TRAFICOM
Finnish Transport and Communications Agency

The security threat of AI-enabled cyberattacks

Traficom publications
31/2022

Table of Contents

TRAFICOM Finnish Transport and Communications Agency
National Cyber Security Centre

Date of publication
13.12.2022

Title of publication
The security threat of AI-enabled cyberattacks

Author(s)
Matti Aksela, Samuel Marchal, Andrew Patel, Lina Rosenstedt, WithSecure

Commissioned by, date
Finnish Transport and Communications Agency Traficom

Publication series and number	ISSN(online) 2669-8757
Traficom Research Reports 31/2022	ISBN(online) 978-952-311-828-7

Keywords
AI, machine learning, information security, cyber security, cyberattacks

Abstract

The topic of AI-enabled cyberattacks surfaced around five years ago with examples of generative AI models able to automate both spear-phishing attacks and vulnerability discovery. Since then, social engineering and impersonation attacks supported by AI have occurred, causing millions of dollars in financial losses[1]. Current rapid progress in AI research, coupled with the numerous new applications it enables, leads us to believe that AI techniques will soon be used to support more of the steps typically used during cyberattacks. This is the reason why the idea of AI-enabled cyberattacks has recently gained increased attention from both academia and industry, and why we are starting to see more research devoted to the study of how AI might be used to enhance cyberattacks.

A study from late 2019 illustrated that over 80% of decision-makers were concerned with AI-enabled cyberattacks and predicted that these types of attacks may go mainstream in the near future[2]. Current AI technologies already support many early stages of a typical attack chain. Advanced social engineering and information gathering techniques are such examples. AI-enabled cyberattacks are already a threat that organisations are unable to cope with. This security threat will only grow as we witness new advances in AI methodology, and as AI expertise becomes more widely available.

This report aims to investigate the security threat of AI-enabled cyberattacks by summarising current knowledge on the topic. AI technology is currently able to enhance only a few attacker tactics, and it is likely only used by advanced threat actors such as nation-state attackers. In the near future, fast-paced AI advances will enhance and create a larger range of attack techniques through automation, stealth, social engineering or information gathering. Therefore, we predict that AI-enabled attacks will become more widespread among less skilled attackers in the next five years. As conventional cyberattacks will become obsolete, AI technologies, skills and tools will become more available and affordable, incentivising attackers to make use of AI-enabled cyberattacks.

The cybersecurity industry will have to adapt to cope with the emergence of AI-enabled cyberattacks. For instance, biometric authentication methods may become obsolete because of advanced impersonation techniques enabled by AI. New prevention and detection mechanisms will also need to be developed to counter AI-enabled cyberattacks. More automation and AI technology will also need to be used in defence solutions to match the speed, scale and sophistication of AI-enabled cyberattacks. This may lead to an asymmetrical fight between attackers having unrestricted use of AI technologies and defenders being constrained by the upcoming regulation on AI applications.

Contact person	**Language**	**Confidence status**	**Pages, total**
Markus Mettälä, Juhani Eronen	English	Public	30

Distributed by	**Published by**
Transport and Communications Agency, National Cyber Security Centre Finland	Transport and Communications Agency, National Cyber Security Centre Finland

[1]https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/
[2]Forrester – The emergence of offensive AI (2019)

TRAFICOM Liikenne- ja viestintävirasto
Kyberturvallisuuskeskus

Julkaisupäivämäärä
13.12.2022

Julkaisun nimi
Tekoälyn mahdollistamat kyberhyökkäykset

Tekijät
Matti Aksela, Samuel Marchal, Andrew Patel, Lina Rosenstedt, WithSecure

Toimeksiantaja ja asettamispäivämäärä
Liikenne- ja viestintävirasto Traficom

Julkaisusarjan nimi ja numero	ISSN(online) 2669-8757
Traficomin tutkimuksia ja selvityksiä 31/2022	ISBN(online) 978-952-311-828-7

Asiasanat
Tekoäly, koneoppiminen, kyberturvallisuus, tietoturva, tietosuoja, kyberhyökkäykset

Tiivistelmä

Tekoälyn mahdollistamat kyberhyökkäykset nousivat esille noin viisi vuotta sitten generatiivisten tekoälymallien vauhdittamana. Tällaiset mallit kykenevät aikaisempaa paremmin automatisoimaan sekä kohdennettuja tietojenkalasteluhyökkäyksiä että haavoittuvuuksien etsimistä. Sittemmin tekoälyn tukemia sosiaalisen manipuloinnin ja imitaatioon perustuvia hyökkäyksiä on tapahtunut, mikä on jo aiheuttanut miljoonien dollarien taloudellisia menetyksiä[1]. Tekoälytutkimuksen tämänhetkinen nopea edistyminen yhdistettynä lukuisiin uusiin käyttötarkoituksiin antaa syytä uskoa, että tekoälytekniikoita tullaan pian käyttämään tukemaan niitä vaiheita, joita tyypillisesti suoritetaan manuaalisesti kyberhyökkäysten aikana. Tästä syystä ajatus tekoälyn tukemista kyberhyökkäyksistä on viime aikoina saanut enemmän huomiota sekä tiedemaailmassa että teollisuudessa. Vaikka ei olekaan todennäköistä, että tekoäly vielä loisi täysin uudenlaisia hyökkäyksiä, näemme jatkuvasti enemmän tutkimusta siitä, miten tekoälyä voitaisiin käyttää kyberhyökkäyksien radikaaliinkin tehostamiseen ja skaalaamiseen.

Vuoden 2019 lopulla tehty tutkimus osoitti, että yli 80 % päättäjistä oli huolissaan tekoälyn mahdollistamista kyberhyökkäyksistä ja ennusti, että tämäntyyppiset hyökkäykset voivat yleistyä lähitulevaisuudessa[2]. Nykyiset tekoälytekniikat tukevat jo monia tyypillisen hyökkäysketjun alkuvaiheita. Kehittynyt käyttäjän manipulointi ja tiedonkeruutekniikat ovat tällaisia esimerkkejä. Tekoälyn tukemat kyberhyökkäykset ovat jo uhka, josta monet organisaatiot eivät pysty selviytymään. Tämä turvallisuusuhka vain kasvaa, kun näemme uusia edistysaskeleita tekoälymenetelmissä ja kun asiantuntemus tekoälystä tulee laajemmin saataville.

Tämän raportin tarkoituksena on esitellä tekoälyn mahdollistamien kyberhyökkäysten turvallisuusuhkaa tekemällä yhteenveto aiheesta olemassa olevasta nykyisestä tiedosta. Tekoälyteknologia pystyy tällä hetkellä parantamaan vain muutamia hyökkääjän taktiikoita, ja sitä käyttävät todennäköisesti vain edistyneet uhkatoimijat, kuten kansallisvaltioiden hyökkääjät. Lähitulevaisuudessa tekoälyn nopeatempoinen kehitys todennäköisesti parantaa ja luo laajemman valikoiman mahdollisuuksia hyökkäysten automatisoinnin, käyttäjämanipuloinnin ja tiedonkeruun saralla. Näin ollen voidaan ennustaa, että tekoälyn tukemat hyökkäykset yleistyvät vähemmän taitavien hyökkääjien keskuudessa seuraavan viiden vuoden aikana. Kun tavanomaiset kyberhyökkäykset vanhenevat, tekoälyteknologiat, -taidot ja -työkalut tulevat helpommin saataville ja edullisemmiksi, mikä kannustaa hyökkääjiä hyödyntämään tekoälyn tukemia kyberhyökkäyksiä.

