HADMÉRNÖK
VÉDELEMINFORMATIKA

20. évfolyam (2025) 4. szám 53–67.
DOI: 10.32567/hm.2025.4.4

Zoltán Kovács[1]

The Use of Artificial Intelligence in Cyberattacks, Part 2

Phases 1–4 of the Cyber Kill Chain Model

Abstract

The first part of this series of articles provided an overview of artificial intelligence (AI) and its various subfields (e.g. machine learning, generative AI, etc.), and showed that the Cyber Kill Chain (CKC) model, despite all its limitations, is suitable for achieving the goal of this series of articles, i.e. it can be used to demonstrate how attackers can use AI in cyberattacks. In order to develop adequate cyber defence against AI-assisted cyberattacks, it is necessary to know what AI-assisted tools attackers can use in each phase of the attack. This article focuses on the first four phases of the CKC model (reconnaissance, weaponization, delivery and exploitation) to examine where and how attackers are already using artificial intelligence in the first four phases of the Cyber Kill Chain model to achieve their goals, and how this helps attackers.

Keywords: artificial intelligence, cybersecurity, cyberattack, Cyber Kill Chain, OSINT, exploit, evasion, phishing, malware

Introduction

The emergence of AI in cybersecurity is a double-edged sword. On the one hand, it assists defensive personnel by offering significant potential for strengthening cyber defence systems and automated threat detection, On the other hand, however, it also provides malicious actors with an extremely effective tool for executing sophisticated, adaptive, and difficult-to-detect cyberattacks.[2] In order to build adequate

[1] Senior Lecturer, Ludovika University of Public Service, e-mail: zkovacs.24@gmail.com
[2] Abbadi–Lachkar 2024.

and effective defence against sophisticated AI-powered cyberattacks, it is necessary to understand what means and methods can be used by attackers. In this regard, Sun Tzu's words in *The Art of War* ring true:

"If you know the enemy and know yourself, you need not fear the result of a hundred battles. If you know yourself but not the enemy, for every victory gained you will also suffer a defeat. […] If you know neither the enemy nor yourself, you will succumb in every battle."[3]

This series of articles therefore focuses on providing a summary description of the AI-supported forms of attack currently in use in each phase of cyberattacks according to the Cyber Kill Chain model.

Numerous scientific articles, reports prepared by cybersecurity companies, blog posts written by experts, etc. have been published on AI-assisted attacks. However, based on the topic and main objective of this series of articles, they can typically be classified into three categories. The first category includes those that analyse an AI-assisted attack in depth. These descriptions help defenders prepare for a given attack and take preventive measures in the appropriate elements of their overall defence system (e.g., implementing new firewall rules, searching for indicators of compromise, or IoCs,[4] and loading them into defence tools for blocking, etc.). At the same time, defenders need to deal with each of these individually when developing their defences, gathering and synthesising the knowledge necessary to develop general defence principles.

The second category includes descriptions that provide a more general, comprehensive overview of several AI-supported attack types in a single document. These typically contain fewer technical details and instead provide an overview of several types of attacks. These documents help defenders develop more general defence methods and identify missing defence elements. However, these descriptions do not typically present AI-assisted attacks following the steps defined by the CKC, so defenders need to break them down and classify the individual parts of the attack method into the CKC phases if they want to take advantage of the defences built on this basis. The third category includes documents that describe AI-assisted attack forms according to the CKC model. On the one hand, the number of such works is low, and on the other hand, they typically do not attempt to summarise previously published cases. Such documents are typically published by cybersecurity companies, which present their own detected cases in this way. These documents help defenders develop effective defence systems according to the CKC, but even in this case, defenders need to process and synthesise many similar documents to achieve their goal of effective defence.

Based on the above, a document that synthesises and summarises numerous and wide-ranging AI-supported cyberattack methods based on the CKC could fill a gap, providing defenders with a general and comprehensive picture that they can use to

[3] Giles 2013.

[4] IoC (Indicator of Compromise) refers to digital traces that indicate the presence of a cyberattack, such as hacking or malware infection, on a system or network. These can include suspicious IP addresses, unknown file hashes, unusual network traffic or changes to system files.

develop effective defences. At the same time, further detailed studies are needed to develop defences (e.g. where and how traditional, i.e. non-AI-supported defence tools can still be used, what AI-supported cyber defence tools are available, how their operation can be coordinated with the help of, for example, SOAR[5], etc.). However, due to the depth and scope of the topic, these will be the subject of another series of articles. This series of articles focuses on summarising AI-supported attacks and breaking them down according to CKC, which is necessary to build a foundation for this.

The second part of this series of articles examines the options currently available to attackers in the first four phases of CKC if they wish to use AI-supported tools for their attacks and shows how these tools can help attackers achieve their goals.

