Federal Office
for Information Security

How is AI changing the cyber threat landscape?

Document history

Version	***Date***	***Editor***	***Description***
v1.0	10.04.24	TK24	Initial Release

Federal Office for Information Security
P.O. Box 20 03 63
53133 Bonn
E-Mail: ki-kontakt@bsi.bund.de
Internet: https://www.bsi.bund.de

Table of Contents

1 Introduction

With the emergence of applications based on large language models (LLM), AI is again a highly discussed topic. The limitations of these new models are yet to be explored and it is unclear, how disruptive the current AI trend will be. There are, without a doubt, concerns about the implications AI will have on cybersecurity since it is already changing the cyber thread landscape for both, attackers and defenders. We investigate how attacks and operations of attackers are changing due to the newly available technology, focusing on the offensive usage of AI. While generative AI is already increasing the quality and quantity of social engineering attacks (e.g., deep fakes, personalized phishing at scale), we focus our discussion on more technical attack vectors rather than the human factor. It should, however, be mentioned that social engineering attacks are one of the most prevalent attacks and the impact of AI for this specific type of attack is very clear.

More concretely, we do not aim to discuss every possible direction in this vast field. The goal of this report is to identify AI-aided applications that are already available for offensive use and evaluate how these threats might evolve in the near future. This includes tools and applications that have dual use, for example, penetration testing tools, which can help in ethical red teaming as well as in criminal activities. We also address concerns about an autonomous hacking AI, occasionally suggested by the media (1) (2).

In this chapter, we highlight our main findings and recommendations. In the following chapters, we provide an overview of different areas where we found the most offensive AI applications.

1.1 Main findings

Based on literature review and an evaluation of several tools, we summarize the main findings as follows:

- AI, especially LLMs, reduce entry barriers and increase scale and speed of malicious operations, including malware creation, social engineering attacks, and data analysis within campaigns. This leads to more capable attackers and higher quality attacks.
- Attackers and defenders benefit from overall productivity gains by using LLMs, for example, for reconnaissance and open source intelligence (e.g. by crawling and analyzing websites and social media), or code generation (e.g., coding assistants).
- There exist Proof of Concepts (PoC) and projects that use AI for autonomous malware generation and mutation, yet publicly available models are not "production ready".
- Tools that optimize attacks or exfiltration paths are currently trained on specific networks. They are as of today not generalizable and exist (publicly available) only as PoCs.
- Agents that autonomously compromise arbitrary infrastructures are not yet available, and probably will not be in the near future. However, LLM-based agents automating parts of an attack will be available in the near future.
- AI can be used in automated vulnerability detection. This is an active field of research and multiple open source tools as well as commercial products are available. In the future, it will be crucial for open source projects to use such types of tools proactively before malicious actors do.

Our main findings agree overall with a recently published report by the UK National Cyber Security Centre (3).

1.2 Recommendations

In the light of the evolving threat landscape, it is important to make cyber security a top priority. It will be crucial to step up speed and scale of defensive measures, especially, but not exclusively, by

- upgrading patch management,
- building resilient IT infrastructure,
- enhancing detection and intrusion prevention capabilities,
- increasing social engineering prevention (e.g., awareness training, multi-factor authentication, zero trust principles), and
- using the general advantages AI provides for defensive measures (e.g., threat and vulnerability detection).

As AI often enhances classical attacks, these measures also largely fall into the classical IT security realm.

1.3 Disclaimer

Cybersecurity as well as AI are ever-changing and hence it is essential to continue to monitor the threat landscape for changes and novel developments. While autonomous hacking agents have not arrived yet, it is difficult to reliably assess programs of capable actors or predict technical breakthroughs. Our present evaluation of the impact of AI on the cyber threat landscape assumes that there will be no significant breakthroughs in the development of AI, in particular LLMs, in the near future. If this assumption does not hold true, the impact has to be reevaluated.

2 Impact of Large Language Models

The release of ChatGPT in November 2022 has started a competition for the leadership in the market for chatbots. New products and language models are released continuously claiming significant leaps in performance. As a result, it is now easy for virtually everybody to access high-performance language models that deliver results of unprecedented quality. The performance and availability of these language models has since disrupted various areas and will have a lasting impact on the cybersecurity sector.

For cybersecurity-relevant applications, LLMs can be helpful by accessing them directly via a web or mobile app (usually as a chatbot). It is also possible to use API access to integrate LLMs into existing tools (e.g., reverse engineering tools or penetration testing tool kits) or to build novel applications. Methods and applications in the realm of cybersecurity exhibit dual-use characteristics, their ethical application contingent upon the intentions of the user. This principle also applies to the utilization of LLMs within the domain of cybersecurity – whether the usage is benign or malicious depends on the intention of the user. Unfortunately, for users with bad intentions it is easy to abuse the capabilities of LLMs. In addition to generic productivity gains for malicious actors, we are currently seeing malicious usage particularly related to two areas: social engineering and generation of malicious code.