Kyberturvallisuusalan on sopeuduttava selviytyäkseen tekoälyä hyödyntävistä kyberhyökkäyksistä. Esimerkiksi biometriset todennusmenetelmät voivat vanhentua tekoälyn mahdollistamien kehittyneiden imitaatiotekniikoiden vuoksi. Uusia ehkäisy- ja havaitsemismekanismeja on myös kehitettävä tekoälyn tukemien kyberhyökkäysten torjumiseksi. Lisää automaatiota ja tekoälyteknologiaa on käytettävä myös puolustusratkaisuissa, jotta ne vastaisivat tekoälyn tukemien kyberhyökkäysten nopeutta, laajuutta ja kehittyneisyyttä. Tämä voi johtaa epäsymmetriseen taisteluun tekoälytekniikoita rajoittamattomasti käyttävien hyökkääjien ja puolustajan välillä, jota rajoittaa tekoälyä koskevat lait ja säännökset.

Yhteyshenkilö	**Raportin kieli**	**Luottamuksellisuus**	**Kokonaissivumäärä**
Markus Mettälä, Juhani Eronen	englanti	Julkinen	30

Jakaja	**Kustantaja**
Liikenne- ja viestintävirasto Traficom, Kyberturvallisuuskeskus	Liikenne- ja viestintävirasto Traficom, Kyberturvallisuuskeskus

[1]https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/
[2]Forrester – The emergence of offensive AI (2019)

TRAFICOM Transport- och kommunikationsverket
Cybersäkerhetscentret

Utgivningsdatum
13.12.2022

Publikation
AI-aktiverade cyberangrepp

Författare
Matti Aksela, Samuel Marchal, Andrew Patel, Lina Rosenstedt, WithSecure

Tillsatt av och datum
Transport- och kommunikationsverket Traficom

Publikationsseriens namn och nummer	ISSN(webbpublikation) 2669-8757
Traficoms forskningsrapporter och utredningar 31/2022	ISBN(webbpublikation) 978-952-311-828-7

Ämnesord
Artificiell intelligens, maskininlärning, cybersäkerhet, informationssäkerhet, dataskydd, cyberangrepp

Sammandrag

Cyberangrepp som artificiell intelligens (AI) möjliggör lyftes fram för cirka fem år sedan och fick fart från generativa modeller för artificiell intelligens. Sådana modeller kan bättre än tidigare automatisera både riktade nätfiskeangrepp och letande efter sårbarheter. Senare har det skett angrepp som baserar sig på AI-stödd social manipulering och imitation, vilket redan har orsakat ekonomiska förluster på flera miljoner dollar[1].

Det snabba framskridandet av forskningen av artificiell intelligens kombinerat med flera nya användningsändamål ger anledning att tro att AI-teknologier mycket snart kommer att användas som stöd för de faser som i allmänhet görs manuellt under cyberangrepp. Av denna anledning har tanken på AI-stödda cyberangrepp under de senaste tiderna fått mer uppmärksamhet både i den vetenskapliga världen och inom industrin. Även om det inte är sannolikt att artificiell intelligens skulle skapa helt nya typer av angrepp ser vi hela tiden mer forskning om hur artificiell intelligens skulle kunna utnyttjas för radikalt effektivare och skalbara cyberangrepp.

Undersökningen från slutet av 2019 visade att över 80 procent av beslutsfattarna oroade sig över de cyberangrepp som artificiell intelligens möjliggör och förutspådde att angrepp av detta slag kan bli vanligare inom den närmaste framtiden[2]. Dagens AI-teknologier stöder redan flera inledningsskeden i en typisk kedja av angrepp. Avancerad manipulering av användare och datainsamlingstekniker är exempel på sådana. AI-stödda cyberangrepp är redan ett hot som många organisationer inte kan klara av. Detta säkerhetshot ökar när vi ser nya framsteg i AI-metoder och när sakkunskap om artificiell intelligens blir tillgänglig på ett mer omfattande sätt.

Syftet med denna rapport är att presentera den säkerhetshot som artificiell intelligens medför i form av en sammanfattning av den information som finns tillgänglig i dag. AI-teknologin kan för tillfället endast förbättra några av angriparens taktiker och används sannolikt endast av avancerade hotaktörer, t.ex. angripare i nationalstater. I den närmaste framtiden kommer den snabbare utvecklingen av artificiell intelligens sannolikt att förbättra och skapa ett bredare urval av möjligheter för automatisering av angrepp, manipulering av användare och datainsamling. Därför är det möjligt att förutspå att AI-stödda angrepp blir allt vanligare hos mindre skickliga angripare under de följande fem åren. När de vanliga cyberangreppen blir föråldrade, blir AI-teknologier, -kunskaper och -verktyg mer lättillgängliga och förmånligare, vilket uppmuntrar angripare att utnyttja AI-stödda cyberangrepp.

Cybersäkerhetsbranschen måste anpassa sig för att kunna klara av de cyberangrepp som utnyttjar artificiell intelligens. Till exempel biometriska autentiseringsmetoder kan föråldras på grund av avancerade imitationsteknologier som artificiell intelligens möjliggör. Man ska också utveckla nya mekanismer för förebyggande och detektering i syfte att avvärja AI-stödda cyberangrepp. Ytterligare automatisering och AI-teknologi ska också användas för försvarslösningar så att de skulle motsvara snabbheten, omfattningen och utvecklingen av AI-stödda cyberangrepp. Detta kan leda till en asymmetrisk bekämpning mellan angripare som använder AI-teknologier utan begränsningar och försvaret som begränsas av lagar och bestämmelser om artificiell intelligens.

Kontaktperson	**Språk**	**Sekretessgrad**	**Sidoantal**
Markus Mettälä, Juhani Eronen	engelska	offentlig	30

Distribution	**Förlag**
Transport- och kommunikationsverket Traficom, Cybersäkerhetscentret	Transport- och kommunikationsverket Traficom, Cybersäkerhetscentret

[1]https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/
[2]Forrester – The emergence of offensive AI (2019)

List of abbreviations

AI	Artificial Intelligence
CVE	Common Vulnerabilities and Exposures
C&C	Command-and-Control
DL	Deep Learning
DNN	Deep Neural Networks
GAN	Generative Adversarial Network
HIDS	Host Intrusion Detection System
ML	Machine Learning
NIDS	Network Intrusion Detection System
NLG	Natural Language Generation
NSFW	Not Safe for Work
OSINT	Open-Source Intelligence
5-6G	5^{th}–6^{th} generation mobile network

AI capabilities for cyberattacks

Artificial intelligence, machine learning and deep learning are overlapping fields that have benefitted from numerous recent advances, providing new capabilities and enabling new applications. While these capabilities have been designed for benign applications such as prediction, generation, data analysis, and information retrieval capabilities, such functionalities could also be used to improve conventional cyberattacks.

Artificial Intelligence (AI) refers to the ability of a machine, such as a computer system or a computer program, to perform tasks that are typically associated with intelligent beings such as humans or animals. Intelligent capabilities associated with AI systems include the ability to reason, solve problems, discover meaning, generalise, plan, and learn from experience. These capabilities are used by human attackers when designing and launching cyberattacks against information systems. On a conceptual level, one might imagine how a sufficiently intelligent AI could be used to craft and launch cyberattacks by replacing the currently manual process of finding vulnerabilities and designing attacks to exploit them. However, AI is a broad concept encompassing many subfields such as expert systems, robotics and fuzzy logic.

Most current AI subfields do not represent anything close to human-level intelligence and would not be able to automatically craft or launch cyberattacks. On the other hand, the AI subfield of machine learning, which has received recent attention due to tremendous progress, is able to outperform humans in several “intelligence” tasks such as image classification, text translation, and playing games such as Chess or Go. Most of the current hype surrounding AI is related to machine learning applications, and the term AI is often used as a more generic and shortcut term to describe machine learning.

Machine learning is the term used to describe a type of expert system that uses data to learn, make decisions and improve through experience without following explicit, hard-coded instructions. It uses algorithms and statistical models to analyse and draw inferences from patterns in data. Machine learning is different from most AI subfields which require explicit, imperative instructions or rules to produce outputs and results. In contrast, machine learning uses adaptive algorithms which learn their behaviour from data in an autonomous manner. Machine learning is divided into three prominent types, namely supervised learning which is designed to perform or replicate known tasks (task-driven), unsupervised learning which is designed to extract hidden information from data (data-driven) and reinforcement learning which is designed to learn new tasks via trial-and-error while trying to maximise a defined reward (trial-error-driven).