The application of AI in the individual phases of the Cyber Kill Chain

The capabilities of artificial intelligence, and thus its use, significantly enhance attackers' capabilities in cyberattacks, enabling them to accelerate, automate and refine their execution. The following chapters describe in detail how attackers utilise artificial intelligence in each phase of the CKC (i.e. the cyberattack chain) and what specific technologies can be used by attackers to achieve their malicious goals. By using AI capabilities, attackers have achieved spectacular results in their attacks, for example, by developing phishing, creating polymorphic malicious code, or carrying out highly deceptive identity theft. Based on data reported by various organisations, the speed of detected cyberattacks has increased by 250% over the past 3 or 4 years, with detected security incidents taking less than an hour from compromise to exfiltration in 20% of cases.[6]

Phase 1: Reconnaissance

Reconnaissance is the first and essentially foundational phase of a cyberattack, the main purpose of which is to select the target(s) and gather relevant information about them in order to plan the attack strategy.[7] AI can significantly speed up and improve the efficiency of this process, allowing attackers to gain deeper and more accurate insight into the target(s)' system(s) and more easily find their potential weaknesses.[8] Such AI-supported tools and methods may be for example:

- *Automated open-source intelligence (OSINT) gathering and target profiling:* AI-based tools, especially natural language processing (NLP) and machine learning (ML), enable attackers to automatically collect and analyse vast amounts of publicly available data (e.g. social media profiles, company websites, patent databases, (cyber)security reports, job offers, news articles,

[5] SOAR: security orchestration, automation and response tool enabling coordinated, automated management and response to security incidents.
[6] Verton 2025.
[7] Hutchins et al. 2011.
[8] Schröer et al. 2025.

public code repositories, dark web, forums, etc.) and create profiles. From the information collected in this way, AI can, for example, identify the hierarchy of employees, email addresses and naming conventions, the technological infrastructure and software being used, network topologies, and even information about the organisational culture at the target company. This is why automated OSINT is critically important during reconnaissance, and AI can exponentially increase the efficiency of this process. Consider, for example, a case where AI-supported entity recognition can automatically identify key individuals and roles in a company's hierarchy and correlate them with public profiles (e.g. social media), which can form the basis for (even automated) targeted phishing attacks (e.g. spear-phishing,[9] whaling[10]).[11]

- *Network topology and infrastructure mapping:* AI can passively monitor network traffic and publicly available DNS records to automatically build the topology of the target network. The AI algorithms used by attackers can identify active services, open ports, operating system types and network device characteristics, often even when defenders attempt to hide them. This ability allows attackers to identify key vulnerabilities, recognise how to achieve a persistent presence in the compromised network and choose more effective attack vectors.[12]
- *Vulnerability detection and prediction:* Traditional vulnerability scanners used (also) by attackers are often signature-based and therefore easily detected by defenders. Instead, AI-based vulnerability scanners use machine learning algorithms to identify patterns in network protocols or system configurations that may indicate potential zero-day vulnerabilities. So-called *hunting* AIs are increasingly able to automatically identify these potential vulnerability patterns in code bases. In addition, AI-supported attack tools are able to search for and find correlations between large vulnerability databases (e.g. CVE) and target system parameters, predicting which vulnerabilities will be able to be successfully exploited by attackers.[13]
- *Analysis of employee behaviour and their weaknesses:* The analytical capabilities of advanced AI-powered attack systems can extend beyond information and communication systems to include the human factor. The AI algorithms used by attackers are capable of analysing employees' online spatial behaviour (e.g. posts, areas of interest), which they can use to identify individuals who are susceptible to psychological manipulation based social engineering attacks or to recognise poor cyber hygiene habits. This can form the basis for personalised attacks. The combination of AI-driven OSINT and AI-supported behavioural analysis provides attackers with extremely effective, *human-centric* vulnerability detection capabilities. This possibility shifts the focus of attacks

[9] Spear phishing is a type of cybercrime where attackers try to trick their victims (a specific person or organisation) with personalised messages to get confidential info. It's different from traditional phishing because it's way more targeted and personalised.

[10] Whaling is a highly targeted form of phishing that specifically targets individuals in senior positions (CEOs, CFOs, etc.), who are referred to as *whales*.

[11] Mirsky et al. 2023; Dees 2025.

[12] Yamin et al. 2021.