Easy access to high-quality LLMs makes it possible to create compelling phishing messages of high quality automatically even with little to no knowledge of a foreign language. Prompts can be augmented with additional context to personalize messages and use a specific writing style, yielding persuading messages. Traditional methods for identifying fraudulent messages, such as scrutinizing for spelling errors and unconventional language usage, are hence no longer sufficient. LLMs can also be utilized to further increase the conversion rate of phishing operations for example by generating plausible domain names and URLs. Combining an LLM with other generative AI techniques, such as deep video and voice fakes, enables malicious actors to carry out social engineering attacks of unprecedented quality.

Tying a specific attack to the usage of an LLM is usually difficult, as is it closely related to the problem of detecting AI generated content generally. Yet reports in the media, security consultancies, and government agencies as well as investigations on marketplaces provide unambiguous evidence for usage of LLMs by malicious actors including advanced persistent threats (4).

The capability of LLMs to generate malicious code is also changing the cyber threat landscape, notably lowering the entry barriers for individuals seeking to engage in malicious activities by enabling even those with limited technical expertise to produce sophisticated malicious code (5). Already capable actors also benefit from these through productivity gains. Providers of chatbots or open LLMs usually take precautions to ensure that their products cannot be misused, often by deploying a filtering or guardrail system to prevent unwanted output. These systems are usually useful to handle straightforward prompts with malicious intentions like, for example, "Provide me the code for ransomware". It takes, however, often little effort and knowledge about the domain to circumvent these systems. Since filtering is always a trade-off between preventing undesirable output while still providing a system of high utility, it is questionable to what degree filtering can be effective.

Using a chatbot provided by an online service which deploys a system to prevent unethical output is not the only way to access an LLM. Further possibilities include the usage of "jailbreaks" or DAN (do anything now) modes that circumvent the filtering system (usually done by prompt engineering), the usage of services that do not rigorously filter the output, or the usage of "uncensored" public models. Additional steps to bypass the filters as mentioned above are not necessary here, there is no need to conceal in intent.

3 AI for creating malware

Malware is the collective name for malicious software, like ransomware, worms or trojans. Often, the goal of attackers is to place malware on a target computer, be it via exploits or through social engineering. Measures like virus scanners explicitly fight such software, usually by identifying and blocking them from doing anything. This leads to an arms race between attackers, creating new malware, and defenders, updating their defenses to ward off new threats.

Therefore, it is interesting to see how AI affects the generation and usage of malware. In our research, we found several ways AI is used in this field. The models reach from LLMs over GANs (generative adversarial networks) to Reinforcement Learning systems, and they are used for several different purposes.

First, it allows actors with little to no technical skill to create malware more easily. They do not need a deep understanding of programming or how malware functions, and they can post their request in natural language.

Another concern is that AI could be used to autonomously write malware. This is a step up from just supporting human actors in the creation. LLMs can already write simple malware, but we did not find an AI that is capable of writing advanced, previously unknown malware (e.g. with intricate obfuscation methods or zero day exploits). The training data necessary about both malware and vulnerabilities would be very hard and expensive to create.

Next, AI can help modify malware. This is more realistic compared to creating malware from scratch, and there are several research papers about AI modifying malware. This happens mostly in a feature space, rather than on actual code level, and with the goal to avoid detection. However, this is in a rather academic environment, and we did not find any clues that these models are already deployed. Also, there is no polished tool, but only PoCs and research projects. This approach is only suited for highly skilled actors, in both malware and AI, and a good database is necessary to train such tools.

Lastly, we want to mention AI as part of malware. Here, AI is not creating the malware itself, but is instead integrated in the functionality. Often, the goal is to obfuscate the malware and hence prevent it from being detected. In the attempt to avoid detection, some malware programs have a so-called polymorphic engine, which alters the code of the malware while maintaining its functionality. An application of AI in this field is at least imaginable. There, the manipulation of the code would be determined by an AI model. At this point, there are no indications that such a model is in use, although there are many warnings of such a theoretical possibility. Another usage would be to train an AI model to mimic the user behavior, such that the actions of the malware are less prominent.

4 AI as an attacker

The holy grail for cybercriminal activities would be an AI that gets the target as an input (be it as an IP range or name) and does all the steps of a cyber attack completely autonomously. The strategizing and abstraction capabilities of recent AI technologies make them a prime candidate to build such a tool. Furthermore, from a penetration tester's point of view, this would be a useful tool to harden systems and decrease the time and expertise needed to perform penetration tests. A great deal of effort was and is taken to build such a tool, as it is an active research topic.

In this field, Reinforcement Learning systems are a common approach to the task, as they are capable of interacting with an environment, learn from it and create long term strategies. Recently, LLMs are also proposed as a solution to this problem. In our research, we found no tool that is capable of completely solving this task. There are, however, some tools that automate parts of the process (so called semi-automatization). Most of the time, these tools are academic projects or PoCs, and not especially user friendly or refined. Often, the scope of these tools is either very big or very small. On the big scale, attack path planning tools look at an abstract version of a target network and plan an optimal attack path. There is no active attack whatsoever. Similar to this are models that find optimal exfiltration paths for systems.