Deep Learning or Deep Neural Networks (DNNs) are a type of machine learning algorithm that attain high performance in the automated processing of natural data such as text, images, sound or video. Recent advances in deep learning are the main reason for the current hype around AI and machine learning. Deep learning techniques achieve unmatched and often superhuman performance in complex tasks such as image classification, text translation or playing complex games. Machine learning and deep learning actually deliver on many long-lasting promises expected from AI systems. They can reason, solve problems, discover meaning and improve through experience in an autonomous manner, only using data. Algorithmic improvements, large data availability and cheap processing power have been the key factors in driving recent progress in deep learning.

The increasing diversity and quantity of available data together with cheap processing power provided by cloud services now make it plausible for adversaries to augment their cyberattacks with machine learning techniques.

Considering all available machine learning capabilities, the following tasks are most likely to enhance cyber attacker capabilities.

- **Prediction** is the task of forecasting the likelihood of a particular outcome based on previously observed data. Classification, anomaly detection, and regression are common examples of prediction tasks. Prediction can be used for offensive purposes, including the identification of keystrokes on a smartphone based on motion, the selection of the weakest target to attack, or the identification of software vulnerabilities to be exploited.
- **Generation** is the task of creating content that fits a target distribution. Generation can be used for offensive purposes, including tampering with media evidence, guessing passwords, or shaping network traffic to avoid detection. Another instance of offensive AI generation is deepfakes, which are believable video media created by deep learning models. This technology can be used to impersonate a target by mimicking their voice, face, and body language to perpetrate phishing or disinformation attacks.
- **Data analysis** is the task of mining or extracting useful insights from data, without knowing a priori what it is being looked for. Data analysis can be used for offensive purposes to identify how to better hide artifacts in malware or to identify assets or targets for social engineering within an organisation.
- **Information retrieval** is the task of finding content that matches or is semantically similar to a given query. Information retrieval methods can be used to track an object or an individual in a compromised surveillance system, to find a disgruntled employee (as a potential insider) via semantic analysis of social media posts, or to summarise lengthy documents during open-source intelligence (OSINT) gathering in the reconnaissance phase of an attack.

Changes caused by AI in cyberattacks

Intelligent automation provided by AI systems will improve conventional cyberattacks by increasing their speed, scale, coverage and sophistication, leading to an overall increase in their success. These improvements affect virtually every attacker tactic in the cyber kill chain. Many new attack techniques will be enabled by the emergence of new AI capabilities.

Benefits and improvements

Conventional attackers use manual effort, expert knowledge and basic attack tools to launch cyberattacks. Disruptions from AI-enabled cyberattacks are three-fold. First, AI techniques can be used to better automate manual attack tasks. Second, AI enhances basic attackers' tools. Third, AI brings completely new capabilities that attackers did not have before. By using AI, an attacker can increase the effectiveness and success of their attacks using the benefits that AI techniques provide. These can be summarised as follows:

- **Speed**: AI can be used to automate tasks that are currently performed manually, such as extraction of credentials, discovery of vulnerabilities in software, and password guessing. These tasks can now be run on a machine and performed at a much faster rate, allowing attackers to reach their goals in much less time. This, in turn, also decreases the time an attacker must spend in the victim's system, thus lowering their chances of being detected.
- **Scale**: AI can be used to scale up an attacker's operation by parallelising and launching automated attacks against many targets simultaneously. The value of AI is most prominent for target-specific attacks, such as spear phishing attacks, which can be personalised at scale for many victims. AI can enable an adversary to target more victims with higher precision and much lower manual effort.
- **Coverage**: AI makes attacks more pervasive and allows for more comprehensive attack coverage. AI-enabled cyberattacks can analyse and reason upon larger amounts of OSINT data, can explore more attack vectors, and can reach more assets to gain a stronger foothold. While conventional attackers can miss information and potential means for attacks, AI optimises the search for and exploitation of vulnerabilities to leave no stone unturned.

Disruptions from AI-enabled cyberattacks are three-fold. First, AI techniques can be used to better automate manual attack tasks. Second, AI enhances basic attackers' tools. Third, AI brings completely new capabilities that attackers did not have before.

- **Sophistication**: AI enables intelligent automation. This increases the sophistication of cyberattacks, making them better and more successful than those skilled human attackers could achieve alone. This increased sophistication comes in three forms:
  - **Contextualisation**: Generative AI capabilities enable an attacker to learn information about a victim or target system and to reuse this information while generating content. Rather than having a one-size-fits-all attack or manually contextualised attacks, AI-enabled attacks can be automatically tailored to their targets. Spear phishing emails can be personalised to their victims, malicious communications can be adapted to the network they occur in, and malware behaviour can be tailored to blend into the operation of the system they compromise.
  - **Adaptiveness**: With the ability to learn and relearn the target's environment automatically, attacks can autonomously adapt, in real time, to observed changes in the system, the target or the victim of the attack. Contextualisation is a long-lasting ability rather than a one-time feature.
  - **Evasiveness**: AI-enabled attacks are stealthier and more difficult to detect than traditional attacks. AI enables clever, optimised and stealthy reconnaissance through data analysis and information retrieval. It can be used to automatically discover attack vectors and places for compromise through prediction, thus reducing the number of required interactions with the target system. Through generation, AI mechanisms can learn and mimic the behaviour of systems and networks they compromise. Finally, by bringing autonomy to malicious programs, AI can reduce the need for communication with a command-and-control entity, thus improving stealth.

These benefits increase the overall success of AI-enabled attacks compared to conventional cyberattacks. AI-enabled attacks can be run faster, target more victims and find more attack vectors than conventional attacks because of the nature of intelligent automation and the fact that they replace typically manual tasks. AI-enabled attacks are also more sophisticated – they are more personalised to their target and can adapt in real-time, thus becoming harder to detect.

AI-enabled attacks can be run faster, target more victims and find more attack vectors than conventional attacks because of the nature of intelligent automation and the fact that they replace typically manual tasks.

AI-enabled attacks are also more sophisticated – they are more personalised to their target and can adapt in real-time, thus becoming harder to detect.

Offensive AI capabilities

One way to understand how AI can improve attacker tactics is by categorising the offensive use of AI into capabilities and then identifying how they improve the cyber kill chain. The MITRE ATT&CK framework breaks down stages of a cyberattack and the attacker's goals into 14 separate tactics. These encompass all techniques used during cyberattacks. Example tactics include reconnaissance, initial access, persistence, defence evasion, credential access, and lateral movement.

AI can support many of these tactics and provide new techniques to better achieve the attackers' goals through six main classes of offensive capabilities[3]:

- **Automation** enables speed, scale, coverage gain, and adaptiveness. Automation reduces the manual effort required by an adversary and increases the autonomy of cyberattacks. Automation mostly benefits the reconnaissance, initial access, lateral movements and impact phases of the ATT&CK framework. It provides techniques such as attack adaption to unknown and evolving environments. It also enables attack coordination to find the most vulnerable target, the best vector to exploit, and the best time to attack. AI can also be used to control the collaboration of bots in botnets through swarm intelligence. Automation also enables better attack campaigns such as more scalable and sophisticated phishing campaigns.
- **Stealth** is a key requirement for attack success. Stealth is achieved by an AI's ability to generate content that resembles a distribution it learned from. AI can therefore cloak malicious behaviour, making it resemble benign behaviour. Cyberattacks utilise multiple steps, all of which must avoid detection to complete a successful attack. The AI stealth capability benefits many tactics, including reconnaissance, initial access, persistence, lateral movement, collection, and exfiltration. It provides techniques for evasion of detection to defeat systems such as Host and Network Intrusion Detection Systems (HIDS and NIDS), email filters and malware detectors. Stealth provides techniques to hide scans and propagation into adjacent systems, and data exfiltration techniques that can blend into normal network activity.
- **Campaign resilience** ensures that attackers maintain a foothold in the systems they have already compromised, can compromise new systems, and are able to launch the next steps of their operation. Campaign resilience capabilities mostly benefit persistence and defence evasion tactics. AI can provide techniques for campaign planning through cost-benefit analysis, the automated identification of tools and resources required for an attack against a chosen target, and the simulation of an attack environment to test-run a planned campaign. It also enables malware obfuscation techniques, helping the attacker select the best place to hide a malicious piece of code in an existing piece of software, or a backdoor in a system. AI also helps in the identification of virtualisation environments, enabling the attacker to disable the execution of attack code within them, thus making both detection and reverse engineering efforts more difficult.