[13] Amster [s. a.]; Glynn 2025.

from merely detecting and exploiting technical vulnerabilities to detecting and exploiting human and organisational weaknesses, thereby bypassing traditional perimeter-based cyber defence systems and therefore making them less effective. Social engineering attacks based on psychological manipulation, including BEC (Business Email Compromise) threats, email scams and phishing, are becoming increasingly sophisticated thanks to AI-generated, deceptively realistic content. This also means that attacks no longer rely solely on technical weaknesses and vulnerabilities in systems, but increasingly exploit user inattention and poor digital hygiene, highlighting the importance of focusing on the human factor in defence.[14]

- *Acoustic side-channel attacks:* AI can also be used for acoustic side-channel attacks, for example, by analysing the sounds of keyboard keystrokes, it is possible to deduce which keys were originally pressed and thus recover the information entered in this way. This capability can be effective even with limited input data and provides new attack surfaces during the CKC reconnaissance phase.[15]

Overall, it can be said that the use of AI by attackers in the CKC reconnaissance phase significantly helps them to drastically reduce the manual effort and time required to detect targets, enabling them to identify their targets and vulnerabilities in greater numbers and with greater precision. This increase in efficiency allows attackers to move forward with their attacks more quickly, increasing the scale and number of their offensive operations, thereby further reducing the time available to defenders to identify and fix their own vulnerabilities.

Phase 2: Weaponization

In this phase, based on the information gathered during the reconnaissance phase, the attacker prepares and packages the malicious tool (exploit + payload) needed for the attack, i.e. prepares their cyber weapon. AI greatly increases the customisation, sophistication and resistance to detection by cyber defence tools of this cyber weapon.[16]

- *AI-based malware creation and adaptation:* Generative AI models, especially generative adversarial networks (GANs), could revolutionise the development of malicious code. GANs consist of two neural networks (generator and discriminator) that compete with each other and ultimately produce better output (results) than if only one network were operating and providing results for a given question or task.[17] In terms of malware, this means that the generator creates new malicious code, while the discriminator attempts to distinguish the newly generated code from existing malicious code or the

[14] Al-Azzawi et al. 2025; Cybersecurity Forecast 2025.
[15] Park et al. 2025.
[16] Navigating a New Threat Landscape 2024.
[17] Goodfellow et al. 2020.

original malware on which the development is based, feeding back the results so that the generator can continuously improve the newly generated malicious code. Within that, the main focus is on the new malware's ability to remain hidden from security systems. AI also enables the creation of polymorphic[18] and metamorphic[19] malware that continuously changes its code or structure, thereby increasing the likelihood that traditional cyber defence tools, such as signature-based antivirus software, will fail to detect it. In addition, AI can optimise the code of newly generated malware based on specific characteristics of the target system, such as the version of the operating system used in the target system, the security software installed, or even the hardware architecture, thereby increasing the likelihood of a successful attack. By using GANs, attackers can create new malware that is not yet included in the databases of defence systems, making traditional signature-based detection less effective against them. This capability fundamentally changes the nature of cyber threats, shifting the focus from static, relatively easy-to-identify patterns to dynamic, evolving and difficult-to-detect threats.

- Malicious, especially generative AI models specifically designed for this purpose (Generative Pre-trained Transformers, GPTs), such as WormGPT or FraudGPT, are capable of generating functional malware, including obfuscated[20] and polymorphic variants that can evade traditional detection methods. In addition to creating new malware, these large language models are also capable of modifying and rewriting the code of publicly available malicious software, thereby creating new variants. They can do this by translating them into other programming languages or by adding new features to existing malware, thereby modifying it. An example of the latter is adding AES encryption to existing code, which makes it more difficult to detect, track and analyse the new variant generated in this way.
- Due to their easy accessibility and simplicity of use, tools such as FraudGPT and WormGPT can serve as a kind of *cybercriminal starter kit* as they are capable of generating hard-to-detect malicious code, phishing sites and other hacking tools based on a few entered commands. This significantly lowers the

[18] Polymorphic malware: malicious software that can change its own code with each infection in order to make it more difficult to detect by traditional defence tools, such as antivirus software. Polymorphic malware changes its original code so that each time it infects a system, it has a different, unique code fragment and binary code structure, while its malicious function remains unchanged. Malware often includes a mutation engine that automatically generates new encryption keys and algorithms (translated by the author). itszótár.hu 2025.

[19] Metamorphic malware: in addition to the characteristics of polymorphic malware, it is also capable of changing, rewriting, translating, and editing its own code, so that each new version differs from the previous one, all without using an encryption key. This is a more advanced technique than polymorphic malware, which uses a key to modify the code. Metamorphic malware is more difficult to detect and identify because its code is constantly changing, it has no constant part that would identify it, and it does not return to its original form; each distributed version differs from the previous ones. This rewriting may involve changing the order of the code, adding unnecessary instructions (so-called *garbage code*), or replacing existing instructions with instructions that have equivalent functionality but different code (translated by the author). itszótár.hu 2025.

[20] Obfuscated code: the source code is deliberately made difficult to understand and unreadable, i.e. it is modified (e.g. by using meaningless variable names) to make it difficult to analyse, but the programme continues to function unchanged.

threshold for entering the world of cybercrime, meaning that even those with little or moderate IT knowledge can successfully enter this branch of crime.[21]

- *Exploit selection:* AI is capable of matching existing or newly generated malicious code with known exploits needed to benefit from it, as well as automating the selection of documents used as bait, thereby significantly increasing the effectiveness and speed of attacks.[22]
- *Intelligent exploit generation:* AI can also accelerate the discovery of new code-level vulnerabilities by analysing patterns in existing vulnerabilities. AI can also help attackers by automating the creation of exploit codes for previously known or newly discovered vulnerabilities. LLMs have been proven in a research environment to be capable of identifying vulnerabilities in various network and software systems and can even generate Proof-of-Concept (PoC) code for real exploits. Although this is still largely in the research phase, reinforcement learning (RL) and deep learning (DL) have the potential to be used to generate autonomous exploits, i.e. exploits that operate independently, without human intervention, and are capable of automatically intruding into a system and exploiting its vulnerabilities. RL agents can learn how to search for code patterns that may contain potential vulnerabilities in software and how to develop exploits that can exploit them (e.g. buffer overflow, format string vulnerability). AI can therefore be capable of code analysis, debugging and generating effective exploit code by simulating possible exploit vectors, which can drastically reduce the time and expertise required for exploit development.[23]
- *Generating spear-phishing and social engineering content:* natural language processing (NLP) and generative AI models (especially LLMs) are extremely effective at automatically generating hyper-personalised and realistic, i.e., customised, convincing and grammatically flawless phishing emails, messages (smishing[24]) or even voice calls (vishing[25]). Based on information obtained in the earlier stages of the attack, AI is able to mimic the target's communication style, incorporate personal or corporate information collected during the reconnaissance phase into messages intended for the target, and even dynamically change the type of content (e.g. urgent financial request, HR notification, IT support). AI-based deepfake technology, which enables the creation of extremely realistic but fake video and audio content, is a great help to attackers in producing such content. Attackers often use Generative Adversarial Networks (GANs), described earlier, to produce deepfake content.
- Generative AI also facilitates transnational and translingual cybercrime by enabling attackers from countries that do not speak the language of the target

[21] Usman et al. 2024; Arif et al. 2024.
[22] Salem–Mrian 2025.
[23] Zhu et al. 2025; Haurogné et al. 2024.
[24] Smishing: a form of phishing that takes place via text messages. Fraudsters use text messages to try to trick recipients into revealing confidential information, such as personal details, bank details or access codes.
[25] Vishing: a word combining the English words voice and phishing. An online fraud method in which attackers attempt to obtain sensitive data from victims over the phone.

country to bridge significant language gaps and create convincing phishing attacks or other attackable content with perfect grammar.
- All of these factors significantly increase the likelihood of successful social engineering attacks.[26]
- *Integration of AI-driven defence evasion techniques:* As a part of preparing the cyber weapon used for the attack, artificial intelligence elements can be integrated into the payload so that it can intelligently evade the deployed cyber defence systems. This may include the ability for AI to adaptively modify the attack execution logic, the file size of the attack code, its hash, or even its network communication in order to circumvent defences such as antivirus software, intrusion detection systems (IDS[27]), intrusion prevention systems (IPS[28]), or even sandboxing environments. AI can recognise and learn the behaviour of the security tools of the target system, even if they are AI-controlled, and carry out adversarial attacks that can deceive even defensive AI models. This ability allows attackers to not only evade existing security protocols, but also actively manipulate the AI-based detection mechanisms of the defence systems.[29]

Overall, it can be said that by exploiting the capabilities of AI, attackers will become more effective in the second, weaponization phase of CKC. AI significantly simplifies the creation of new variants or even completely new malicious code and cyber weapons, and its built-in functions help them remain hidden from security systems. AI is also a great help in producing deceptive content that can be used for attacks. This fundamentally transforms the weaponization phase, as attackers can develop more effective and adaptable tools with less effort, making traditional signature-based detection methods increasingly obsolete.

Phase 3: Delivery

In the delivery phase, the attacker delivers the weapons prepared in the previous phase (e.g. malware, exploit) to the target system or user. AI also assists the attacker in this phase of the attack by making delivery as covert and effective as possible. It increases the effectiveness of social engineering or phishing attacks used by the attacker.[30]

- *Optimised delivery channels and timing:* AI can analyse the network traffic of target systems, the security configurations of target systems and user behaviour patterns in order to choose the least conspicuous and least protected delivery route for delivering malicious code. For example, machine learning

[26] Falade 2023; Yu et al. 2024.
[27] IDS: intrusion detection system, a system that monitors and analyses network traffic in order to detect malicious activity and policy violations and then issues an alert when something is detected.
[28] IPS: intrusion prevention system, a security technology that monitors network traffic in real-time for malicious activity and automatically takes action to block or prevent threats like malware, exploits and unauthorised access.
[29] Fritsch et al. 2022; Singh–Cheema 2024.
[30] 'What is the cyber kill chain?', Microsoft [s. a.].

(ML) can identify periods that may be more favourable for carrying out an attack (e.g. outside of working hours, on weekends when users are less attentive or during periods when security systems are under higher load). AI can also dynamically select the delivery method tailored to the target, whether it be email, an infected website, a software update, or even a suggestion to use physical data storage. AI can also help attackers configure and deploy malware. AI-driven adaptive attack strategies allow attackers to adapt in real time to changes in the defence mechanisms and infrastructure of target systems, thereby helping to maximise the success and impact of the attack.[31]

- *Hiding delivery using steganography*[32] *and polymorphism:* AI can also help attackers deliver their cyber weapons (malicious payloads) to their targets in a way that is difficult to detect, for example by hiding them in legitimate files (such as images or audio files) using steganography techniques. Polymorphic code generation has already been discussed in the section on the *Weaponization* phase. However, this is also an effective aid for attackers during delivery, as the ability of malware to change its form allows it to avoid detection by signature-based detection tools, such as network intrusion detection systems (NIDS[33]), during delivery.[34]
- *Automated password cracking:* Even if the attacker already has certain data (e.g. username) to access the target system, they may still need additional valid authentication data (e.g. password). AI-based password cracking tools (e.g. PassGAN) can effectively assist the attacker, as they can quickly generate likely passwords by learning from databases of leaked credentials. This generative AI tool models the distribution of real passwords and produces highly accurate results without human intervention or input rules. This increases the speed at which certain authentication data can be cracked.[35]
- *Execution of intelligent phishing campaigns:* The capabilities of generative AI and natural language processing (NLP) are not only useful for generating personalised malicious content for attacks but are also extremely important for attackers in executing attack campaigns. AI can refine the delivery pattern of already hyper-personalised and realistic phishing messages in real time, taking into account the reactions of recipients and the defences of the attacked system (e.g. number of opens, number of clicks, ending up in the spam folder, etc.). AI can automatically modify the subject line, sender name, or even the structure of embedded malicious links if it detects that previous attempts have been unsuccessful and/or have been detected by the target's security systems. This adaptive behaviour significantly increases the effectiveness of attack campaigns and the likelihood of avoiding detection. Attack campaigns

[31] Potter et al. 2025; Kumar–Chauhan 2025.

[32] Steganography is a branch of computer science/cryptography that aims to hide secret messages, not by encrypting the message itself (using cryptography), but by hiding the fact of communication from others, most often by embedding it in media files (e.g. images, sounds).

[33] NIDS: network intrusion detection system, which monitors network traffic and issues alerts when it detects suspicious activity or violation of rules.

[34] Fadhil 2025; Potter et al. 2025.

[35] Hitaj et al. 2019.

thus become more resistant to the initial defensive responses of target systems, creating a continuous learning loop for attackers and obsoleting static cyber defence mechanisms.
- AI-generated phishing emails have a worrying success rate, and fully AI-based, automated phishing attacks now have a success rate that has overtaken attacks carried out with human involvement.
- In this phase, AI-based chatbots also assist attackers, as they can be used to automate real-time communication with targets in specific cases in such a way that they are almost indistinguishable from a human being on the other end of the conversation. This greatly facilitates the collection of authentication data, for example.[36]
- *Attacks against the supply chain:* Attackers can also use AI in supply chain attacks, for example to inject malicious code into the software supply chain. Vulnerabilities in AI models, such as data poisoning or Trojan models, can also help attackers deliver malicious code if they use public AI models at the target organisation.[37]

Overall, it can be said that in the third phase of CKC, the use of AI in delivery effectively increases the success rate of attacks by facilitating evasion, supporting password cracking, and enabling highly targeted and persuasive campaigns that overcome human perception and language barriers. The latter, as AI-generated content becomes increasingly indistinguishable from legitimate communication, further reinforces the fact that humans will be the most vulnerable part of the defence of information and communication systems.

Phase 4: Exploitation

In the exploitation phase, the cyber weapon that has been prepared and delivered becomes active and exploits one (or more) vulnerabilities in the target system in order to gain access to the target network or one of its components. In this case, AI can also increase the speed and accuracy of the attack, help evade the target network's defence systems, combine vulnerabilities that can be exploited by the attacker, and improve the chances of obtaining the authentication data needed for impersonation, thereby helping the attacker to raise their level of authorisation.[38]

- *Autonomous exploit execution and optimisation:* AI, especially reinforcement learning (RL), and AI-driven attack strategies such as adaptive exploit execution, enable attackers to adapt in real time to the defence mechanisms of the target system and changes in the infrastructure of the victim, thereby significantly increasing the likelihood of a successful attack. Based on the responses of the target system, AI is capable of dynamically modifying the execution of the

[36] Dean 2025; Hoxhunt [s. a.].
[37] Fernández 2025; Blake 2025.
[38] Hutchins et al. 2011.

exploit in real time. If an exploit does not work the first time, AI can analyse the errors, the error codes returned, or even the behaviour of the system, and modify the attack parameters accordingly to ensure the attack is successful. This capability minimises the need for human intervention and speeds up the exploitation phase of the attack.[39]

- *Intelligent evasion techniques during the exploitation phase:* In the Weaponization phase, attackers can use the elements integrated into the payload described in Phase 2. AI-controlled exploits are capable of detecting and adapting to the target system and its security elements, actively circumventing cybersecurity elements such as antivirus (AV) software, intrusion detection systems (IDS/IPS) and sandbox environments. This is possible because AI can learn the detection mechanisms of defence systems and carry out adversarial attacks that deliberately deceive even machine learning-based detection models. For example, an attacking AI can change the exploit code, or the sequence of system calls to appear legitimate to the defensive AI or even manipulate the inputs of defensive AI models to cause them to make incorrect classifications or decisions.
- In order to evade detection by security systems used in the target system, they may also attack AI-based protection solutions themselves, for example by manipulating input data, which leads to incorrect predictions or event classification. Such attacks may include poisoning training databases, continuously and subtly altering input data (e.g. by adding noise), or modifying the parameters of pre-trained models (model manipulation). Attackers do all this in order to negatively influence the accuracy of defence AI-based systems, thus also resulting in incorrect decision-making and/or event classification. These, in turn, directly affect the ability of defence AI to detect and report attack attempts. These attacks are already effectively a direct *AI against AI* type of ongoing battle, which will intensify in the near future. On both the offensive and defensive sides, speed, the ability to respond to changes, and ultimately adaptive learning will make the use of AI alongside humans indispensable.[40]
- *Application of polymorphism and metamorphism during exploitation:* Malware generated with the help of AI can change its code or behaviour in real time if it detects signs or attempts at detection. This makes it even more difficult for defence mechanisms such as dynamic analysis or behaviour-based detection to work, as the malicious software constantly *changes its shape* while running. Attackers can also use this feature of malware to their advantage during the exploitation phase.[41]

AI can also provide significant assistance to cyber attackers during the exploitation phase of a CKC. Autonomous exploit execution, the use of intelligent evasion techniques, and the dynamic, autonomous adaptation of malicious code to a given target

[39] Rohlf 2025; Luong et al. 2025.
[40] Nobles 2024; Syed 2025.
[41] Alrzini–Pennington 2020; SentinelOne 2025.

system all make detection significantly more difficult and increase the likelihood of a successful intrusion. In addition, adaptive exploitation means that even patched vulnerabilities can be re-exploited by attackers through new, AI-generated attack vectors, requiring continuous, real-time vulnerability assessment and remediation on the defence side. This requires a lot of resources from them.

Conclusions

The second part of this series of articles examined where and how attackers use artificial intelligence in the first four phases of the Cyber Kill Chain model to achieve their goals, and how this helps them. In 2024, major cybersecurity companies wrote in their annual studies that attackers typically used artificial intelligence in the first two phases of the CKC, i.e. reconnaissance and weaponization, for example, to create fake profiles, processing large amounts of data collected from stolen or publicly available information when mapping targets, phishing, creating deepfake videos, generating malicious code, and possibly controlling DDoS attacks in later phases. Based on an examination of the first four phases, it can already be said that the development of AI over the past 1.5–2 years has also led to significant advances in its malicious use. In order for the defence side to keep up with this, it is necessary to understand where and how attackers use AI in the later phases, and further research is needed to examine what options are available to defenders and where further developments are needed to create effective defences. The next part of the series continues this investigation and examines AI-assisted attacks in the last three (plus one) phases of the CKC. It then discusses the current challenges and trends in the use of AI on the offensive side, making recommendations for the direction of further research.

References

Abbadi, Driss – Lachkar, Abdelkader (2024): Cyber Threats in the Age of Artificial Intelligence. Exploiting Advanced Technologies and Strengthening Cybersecurity. *International Journal of Science and Research Archive,* 13(1), 2576–2588. Online: https://doi.org/10.30574/ijsra.2024.13.1.1961

Al-Azzawi, Mays – Doan, Dung – Sipola, Tuomo – Hautamäki, Jari – Kokkonen, Tero (2025): Red Teaming with Artificial Intelligence-Driven Cyberattacks: A Scoping Review. *arXiv:2503.19626.* Online: https://doi.org/10.48550/arXiv.2503.19626

Alrzini, Joma – Pennington, Diane (2020): A Review of Polymorphic Malware Detection Techniques. *International Journal of Advanced Research in Engineering and Technology (IJARET),* 11(12), 1238–1247. Online: https://doi.org/10.34218/IJARET.11.12.2020.119

Amster, Alex [s. a.]: Automating Vulnerability Detection in Networks with AI. *AllStarsIT,* s. a. Online: www.allstarsit.com/blog/automating-vulnerability-detection-in-networks-with-ai

Arif, Aftab – Khan, Muhammad Ismaeel – Khan, Ali Raza A (2024): An Overview of Cyber Threats Generated by AI. *International Journal of Multidisciplinary Sciences and Arts,* 3(4), 67–76. Online: https://doi.org/10.47709/ijmdsa.v3i4.4753

Blake, Harrison (2025): *AI-Powered Threats in Supply Chains: A Looming Cybersecurity Challenge.* ResearchGate. Online: www.researchgate.net/profile/Harrison-Blake-2/publication/389274676_AI-Powered_Threats_in_Supply_Chains_A_Looming_Cybersecurity_Challenge/links/67bc8c29461fb56424e8923e/AI-Powered-Threats-in-Supply-Chains-A-Looming-Cybersecurity-Challenge.pdf

Cybersecurity Forecast 2025 (2025): Google Cloud Security. Online: https://cloud.google.com/blog/topics/threat-intelligence/cybersecurity-forecast-2025

Dean, B. (2025): New Report: Over 80% of Cyberattacks Now Use AI. *Programs.com*, 8 August 2025. Online: https://programs.com/resources/ai-cyberattack-stats/

Dees, Mels (2025): CrowdStrike Introduces Tools to Block Malicious AI Models. *Techzine Global,* 30 April 2025. Online: www.techzine.eu/news/security/130990/crowdstrike-introduces-tools-to-block-malicious-ai-models/

Fadhil, Ammar (2025): Enhancing Data Security: A Hybrid Approach of AI-Driven Steganography and Encryption. *The Indonesian Journal of Computer Science,* 14(2). Online: https://doi.org/10.33022/ijcs.v14i2.4759

Falade, Polra V. (2023): Decoding the Threat Landscape: ChatGPT, FraudGPT, and WormGPT in Social Engineering Attacks. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology,* 9(5), 185–198. Online: https://doi.org/10.32628/CSEIT2390533

Fernández, Rodrigo (2025): AI-Driven Supply Chain Attacks: The New Cyber Risk in 2025. *NeuralTrust,* 25 September 2025. Online: https://neuraltrust.ai/blog/ai-driven-supply-chain-attacks

Fritsch, Lothar – Jaber, Aws – Yazidi, Anis (2022): An Overview of Artificial Intelligence Used in Malware. In Zouganeli, Evi – Yazidi, Anis – Mello, Gustavo – Lind, Pedro (eds.): *Nordic Artificial Intelligence Research and Development.* Cham: Springer International Publishing, 41–51. Online: https://doi.org/10.1007/978-3-031-17030-0_4

Giles, Lionel (2013): *Sun Tzu on the Art of War.* London: Routledge. Online: https://doi.org/10.4324/9781315030081

Glynn, Fergal (2025): AI Vulnerability Scanner: 6 Practical Metrics Every Security Team Should Monitor. *Mindgard,* 25 August 2025. Online: https://mindgard.ai/blog/ai-vulnerability-scanner-metrics

Goodfellow, Ian et al. (2020): Generative Adversarial Networks. *Communications of the ACM,* 63(11), 139–144. Online: https://doi.org/10.1145/3422622

Haurogné, Jean – Basheer, Nihala – Islam, Shareeful (2024): Vulnerability Detection Using BERT based LLM Model with Transparency Obligation Practice towards Trustworthy AI. *Machine Learning with Applications,* 18. Online: https://doi.org/10.1016/j.mlwa.2024.100598

Hitaj, Briland – Gasti, Paolo – Ateniese, Giuseppe – Perez-Cruz, Fernando (2019): PassGAN: A Deep Learning Approach for Password Guessing. *arXiv:1709.00440.* Online: https://doi.org/10.48550/arXiv.1709.00440

Hutchins, Eric M. – Cloppert, Michael J. – Amin, Rohan M. (2011): Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains. *Leading Issues in Information Warfare & Security Research,* 1(1), 1–14.

ITszótár.hu (2025): Metamorf és polimorf kártevők: Ezen kártékony szoftverek működésének magyarázata. *ITszotar.hu,* 15 May 2025. Online: https://itszotar.hu/metamorf-es-polimorf-kartevok-ezen-kartekony-szoftverek-mukodesenek-magyarazata/

Kumar, Ankit – Chauhan, Nidhi (2025): AI-Driven Optimization for Enhancing Performance, Efficiency, and Personalization in Content Delivery Networks. *International Journal of Computer Techniques,* 12(3), 1–9. Online: https://ijctjournal.org/wp-content/uploads/2025/06/AI-Driven-Optimization-for-Enhancing-Performance-Efficiency-and-Personalization-in-Content-Delivery-Networks.pdf

Luong, Phung D. et al. (2025): xOffense: An AI-driven Autonomous Penetration Testing Framework with Offensive Knowledge-Enhanced LLMs and Multi Agent Systems. *arXiv:2509.13021v1.* Online: https://arxiv.org/html/2509.13021v1

Microsoft [s. a.]: What is the Cyber Kill Chain? *Microsoft Security,* s. a. Online: www.microsoft.com/en-us/security/business/security-101/what-is-cyber-kill-chain

Mirsky, Yisroel et al. (2023): The Threat of Offensive AI to Organizations. *Computers & Security,* 124. Online: https://doi.org/10.1016/j.cose.2022.103006

Navigating a New Threat Landscape (2024). Darktrace. Online: www.darktrace.com/resources/navigating-a-new-threat-landscape

Nobles, Calvin (2024): The Weaponization of Artificial Intelligence in Cybersecurity: A Systematic Review. *Procedia Computer Science,* 239, 547–555. Online: https://doi.org/10.1016/j.procs.2024.06.206

Park, Jin H. – Ayati, Seyyed A. – Cai, Yichen (2025): Improving Acoustic Side- Channel Attacks on Keyboards Using Transformers and Large Language Models. *arXiv:2502.09782.* Online: https://doi.org/10.48550/arXiv.2502.09782

Phishing Trends Report (Updated for 2025) [s. a.]. *Hoxhunt,* s. a. Online: https://hoxhunt.com/guide/phishing-trends-report

Potter, Yujin et al. (2025): *Frontier AI's Impact on the Cybersecurity Landscape.* arXiv:2504.05408. Online: https://doi.org/10.48550/arXiv.2504.05408

Rohlf, Chris (2025): AI and the Software Vulnerability Lifecycle. *Center for Security and Emerging Technology,* 8 August 2025. Online: https://cset.georgetown.edu/article/ai-and-the-software-vulnerability-lifecycle/

Salem, Maher – Mrian, Mohammad (2025): *AI-Driven Penetration Testing: Automating Exploits with LLMs and Metasploit-A VSFTPD Case Study.* 2025 International Conference on New Trends in Computing Sciences (ICTCS), Amman, Jordan, 89–96. Online: https://doi.org/10.1109/ICTCS65341.2025.10989363

Schröer, Saskia L. – Pajola, Luca – Castagnaro, Alberto – Apruzzese, Giovanni – Conti, Mauro (2025): Exploiting AI for Attacks: On the Interplay between Adversarial AI and Offensive AI. *arXiv:2506.12519v2.* Online: https://arxiv.org/html/2506.12519

SentinelOne (2025): What is Polymorphic Malware? Examples & Challenges. *SentinelOne,* 20 August 2025. Online: www.sentinelone.com/cybersecurity-101/threat-intelligence/what-is-polymorphic-malware/

Singh, Bhagwant – Cheema, Sikander S. (2024): Emerging Trends in AI-Powered Malware Detection: A Review of Real-Time and Adversarially Resilient Techniques. *Tuijin Jishu/Journal of Propulsion Technology,* 45(4).

Syed, Shoeb A. (2025): Adversarial AI and Cybersecurity: Defending Against AI- Powered Cyber Threats. *Iconic Research and Engineering Journals,* 8(9), 1030–1041.

Usman, Yusuf – Upadhyay, Aadesh – Chataut, Robin – Gyawali, Prashnna K. (2024): Is Generative AI the Next Tactical Cyber Weapon for Threat Actors? Unforeseen Implications of AI Generated Cyber Attacks. *arXiv:2408.12806.* Online: https://doi.org/10.48550/arXiv.2408.12806

Verton, Dan (2025): The 2025 Cybersecurity Pulse Report. *iSMG,* 30 May 2025. Online: https://ismg.io/resource/rsac-2025-pulse/

Yamin, Muhammad M. – Ullah, Mohib – Ullah, Habib – Katt, Basel (2021): Weaponized AI for Cyber Attacks. *Journal of Information Security and Applications,* 57. Online: https://doi.org/10.1016/j.jisa.2020.102722

Yu, Jingru et al. (2024): The Shadow of Fraud: The Emerging Danger of AI-powered Social Engineering and its Possible Cure (Version 1). *arXiv:2407.15912.* Online: https://doi.org/10.48550/ARXIV.2407.15912

Zhu, Yuxuan et al. (2025): CVE-Bench: A Benchmark for AI Agents' Ability to Exploit Real-World Web Application Vulnerabilities. *arXiv:2503.17332v4.* Online: https://arxiv.org/html/2503.17332v4