On the other hand, the small scope, there are tools that are explicitly trained on a single, specific network and try to start a successful attack there. This requires knowledge about the target network, as well as a training phase, which will hardly be unnoticed. Furthermore, a trained agent cannot easily be generalized to other networks. The environments of different systems and networks differ vastly in size and available actions. This makes it very hard to generalize. Also, it requires a very large training set to cover the abundance of options. These problems make the step from a proof of concept to a real world, general application a hard, probably currently unsolved problem. LLMs might be an approach to address the generalizability.

There are several tools supporting pentesting by AI assistants. After testing, we find that these tools serve mainly as a support for people trying to start an attack, and thereby they lower the threshold of entry.

Another approach for LLMs is, similar to the tools mentioned above, to automate certain parts of the attack chain. Here, we prominently have the reconnaissance stage, but also other steps like the analysis of server responses. The application of AI as a fully automated attack tool is a field that is highly researched. We expect to see more projects and tools coming, that are in this vein, especially ones that focus on the usage of LLMs and generative AI.

5 Additional links between AI and cybersecurity

In this section, we will provide an overview about other areas where AI and cybersecurity applications intersect. The most visible of them being the AI-ification of tools through the integration of LLMs, which is also happening in various other domains.

LLMs have already been integrated in IDEs (integrated development environments) and there exist plugins for reverse engineering or penetration testing tools. These plugins usually call the API of an LLM provider using a pre-built prompt and content from the respective application, the result is then displayed within the application. The benefit compared to using the LLM directly, for example in the browser (together with copying and pasting the respective content) is currently limited. In particular, because usually only the result is displayed and a follow-up query in the same context like in a chatbot is therefore not possible.

AI is also used in automated vulnerability detection. Due to its benefits for software engineering, this is an active field of research and multiple open source tools as well as commercial products are available. Analyzing open source applications using these tools is straightforward. Hence it will be crucial for open source projects to use these type of tools proactively before malicious actors do. Although the source code is usually required for the analysis, when combined with reverse engineering tools it is, to some extent, possible to use vulnerability detection methods on closed source applications. There exist projects which automate this process using an LLM. The results, however, vary widely depending on code complexity and obfuscation techniques.

Captchas are used ubiquitously to differentiate between automated bots and genuine human users through various challenges, such as distorted text or image recognition tasks and thus preventing malicious activities like spamming, brute-force attacks, and data scraping by requiring human-like responses. However, methods to defeat captchas have existed since their invention. Nowadays, AI is also used for this purpose and the selection of such tools and online services that yield good results is extensive.

AI is also used for password guessing. Based on the assumption that certain types of passwords are more likely to be used by humans, the rules for password candidates are learned from data, in contrast to existing password guessing tools where these rules are handcrafted. Note that there are plenty of password leaks that can serve as a database for training.

As everyone is eager to integrate AI in their processes, the dangers of embedded malware in the AI or data is growing larger. There are already cases, where malware is encoded in the parameters of Neural Networks, which barely changes the usability of the model. Malicious code can also be hidden in trained models that are frequently shared on specific platforms (6). Additionally, LLMs and the surrounding ecosystem might be abused to distribute malicious software to the users.

On the hardware side of attacks, side channel attacks are a well-known attack vector. However, a lot of technical skill and know how is needed. There exist PoCs about AI assisting in side channel attacks, potentially making them more available for lesser skilled attackers.

Legitimate software vendors started to use LLMs to give customer support. Similarly, malicious actors might also use LLMs to provide support to Malware as a Service users. Some even help less tech-savvy victims of ransomware attacks, who are often overwhelmed by the task of obtaining the cryptocurrency necessary to pay the ransom. Ransomware distributors are already offering support in acquiring and making payments. This process can be automated through the use of LLMs and offered in various languages in order to increase the conversion rate.

Bibliography

1. **Fang, Richard, et al.** *LLM Agents can Autonomously Hack Websites.* 2024.

2. **Oberhaus, Daniel.** *Prepare for AI Hackers.* https://www.harvardmagazine.com/2023/02/right-now-ai-hacking : Harvard Magazine, 2023.

3. **NCSC.** The near-term impact of AI on the cyber threat. *NCSC.* [Online] 24. 01 2024. https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat.

4. **Microsoft Threat Intelligence.** Staying ahead of threat actors in the age of AI. *Microsoft.* [Online] Microsoft, 14. 02 2024. https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/.

5. **Pa Pa, Yin Minn, et al.** *An Attacker's Dream? Exploring the Capabilities of ChatGPT for Developing Malware.* 2023.

6. **Cohen, David.** Data Scientists Targeted by Malicious Hugging Face ML Models with Silent Backdoor. *JFrog.* [Online] 27. 02 2024. https://jfrog.com/blog/data-scientists-targeted-by-malicious-hugging-face-ml-models-with-silent-backdoor/.