[3]Mirsky, Yisroel, et al. "The threat of offensive AI to organisations." arXiv preprint arXiv:2106.15764 (2021).

- **Social engineering** targets the exploitation of human users, which are often considered the weakest link in an information system. AI systems can learn from humans to better exploit their emotions and trust. This ability has been demonstrated in chatbots and personal voice assistants that can mimic human-like interactions, and in recommender systems designed to target individuals with advertisements, products or media recommendations. AI enhances social engineering attacks in the same manner – by learning from its victims. This AI capability benefits attacker tactics where human victims are involved, such as reconnaissance, initial access, privilege escalation and credential access. It provides techniques for target selection and tracking to select a victim from an organisation and to follow their activities before launching an attack against them. AI provides techniques for automated and personalised interactions with humans, both offline, via automatically generated spear phishing emails, and online, via chatbots. Finally, AI can be used to impersonate existing people using deepfakes and for building fake online personas to establish contact with targeted victims.
- **Credential theft/spoofing** enables illegitimate access to systems that are otherwise secured by compromising their authentication methods. AI can mimic human behaviour by reproducing authentication protocols and guessing credentials. Credential theft/spoofing capabilities are used for both initial access and credential access tactics. AI also provides techniques for impersonating the voice and face of a user to spoof biometric authentication systems. It provides techniques to defeat implicit key logging systems that rely on user actions, such as keystroke patterns, eye movements and device motion for authentication, by learning and mimicking those human behaviours. Generative AI models can guess passwords with low entropy, or those that include personal information about their targets.
- **Information gathering** enables an adversary to take sensible actions during an attack while ensuring the success of those actions by limiting the number of attempts or queries the attacker must make. AI can autonomously mine large amounts of data and extract relevant information from it. This capability is well suited to the collection of relevant information for an attack. This information gathering capability is beneficial to reconnaissance, credential access, collection and impact tactics. AI provides techniques suitable for the collection and mining of Open-Source Intelligence (OSINT) data. Stealth techniques can be used to camouflage both data collection and the probing of targeted systems. Incomplete intelligence information can be complemented using generative machine learning models that are able to fill gaps in missing data. Sensible information can be identified and extracted using natural language processing or graph mining techniques, which can also be used to identify the most valuable data to exfiltrate in a compromised system. Information gathering capabilities also provide techniques for espionage and target tracking through deep learning methods for large-scale image and voice processing.

Offensive AI capability	Cyber kill chain tactics	AI attack techniques
Automation	• Reconnaissance • Initial access • Lateral movement • Impact	• Attack adaptation • Attack coordination • Attack campaigns • Vulnerability discovery
Stealth	• Reconnaissance • Initial access • Persistence • Lateral movement • Collection • Command & Control • Exfiltration	• Evasion of detection • Scanning • Propagation • Data exfiltration
Campaign resilience	• Persistence • Defence evasion	• Campaign planning • Malware obfuscation • Identification of virtualisation
Social engineering	• Reconnaissance • Initial access • Privilege escalation • Credential access	• Target selection • Target tracking • Spear phishing • Impersonation • Fake persona building
Credential theft / spoofing	• Initial access • Credential access	• Biometric spoofing • Implicit key logging • Password guessing
Information gathering	• Reconnaissance • Credential access • Collection • Impact	• OSINT mining • Target selection • Target tracking • Espionage

The six AI offensive capabilities described in the above table largely benefit reconnaissance and defence evasion steps of the attack kill chain, followed by resource development, impact, discovery and collection. Current AI techniques do not significantly improve privilege escalation, execution, or persistence steps. Overall, AI offensive capabilities largely benefit the early and late stages of a cyber kill chain.

Examples of AI attack techniques

Concrete examples of AI attack techniques exist and have been used to demonstrate how AI can increase the success of cyberattacks. Current AI technologies are mature enough to be used for both stealth and social engineering applications. AI-based spear phishing and impersonation attack tools have already been developed. Some available AI techniques can already be combined to enhance several stages of the end-to-end cyberattack chain.

Spear phishing with target selection

AI can be used to support the selection of phishing victims by identifying and targeting specific characteristics via user profiling. To profile individuals, the adversary first collects online profiles from social media platforms (Twitter, Facebook, LinkedIn, etc.). During reconnaissance, this collection process may be limited to specific employees within a given organisation. A list of high-value employees can often easily be collected from LinkedIn. Sensible features that can be used to classify users' interests (number of followers, friends, contacts, the age of the account, the number of posts, the number of likes, retweets, reactions, their interests, hobbies, etc.) are then extracted from the collected profiles. These features are subsequently used to cluster potential victims into groups with similar characteristics – in much the same way as how user profiling is performed in recommender systems. The final step consists of identifying and labelling clusters of interest such as "highly gullible" or "high value." These clusters then go on to become the targets of subsequent spear phishing attacks.

During the second stage of the attack, online accounts of the selected victims are scraped and natural language processing techniques are used to extract "topics" representing the interests of each victim. These topics are then fed into a pre-trained natural language generation (NLG) model. Many such high-performance NLG models, such as GPT-3, are freely available online. The pretrained model may also be better contextualised by fine-tuning it against content the victim has written. Finally, the model is used to generate personalised emails or social media posts that mimic the victim's interests and writing style, thereby increasing the attack's chance of success.

An automated tool able to generate realistic phishing tweets, called SNAP_R[4], was developed by researchers at ZeroFox several years ago (prior to the creation of highly performant NLG models such as GPT-3). A real-life experiment concluded that SNAP_R could write tweets that were more successful at triggering victim click-through than human-written tweets. SNAP_R was able to generate new tweets four times faster than humans could write them. Incremental improvements to phishing content generation can be further achieved through conventional machine learning improvement techniques like A/B tests, where different versions of a phishing generator are used to send emails or social media messages to different sets of victims. The version with the highest response rate or click-through rate would thus be selected as the new generator and the baseline for further improvements.

[4]https://www.forbes.com/sites/thomasbrewster/2016/07/25/artificial-intelligence-phishing-twitter-bots/

Spear-phishing with target selection is an example of clever automation, where basic automated phishing message generators are redefined and augmented with the ability to select victims, capture their interest and write a message that relates to these interests in a fully automated manner.

Impersonation

Impersonation is a mechanism used in phishing and vishing (voice phishing) attacks. A technology called deep voice leverages deep learning techniques to impersonate a target's voice and can synthesise speech from text. Training a deep voice model requires audio samples of the victim's voice. This data can be obtained from recorded online meetings and audio of public appearances which are largely available online, especially if the victim is sufficiently high-profile (for instance, a politician or CEO). Several successful vishing attacks have already been publicly reported. In July 2019, the CEO of a UK-based energy company was impersonated in a vishing call that led to a fraudulent money transfer of $243,000[5]. In 2020, a Hong Kong bank director was impersonated using deep voice to order fraudulent money transfers amounting to $35 million, part of which were executed by a bank manager deceived by the vishing call.

Although occurrences of AI-enabled phishing have been anecdotal so far, it is possible to scale vishing campaigns via simple automation. Deep voice techniques can be combined with other deep learning technologies such as speech recognition and chatbots. Using this combination of techniques, vishing calls can be performed in a fully autonomous manner. The resulting program would parse the victim's speech (with a speech recognition algorithm), generate a text reply (using a natural language generation algorithm) and then be converted into the impersonated voice (using a deep voice technique). Deep voice techniques can also be used to fool biometric voice authorisation systems that are commonly used as authentication during phone calls.

Impersonation can be taken to another level using deepfakes, which enable an adversary to impersonate both the voice and face of a target[6]. Deepfakes can be used to impersonate a target during a video call by cloning the voice of the target, syncing their lips to the speech, and performing gestures that the victim might make. Deepfakes can be generated in real-time, converting a video of an actor into a video of the target. Deepfakes leverage generative deep learning models such as Generative Adversarial Networks (GANs). They require several stills of the target's face for training purposes (which can be extracted from frames in a video). Once again, such data can be easily obtained for high-profile targets since videos of their participation in public events are likely available online. Deepfake generation services and software are readily available and easy to use.

Deepfake-based impersonation is an example of new capability brought by AI for social engineering attacks. No prior technology enabled to convincingly mimic the voice, gestures and image of a target human in a manner that would deceive victims.

[5]https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/
[6]https://www.bleepingcomputer.com/news/security/elon-musk-deep-fakes-promote-new-bitvex-cryptocurrency-scam/

Malware communication cloaking

Defenders sometimes look for signs of malicious communication when searching for systems on a network that may have been compromised. Thus, by cloaking or suppressing communication, an attacker can better maintain stealth in the system they've hijacked. AI techniques can be used to blend malicious communications with benign network traffic. As an example, let us consider a malicious program designed to exfiltrate confidential data from a compromised system via a Command-and-Control (C&C) channel. This hypothetical implant is designed to use AI techniques to perform data exfiltration in the stealthiest manner possible. Here's how it might work. After initially being installed into the target system, the implant lays dormant and only performs passive monitoring of network traffic generated by the compromised system, possibly over an extended period of time. The implant clusters the collected traffic according to several characteristics such as ports and protocols used, domain names requested, common destinations, and traffic load during certain time periods. The results of this clustering are then used to identify the most common communication protocols, the most requested domain names, and peak traffic hours. The implant then sends the results of this local analysis to a "default" C&C server during the identified peak hour. The C&C server uses this information to register a domain that looks similar to the victim host's most requested domain name, and which will be used for future communications with the compromised system. It sets up a service using the same port and protocol most used by the compromised system. Once this is done, the malware activates and uses this new communication channel (domain name, service/port, peak hour) for further communication with the C&C. This channel is finally used to conduct stealthy data exfiltration, calibrating the quantity of exported data to match the typical traffic load of the compromised system at the given time.

AI-based malware communication cloaking is an example of enhancement of the basic attacker's tools for stealth. Data mining techniques are used to identify and model the common communication of a victim host. This information is further used to shape the already defined communications that the malware must have to achieve its goals.

An end-to-end AI-enabled cyberattack

Individual AI-based attack techniques can be combined to enhance several stages of an end-to-end cyberattack, including reconnaissance, intrusion, C&C establishment, privilege escalation, lateral movement, and exfiltration.

Automated deep learning CAPTCHA breakers (like GSA Captcha breaker[7]) can be used during reconnaissance (stage 1) to solve challenges while crawling and gathering information from public-facing web pages of a target organisation. Automated chatbots can also be employed to establish initial contact with employees of the target organisation. Further communications with the most responsive victims can then be taken over by human operators to gain specific information and develop trust. The gathered information can then be used during intrusion (stage 2) to craft believable spear-phishing messages against the identified victims using tools similar in nature to SNAP_R. The perimeter of the victim organisation's network can also be regularly scanned and fuzzed to find vulnerabilities using automated vulnerability fuzzing engines such as Mechanical Phish from ShellPhish[8].

Data exfiltration stages	AI attack techniques
Reconnaissance	• Captcha breaker • Automated chatbot
Intrusion	• SNAP_R spear phishing • Mechanical phish vulnerability fuzzing
Command & Control	• Communication clustering • Empire network traffic shaping
Privilege escalation	• CeWL password generator
Lateral movement	• Automated AI planning • Automated execution with MITRE CALDERA
Exfiltration	• Automated recognition of valuable content • Communication clustering • Empire network traffic shaping

[7]GSA Captcha breaker - https://www.gsa-online.de/product/captcha_breaker/
[8] The Mechanical Phish - https://github.com/mechaphish/mecha-docs

To establish a stealthy C&C channel (stage 3), a communications cloaking technique can be used for identifying peak traffic hours, common protocols and domain names used by the compromised host. This learned information can be fed to e.g. the Empire post-exploitation framework[9], to shape stealthy malicious communications that follow patterns learned from the victim's system. Privilege escalation (stage 4) can further be achieved by cracking the admin password of compromised hosts using clever password generators such as CeWL[10]. CeWL generates complex passwords composed of several words extracted from web content it analyses. If fed with social media accounts of a victim or system administrator, CeWL can generate complex passwords combining words related to the victim's interests.

Lateral movement (stage 5) can be planned to infer the optimal path from the currently compromised host to the ultimate target destination using automated AI planning methods. Lateral movement can be further automated using tools such as MITRE CALDERA[11]. This leads to selection and exploitation of the shortest path to complete the attack, reducing the attack duration and the likelihood for detection. To pre-select and reduce the amount of data sent during exfiltration (stage 6), a deep learning model for content recognition can be used to narrow down available content. Machine learning models designed to perform this task, such as "NSFW," already exis[12]. Such models can easily be fine-tuned and repurposed to identify "valuable" information specific to the target. The identified data can then be exfiltrated using the formerly established stealthy C&C channel and blended in with regular business operation communications.

[9]Empire - https://github.com/EmpireProject/Empire
[10]CeWL: Custom Word List generator - https://github.com/digininja/CeWL
[11]MITRE CALDERA - https://github.com/mitre/caldera
[12]Open nsfw model - https://github.com/yahoo/open_nsfw

The current threat of AI-enabled attacks

Attacker profiles come in many shapes and forms, and thus understanding how they might be enabled by AI is a challenging task. A group's technological readiness, availability of AI technical skills and interest in using AI are all relevant factors. Analysing different types of attackers and their motivations can help us project how and when AI will start being used in cyberattacks. Currently, most examples and knowledge related to AI-enabled cyberattacks come from public and private research whose aims are to understand the AI threat and increase our level of readiness when it materialises.

AI for which attacker and which purpose?

AI can be used in cyberattacks in two different ways. The first is defined as direct use, in which it provides new automation and enhances existing attacker tools. Such AI applications are mostly employed in early attack stages such as reconnaissance, discovery, collection, initial access, credential access and defence evasion. These applications currently receive the most research attention, and many examples are already available for direct use. Examples of such tools are CAPTCHA breakers (e.g. from GSA), password guessers (e.g. Cewl), vulnerability finders (e.g. Mechanical Phish from ShellPhish), phishing generators (e.g. SNAP_R), and deepfake generators (e.g. DeepFaceLab).

The second is defined as embedding AI capabilities in malware to enable autonomy and more complex decision-making. AI decision logic could theoretically enable malware to run multiple attack steps, find vulnerabilities and exploit them, all without human intervention. Such AI applications would be used during the latter stages of an attack chain for purposes of maintaining persistence, escalating privileges, lateral movement, command and control functionality, exfiltration, and impact. There are only a few practical examples of these applications, and many conjectures are being made about how AI might be used to create autonomous malware in the future.

We also identify three types of attackers that would exploit AI for different reasons.

- **Individual attackers** would leverage AI through direct use, with their main objectives being to scale and speed up their operations. Individual attackers would automate manual tasks and enhance attacks using mostly readily available technology that requires little effort to develop or adapt. This type of attacker is unlikely to develop their own AI-based attack tools.
- **Organised cybercrime groups** would leverage AI to optimise their business and maximise their profits. They would leverage AI solutions through direct use for most of the early stages of an attack, aiming to identify valuable targets and optimise their monetisation process. Organised cybercrime groups would use AI to perform business studies and identify activity sectors or companies from which they can obtain the most revenue. They would automate information gathering and OSINT campaigns to select the best attack vectors and quickest ways to attack their targets. This type of attacker has the means to both apply existing technologies and adapt them if necessary.

- **Nation-state attackers** would make use of more advanced AI techniques, utilising both direct and embedded mechanisms. They already likely use AI for large-scale data mining and information extraction in reconnaissance campaigns. One important driver for creating autonomous malware is to reduce the need for communications with a command-and-control server. Autonomous malware would be inherently stealthier and more resilient to detection. Another motivation is to prevent investigators' ability to link the malware to its operations, thereby hindering attribution efforts. Nation-state attackers are well-resourced and can easily hire AI experts to research and develop their own AI-enabled attacks, including autonomous malware.

Who currently perform AI-enabled attacks?

Currently, there is little evidence of AI-enabled attacks in the wild. This does not mean that attackers are not using or developing AI techniques to enhance their attacks. It is very likely that nation-state attackers already leverage AI in the early stages of attacks for reconnaissance, OSINT mining purposes and target identification. They likely have the means and skills to design AI-enabled cyberattacks and are probably investing in research related to autonomous malware concepts. It is possible that some better-resourced non-nation-state groups are developing methods to run advanced phishing campaigns.

However, it is difficult to find evidence for such attacks – investigators rarely gain access to attackers' backend systems where their AI-based logic is likely deployed. AI techniques serve only to improve existing attack techniques, leaving little trace that can help differentiate AI-enabled cyberattacks from their conventional counterparts. This explains why the only reports of real AI-enabled attacks are related to impersonation and deepfake/deep voice vishing. There is certainty about the nature of these attacks since they utilise mechanisms that were only recently enabled by cutting-edge AI techniques. However, there have been suspicions that AI played a part in other real-world attacks[13] such as the DDoS attack targeting TaskRabbit in 2019, which was suspected to have been launched by a botnet controlled using some AI capability. It has also been speculated that Instagram was targeted by AI-driven large-scale vulnerability scanning when a bug was exploited and led to a data breach in 2019. However, these are just conjectures without serious evidence to support them.

Most of our currently available knowledge regarding AI-enabled cyberattacks is derived from work performed by security researchers in industry and academia. Most known offensive techniques and tools that leverage AI were developed by research groups aiming to better understand threats and develop counters to them. A few research groups are dedicated to investigating AI-based attacks, like the Offensive AI Lab[14] founded in 2020 at Ben Gurion University in Israel and the Shellphish group[15] from UC Santa Barbara.

Public agencies have also started to consider this threat and have allocated funding for research on offensive AI capabilities. For instance, the DARPA Cyber Grand Challenge[16] was a competition to develop automated, scalable and machine-speed vulnerability detection solutions. The aim of this challenge was focused on the defence side of the problem: developing new defensive tools to find vulnerabilities and patch them before they are exploited, but it clearly called for offensive AI solutions. The European Commission recently launched a research call for "increased cybersecurity"[17], aiming for outputs that increase knowledge about how an attacker might use AI technology to attack IT systems and digital processes as well as products and systems resilient to AI-powered cyberattacks. While public funding exists, its level is still limited for this offensive technology since the threat may not yet be considered prominent enough due to limited evidence of the occurrence of real AI-enabled attacks.

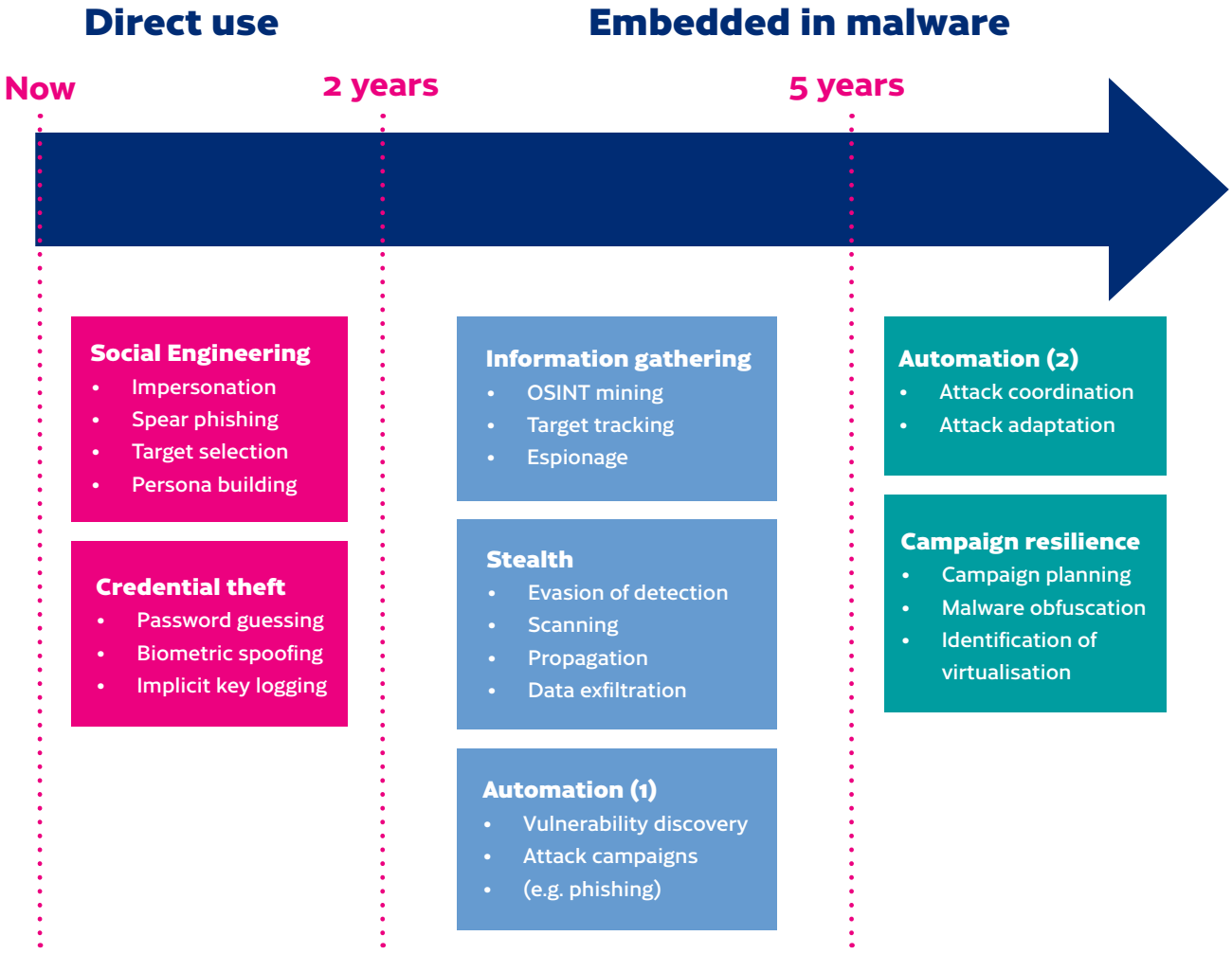
[13]Has an AI cyberattack happened yet? https://www.infoq.com/articles/ai-cyber-attacks/
[14]Offensive AI Lab - https://offensive-ai-lab.github.io/about/
[15]ShellPhish group - https://shellphish.net/
[16]DARPA Cyber Grand Challenge (CGC) - https://www.darpa.mil/program/cyber-grand-challenge
[17]https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/horizon-cl3-2021-cs-01-03

Timeline for AI-enabled attacks threat appearance

While AI can be used to enhance cyberattacks in several ways, only a handful of techniques have been properly demonstrated. Very few reports of real AI-enabled cyberattacks exist. Factors such as technological readiness, data availability, level of AI know-how, and motivation to use AI determine how quickly AI offensive capabilities will be taken up by cyberattackers. Considering the cyber kill chain, in the short-term AI is most likely to benefit the early stages of an attack, where AI algorithms are developed and run offsite. The following timeline highlights a forecast for the emergence of offensive capabilities in AI-enabled cyberattacks. Several blockers and enablers may impact these predictions in the long-term.

Short-term (0–2 years)

Although some current AI methods are mature enough to support direct use in cyberattacks, AI cannot yet be embedded in malware for the purposes of autonomy. The development of AI-based systems is complex – trial-and-error, tuning, validation, and extensive testing are all required to obtain highly performant systems. These processes are performed by experienced machine learning practitioners and through



direct interactions with the models being trained. This is possible in direct use scenarios where models are trained and operated on the attacker's own systems, but not with models running inside a malware on compromised systems. Communications and interactions are restricted in such deployments and can trigger detection events endangering the attack and exposing the malware code to security researchers.

For the same reason, current AI techniques are mature enough to support the early steps of an attack kill chain, such as reconnaissance, initial access or credential access. These steps are typically performed manually by the attacker, and thus trial-and-error approaches can be used while developing these approaches. For later steps in the attack kill chain, such as persistence, lateral movement, exfiltration or impact, which are executed by the malware in a more autonomous fashion, machine learning tuning is complicated.

The early steps of an attack kill chain are also most suitable for leveraging AI because many of them target human users rather than machines. Current AI technology has reached a high level of readiness in human applications like user profiling for marketing purposes, ad targeting, and content and product recommendation. AI also performs well at mimicking human behaviours with speech recognition, language translation, and text and speech generation. Social engineering and credential thefts are offensive AI capabilities which rely on user profiling, artificial human interactions and human behaviour imitation. These capabilities benefit from the technological progress of AI in human-related tasks, where AI technologies developed for legitimate uses can be repurposed into malicious applications through techniques like transfer learning. Voice assistance technology can be repurposed for impersonation; text generation models and chatbots can be repurposed for spear phishing, fake persona building and password guessing; and user profiling and recommendation systems can be repurposed for target selection. A second enabler for these offensive AI capabilities is the availability of data to train models. The large number of detailed user profiles is available on social media sites, and these can be easily scraped and used to build fake personas and user profiling that is subsequently used for impersonation, spear phishing or password guessing. Human-generated content like video recordings, voice recordings, emails or social media posts are also widely available, and can be easily used to train spear phishing generators, fake personas or deep fake models. Moreover, since this same data is already used to train machine learning models used for legitimate purposes, it has often been sanitised and curated; providing high-quality data ready to be used for malicious purposes.

These reasons, coupled with the fact that attackers already use automation for social engineering and credential theft, explain why these offensive AI capabilities are likely to be used by cyberattackers in the short term (0-2 years). As previously discussed, tools already exist for spear phishing generation and target selection. Examples of impersonation attacks using deepfakes have already been documented. AI-enabled social engineering and impersonation are also currently the most significant security threats perceived by both industry and academia. They fear the use of these techniques for the purpose of stealing credentials, gaining initial access and establishing footholds in victim systems.

Mid-term (2–5 years)

In the mid-term, AI-driven information gathering and OSINT mining capabilities are likely to become a realistic threat. Unsupervised learning techniques have not yet undergone the same performance revolution as prediction and generation techniques. Unsupervised learning is challenging because it is often difficult to know what pieces of information contain relevant signals. Thus, there are technical challenges associated with efficiently mining OSINT information using data analysis, preventing its adoption in the short term. On the other hand, we may soon see more targeted information gathering methods which rely on information retrieval, some form of OSINT mining, and target tracking. AI technologies for object recognition and pattern matching in image, video, sound or text are already highly performant and could easily be used for such purposes. However, even though the knowledge for such applications is available, legitimate tools that can be easily repurposed for OSINT mining are missing, increasing the effort needed for such AI-enabled attack techniques.

Stealth and automation are related to direct use scenarios such as vulnerability discovery and attack campaign automation. These will also be likely used by attackers in the mid-term (2-5 years). Both offensive techniques target software running on computers. While AI technologies are mature enough to learn from, profile and replicate computer behaviours, there are challenges related to data availability, data quality and availability of machine learning models able to solve similar problems for use in stealth and automation. There are virtually no publicly available models designed to evade detection systems, propagate attacks or automate attack campaigns, and thus no legitimate models that could be repurposed for malicious applications. There are also comparatively few high-quality datasets available that represent systems behaviour, network traffic or vulnerability discovery. This issue is partly solved for AI embedded in malware, since such models could be deployed in compromised systems and would be able to monitor behaviour and network traffic at will. However, data extracted from such systems could not easily be sanitised or labelled, and would thus not be of high enough quality to be used in the training of high-performance machine learning models.

Nevertheless, information gathering and the automation of vulnerability discovery are perceived as a significant threat by the industry. Vulnerability discovery also has legitimate applications for bug fixing, and there is a notable effort to develop AI-based solutions for this purpose. This may lead to an increased availability of data usable for this purpose, as well as legitimate vulnerability discovery tools that could be repurposed for malicious usage. This trend may lead to AI-enabled vulnerability discovery used in cyberattacks sooner rather than later.

Long-term (> 5 years)

The last of our previously discussed offensive AI capabilities, namely campaign resilience and autonomous malware, will likely only appear in the long term. Both require the construction of algorithms with a very high level of autonomy – something that is impossible to achieve at the current moment. Reinforcement learning is a technique that could be used to create autonomous malware. However, current reinforcement learning techniques suffer from challenges such as the difficulty of defining and computing sensible rewards and the large number of episodes required for training. Trial-and-error approaches cannot be used on a real compromised system, since they would be easily detected or would likely

cause failures in the compromised system itself. Offensive AI techniques like attack adaptation, campaign planning or malware obfuscation use generation techniques. While generators perform well at synthesising natural data such as images, video or speech, where the content has low format constraints they struggle at generating valid content under strong formatting constraints, such as machine language, code or network packets. A last challenge related to the creation of malware with embedded AI logic is the lack of availability of machine learning libraries required for them to run on compromised systems. A wide enough deployment of machine learning libraries in computing systems such as laptops, smartphones or tablets has not yet been reached. Machine learning models and libraries would need to be included in the malware itself, drastically increasing the payload file size. Moreover, ML models enabling autonomy are akin to large language models that are very large and require large amounts of computing power and memory to run. The size and resources consumed by these models prevent their deployment in current systems, not meeting the desired requirements, and they will possibly enable easier detection in the future. Due to these challenges, it is unlikely that we will witness self-planned attack campaigns or intelligent self-propagating worms driven by AI any time soon.

Nation-state attackers will be (or already are) the first likely threat actor to use AI-enabled cyberattacks, because they are deliberate, calculated, well-funded and supported with enough resources to target anything or anyone they deem worthwhile.

Aside from legitimate researchers, nation-state attackers will be (or already are) the first likely threat actor to use AI-enabled cyberattacks, because they are deliberate, calculated, well-funded and supported with enough resources to target anything or anyone they deem worthwhile. After widespread nation-state adoption of AI-based cyber stack tools, the usage of AI in cyberattacks will likely trickle down to less skilled and resourced adversaries.

Threat blockers

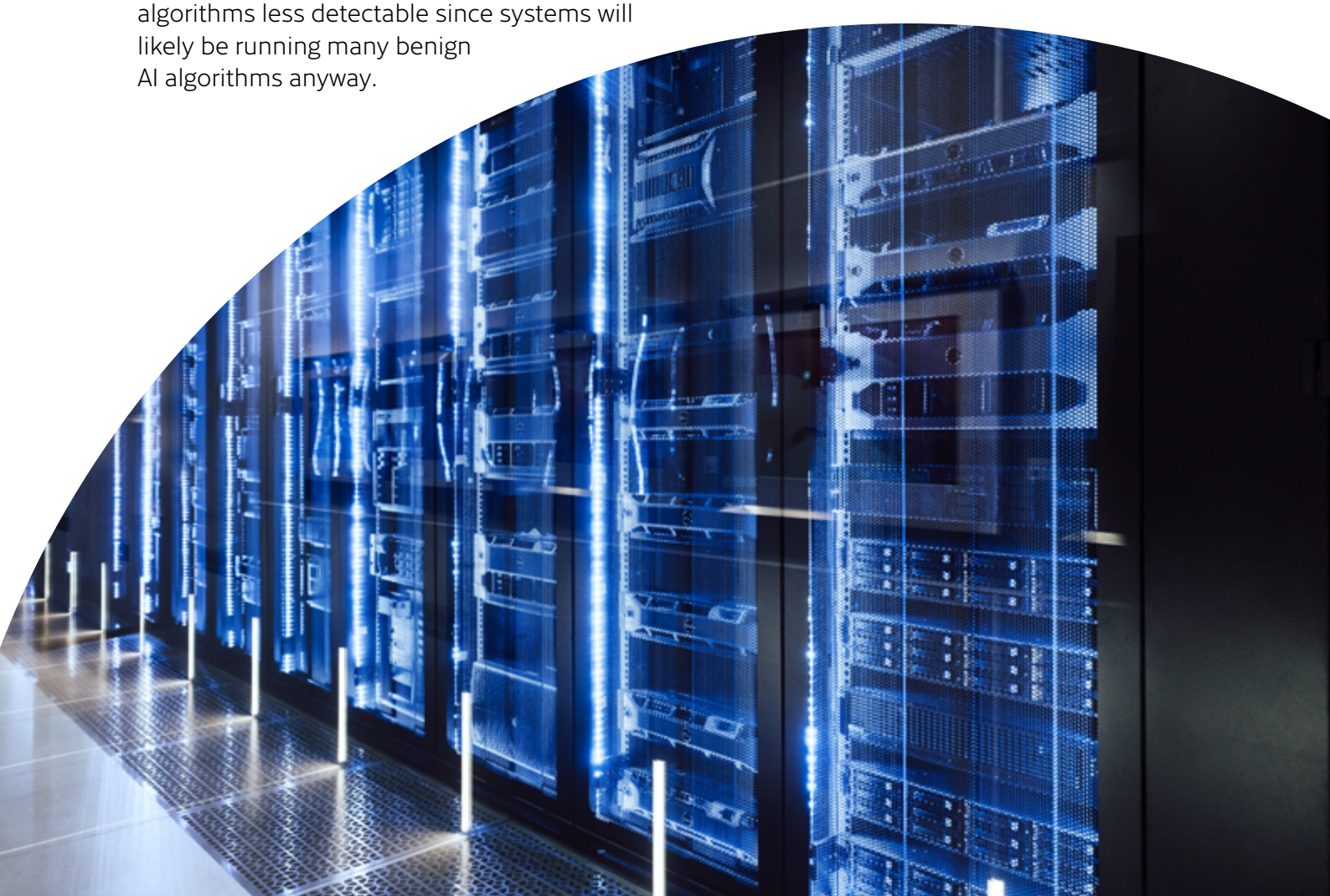
The factors blocking adoption of AI in cyberattacks are as follows:

- **Lack of attacker's motivation**: As long as conventional attack techniques remain effective and enable attackers to reach their goals (e.g. profit), they have little motivation to use AI.
- **Lack of AI skills**: Conventional cyberattackers are not well-versed in AI techniques. This is a barrier for both the identification of how AI might benefit cyberattacks and for the implementation of AI-enabled attacks.
- **Lack of available data**: Besides human-related data such as text, sound, video and images, there is a lack of available high-quality data related to systems and networks and how to attack them. This hinders the development of AI-enabled attacks against systems.
- **AI technical readiness**: AI technology is not yet performant enough in fields like unsupervised learning and reinforcement learning to enable the use of AI in some attack techniques. For instance, current AI techniques cannot yet provide the high degree of autonomy and self-tuning ability required for the creation of autonomous malware.

Threat enablers

Several factors could foster and accelerate the appearance of AI-enabled attacks in the future:

- **Conventional cyberattacks become obsolete**: If security solutions effectively defeat existing attack techniques and prevent attackers from making a sufficient profit, there will be an incentive to use AI to improve the effectiveness of cyberattacks.
- **Decreased cost/benefit of AI-enabled attacks**: Considering the 5G/6G and IoT revolution, more devices are being connected to the Internet, all of which are potential targets. AI-enabled attacks may be the only way to scale up operations and address this larger ecosystem.
- **Widespread availability of AI capabilities in end-devices**: As AI capabilities become more widely available on regular systems, driven by trends in edge computing and ubiquitous AI in 5G/6G, it will be possible to easily embed AI in malware and run it on any system. This will also make malicious AI algorithms less detectable since systems will likely be running many benign AI algorithms anyway.
- **Availability of open-source attack tools**: Research on AI-enabled attacks and benign machine learning applications that can be repurposed for malicious usage will produce tools that can be misused by attackers. Tools for vulnerability detection or penetration testing are such examples. They reduce the effort to run AI-enabled attacks and increase the attackers' incentive for using them.
- **AI skills available in dark web marketplaces**: As AI expertise becomes more widespread, average salaries in the field may decrease. This may incentivise AI experts to market their skills to malicious groups. One can imagine that AI skills and tools might be sold on dark web marketplaces as a Cyberattack-as-a-Service (CAaaS) model. Cyberattacker groups would then purchase AI tools and services to use in their cyberattacks. This model also removes some of the accountability from the AI experts since they would not be the ones ultimately using them for malicious purposes.



Impact on current approaches to cybersecurity

The use of AI will lead to more contextualised, faster, stealthier and more unpredictable cyberattacks. A slow or ineffective response to these attacks will allow adversaries to get deeper into systems and networks before being detected or blocked. Cybersecurity solutions will have to evolve more intelligent automation to address the emergence of AI-enabled cyberattacks. New security approaches will need to be developed to detect and counter AI-enabled cyberattacks, leading to a new arms race that exploits cutting-edge AI technologies for both attack and defence.

Changes to current cybersecurity approaches

As we have addressed, AI will inevitably be used to improve existing attack techniques by increasing their speed and scale. These same characteristics may also be leveraged to pursue secondary attack goals and create decoys. The speed and scale of AI-enabled cyberattacks may simply be used to exhaust and overwhelm existing, relatively slow defensive measures. By triggering many of the rules that are handled by human operators, a conventional cyberattack could be run in parallel to a decoy attack and may go completely unnoticed.

To deal with this evolution, security solutions won't need to be changed drastically – current approaches already have to keep up the increasing speed, scale and sophistication of cyberattacks, despite their lack of use of AI techniques. Security solutions continue to evolve and remain effective in the attacker-defender arms race. AI-enabled cyberattacks will hasten this trend and lead to an increased use of autonomous decision-making mechanisms in security solutions. These will offset the shortcomings of human-based detection and response operations that can take hours or even days. Only automated defence solutions running on endpoints in an organisation will be able to match the speed and scale required to counter AI-enabled cyberattacks. Furthermore, only clever decision-making provided by AI will be able to cope with the high adaptability and sophistication of AI-enabled cyberattackers. Conventional automation using rule-based or signature-based systems is too static and slow to adapt to fast-evolving threats. While the integration of AI in security solutions has been an ongoing trend for over 15 years and most organisations invest in it, they admit that they are not yet ready to cope with AI-enabled cyberattacks. AI-based security solutions need to become more widespread and ubiquitous to cope with this new security threat. Nevertheless, human security operators will have to be kept in the loop, as a practical and ethical requirement, and to control and determine high-level security strategies.

AI-based security solutions need to become more widespread and ubiquitous to cope with this new security threat.

AI will enable completely new attack techniques which will be more challenging to cope with, and which will require the creation of new security solutions. Current techniques used to detect deception and phishing only work against current threats. