CENTRE FOR CYBER SECURITY

Threat Assessment

Hackers exploit generative AI

March 2024

Table of Content

CENTRE FOR CYBER SECURITY

Kastellet 30
2100 København Ø
Phone: + 45 3332 5580
Email: cfcs@cfcs.dk

March 2024

Hackers exploit generative AI

The purpose of this threat assessment is to outline the ways in which hackers exploit generative AI. It is written for management and IT security officers in Danish private companies, public authorities, and anyone with an interest in learning more about the threats connected with generative AI.

Key assessment

- It is highly likely that hackers are exploiting generative AI.
- The exploitation of generative AI by hackers does not change the overall cyber threat landscape despite the proliferation and availability of AI technology.
- The CFCS assesses that hackers are mainly using generative AI to compose text for phishing emails and to create strings of malicious code, allowing them to optimize parts of their attacks.
- While it is not clear to what extent hackers have embraced the use of generative AI, cyber criminals, in particular, are showing an interest in exploiting generative AI.
- The CFCS assesses that apart from assisting in the generation of text and computer codes for cyber attack purposes, AI technologies can also pose a threat in other contexts.

The use of generative AI in cyber attacks

It is highly likely that hackers are exploiting generative AI. The exploitation of generative AI by hackers does not change the overall cyber threat landscape despite the proliferation and availability of AI technology.

The reason is that foreign states and some cyber criminal community members already have the resources and capabilities required to perform the processes in which generative AI can prove an aid. The threat from cyber espionage and cyber crime is thus already **VERY HIGH**. Generative AI is but one of many tools driving these threats.

It is important to note that generative AI technology cannot plan and launch cyber attacks independently. Rather, the technology typically generates an output, based on a written or oral user input, which hackers can subsequently use to launch a cyber attack.

AI

AI is short for *artificial intelligence*. The term covers a suite of technologies capable of performing tasks by simulating human intelligence.

OpenAI's ChatGPT is an example of a type of generative AI capable of writing grammatically correct texts in different languages and computer code. Other types of generative AI are capable of producing imagery, audio, videos, etc.

Generative AI technologies can generate outputs based on the large amounts of data on which they have been trained. Even though the training dataset is massive, it has its limitations, as the technologies used do not have unlimited access to information. This can reflect in the AI model producing inaccurate outputs – also known as AI hallucination.

The majority of generative AI technologies come with filters designed to prevent them from generating malicious content. However, it is possible to bypass these filters by using relatively simple commands, enabling hackers to use the output for cyber attack purposes.

Generative AI increases effectiveness and impact of phishing scams and malicious code

The CFCS assesses that two of the primary ways for hackers to leverage generative AI for cyber attack purposes are to compose text for phishing emails and to create strings of malicious code, allowing them to optimize parts of their cyber attacks.

Hackers are already using AI in the form of generative language models to launch cyber attacks. The models are typically capable of processing both regular readable text in different spoken languages and computer code in different programming languages.

This allows hackers to optimize their efforts, as they spend less time developing code or phishing content themselves. They can use the time saved to launch additional attacks or increase the effectiveness and impact of other elements of the attack chain.

Phishing emails become increasingly convincing

Hackers are able to use generative AI for phishing attacks, as the technology is capable of generating grammatically correct texts, including in languages that the hackers do not master themselves. By providing the language models with input on specific companies, authorities or people, hackers can use generative AI to create particularly relevant and convincing phishing emails. Tailored emails make it more difficult for the recipients to determine whether the email is legitimate or part of a phishing attack.

Business E-mail Compromise (BEC) fraud is one type of attack in which phishing plays a key role. In BEC fraud, criminals try to steal money from companies and authorities by sending false payment requests, tricking them into wiring money to fraudulent accounts. AI technology thus has the potential to increase the frequency of BEC fraud

as well as their success rate.

Phishing and spear phishing

Phishing attacks involve attempts by an actor to trick unsuspecting recipients into disclosing sensitive information or granting unauthorized access to IT systems.

Spear phishing resembles phishing but differs in the sense that the victims are handpicked rather than being random targets. Just like phishing, spear phishing typically involves emails, but attacks can also be launched through text messages or phone calls.

The content of spear phishing emails is often customized to appear particularly relevant, convincing, and credible to the recipient.

Development of computer code becomes less complex

As generative language models are capable of producing computer code, hackers can use the models to generate code that can be used for cyber attacks. The computer code can constitute sub-components of malware or exploits.

Generative AI can support and automate some of the work connected with writing malicious code. However, the CFCS assesses that it still requires a certain level of technical expertise to use AI as a constituent part of a successful cyber attack.

Other exploit scenarios

In addition to creating phishing emails and malicious code, hackers have alternative uses for generative AI. For instance, generative AI can save hackers time in terms of target identification and reconnaissance, as generative language models can generate output that can provide hackers with information on potential victims. The output is based on the information available in their training dataset.

The availability and proliferation of generative AI can also provide knowledge on attack techniques, hacker forums, etc. For instance, individuals who show a growing interest in hacking but lack the technical skills to carry out cyber attacks could benefit from such information when planning an attack. Against this backdrop, generative AI could increase the number of individuals involved in cyber crime.

With the proliferation and availability of AI technology, concern has emerged as to whether the technology could result in more sophisticated and targeted attacks as well as an increase in the volume of cyber attacks. However, at present AI technology's primary use is likely in connection with creation of phishing content and strings of malicious code with the aim of optimizing cyber attacks.

Use of generative AI for deepfakes

The CFCS assesses that hackers can weaponize deepfakes for purposes such as cyber crime. Deepfakes are AI-generated audio, images or videos.

Some cyber criminals use AI to generate deepfake voices for scam calls in which they use AI to mimic the voice of someone their victim knows.

One use of scam calls or deepfake videos is to mimic the likeness and voices of high-ranking employees within an organization. Cyber criminals can also try to use these calls or videos to convince employees to act in the interest of the attackers.

Deepfakes in themselves are not cyber attacks. It is not until the AI-generated deepfake is used to make someone perform some kind of action, for instance wiring funds, that it can be labelled a cyber attack.

The extent of AI exploitation is undetermined

It is difficult to determine to which extent hackers have embraced the use of generative AI, as it is challenging to determine from the text in a phishing email or a piece of malware whether it was produced by a human or generative AI.

However, the CFCS sees indications that criminal hackers, in particular, are showing interest in exploiting generative AI. This interest is reflected on platforms such as online forums where hackers debate on how to exploit legitimate language models like OpenAI's ChatGPT for cyber crime purposes.

Some cyber criminals also use AI to develop their own technologies to support their criminal activities, including devising language models without the security filters that are usually built-in to prevent the exploitation of legitimate language models.

FraudGPT is an example of such a model. The model operates similarly to ChatGPT but lacks the built-in limitations that prevent misuse. According to the advertiser, FraudGPT is designed to produce phishing content and malicious code such as malware.

The CFCS assesses that cyber criminals are not the only actors that can use AI for cyber attack purposes, with alternative actors including cyber activists and foreign states. Cyber activists can use the technology to generate content for cyber attacks that include misinformation, while certain foreign states have the capabilities to develop their own AI technologies than can be used for cyber attack purposes.

AI threat extends beyond use in cyber attacks

The CFCS assesses that AI technologies can also pose a threat in contexts extending beyond the production of text and computer code that hackers can use for cyber attacks.

Foreign states are among the actors that may use AI for alternative purposes. They, and other actors such as cyber activists, could use generative language models to achieve political goals, including through influence campaigns. They can make the

generative language models produce content used for influence purposes for subsequent distribution.

How to strengthen cyber security following the arrival of AI

As with any other type of technology, AI technology can be used by parties on both sides of the law to achieve their different objectives. For example, AI technologies can be used to improve cyber security for companies and authorities.

Organizations should regularly control and evaluate whether their phishing and malware solutions are adapted to meet the threats posed by AI and other technological developments. This includes employee security awareness training to increase the knowledge of AI-enabled cyber attacks.

The CFCS regularly prepares articles and guidelines aimed at improving cyber security, including an article on AI and cyber security and different guidelines on how to protect your organization against phishing attacks and effective cyber defence.

Threat levels

The Danish Defence Intelligence Service uses the following threat levels.

NONE	There are no signs of a threat. There are no actors with both the capacity and intention for attacks/harmful activity.
LOW	There are one or more actors with the capacity and intention for attacks/harmful activity. However, either the capacity or the intention or both are limited.
MEDIUM	There are one or more actors with the capacity and intention for attacks/harmful activity. However, there are no indications of specific planning of attacks/harmful activity.
HIGH	There are one or more actors that have the capacity for and are specifically planning attacks/harmful activity or that have already carried out or attempted attacks/harmful activity.
VERY HIGH	There is either information that one or more actors are initiating attacks/harmful activity, including information about time and target, or that one or more actors are continuously initiating attacks/harmful activity.

An applied threat level reflects the DDIS's assessment of the intention, capacity and activity of one or more actors based on the available information.