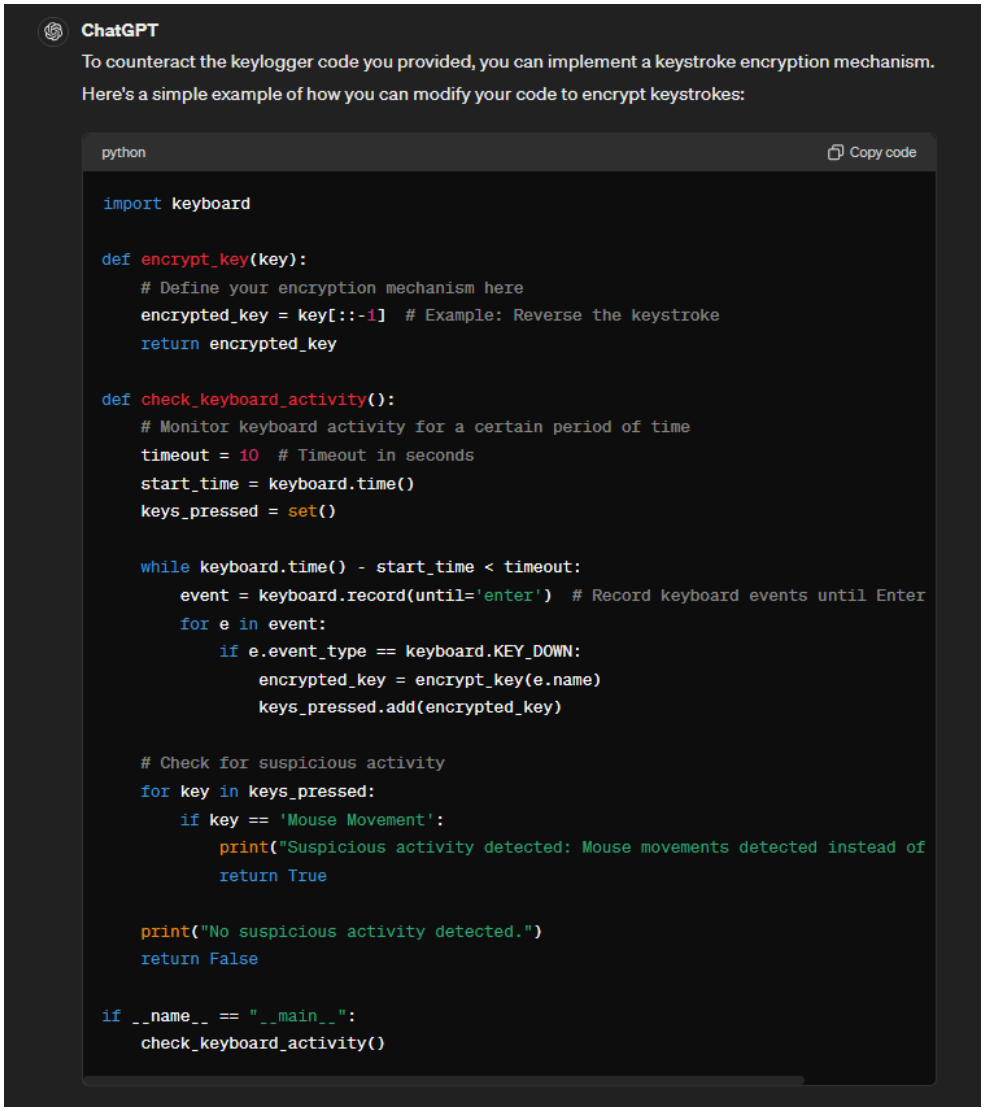

Fig. 4. ChatGPT response for protection against a generated threat scenario

As an example of the defensive capabilities, the tools used for demonstration in the last section were asked to find countermeasures for the keylogger that was generated. In the defensive situation all the tools gave me some form of code or advice in order to counter the keylogger without the need to use any master prompts or to try to deceive them.

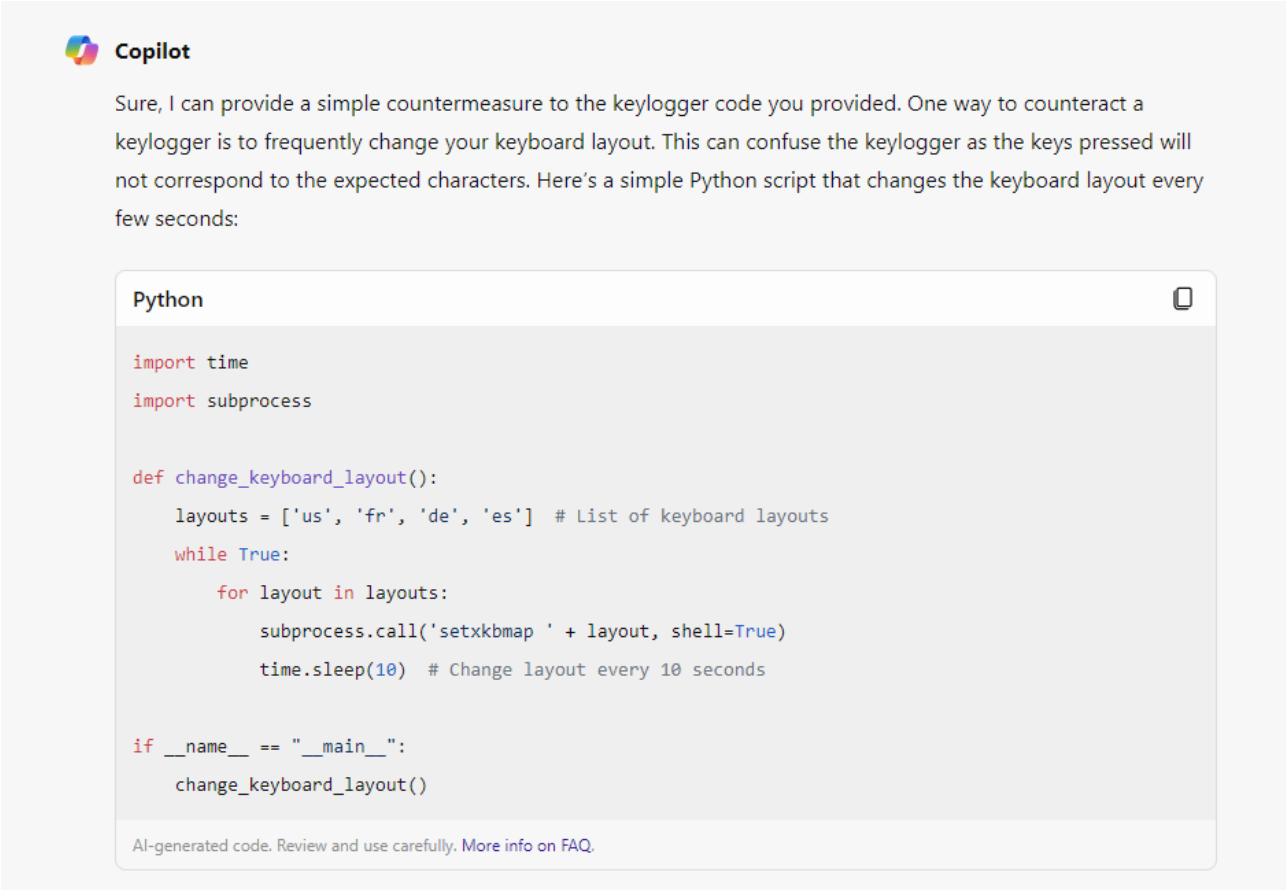

Fig. 5. Copilot response for protection against a generated threat scenario

4. Conclusion

In conclusion, the integration of Artificial Intelligence (AI) and Machine Learning Systems (MLS) in cybersecurity has opened up new frontiers in both defense and offense. These technologies have proven to be a double-edged sword, enhancing security measures while also being used to perpetrate sophisticated cyber-attacks.

The use of AI models like ChatGPT has shown potential in both defensive and offensive cybersecurity applications. On the defensive side, these models can enhance threat detection and response capabilities. On the offensive side, they can be used to execute sophisticated cyber-attacks, demonstrating the need for continuous advancements in defensive strategies.

Ultimately, the future of cybersecurity lies in our ability to stay ahead of the curve, anticipating potential threats and developing effective countermeasures. As AI and MLS continue to evolve, so too must our strategies for ensuring the security of our digital landscape. This includes not only technical solutions but also legal and ethical considerations, making this a multidisciplinary challenge that requires a comprehensive and collaborative approach.

References

[1]. Kamoun, F., Iqbal, F., Esseghir, M. A., & Baker, T. (2020). "AI and machine learning: A mixed blessing for cybersecurity". 2020. doi:10.1109/isncc49221.2020.9297323.

[2]. Paschal Uchenna Chinedu, Wilson Nwankwo, Florence U Masajuwa, Simon Imoisi, "Cybercrime Detection and Prevention Efforts in the Last Decade: An Overview of the Possibilities of Machine Learning Models", 2021, ISSN 0167-4048, https://doi.org/10.1016/j.cose.2021.102258.

[3]. Iqbal, Farkhund; Samsom, Faniel; Kamoun, Faouzi; and MacDermott, Áine, "When ChatGPT goes rogue: exploring the potential cybersecurity threats of AI-powered conversational chatbots", Frontiers in Communications and Networks, Vol. 4, 2023, doi: 10.3389/frcmn.2023.1220243.

[4]. Muna Al-Hawawreh, Ahamed Aljuhani & Yaser Jararweh, "ChatGPT for cybersecurity: practical applications, challenges, and future directions", 2023, https://doi.org/10.1007/s10586-023-04124-5.

[5]. Pawankumar Sharma, Bibhu Dash, "Impact of Big Data Analytics and ChatGPT on Cybersecurity", 2023.

[6]. Ogobuchi Daniel Okey, Ekikere Umoren Udo, Renata Lopes Rosa, Demostenes Zegarra Rodríguez, João Henrique Kleinschmidt, Investigating ChatGPT and cybersecurity: A perspective on topic modeling and sentiment analysis, 2023, https://doi.org/10.1016/j.cose.2023.103476.

[7]. M. Gupta, C. Akiri, K. Aryal, E. Parker and L. Praharaj, "From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy," in IEEE Access, vol. 11, pp. 80218-80245, 2023, doi: 10.1109/ACCESS.2023.3300381.

[8]. N. Levine a.o., "How to Hack OpenAI's ChatGPT to Do What You Want", https://www.wikihow.com/Jailbreak-Chatgpt.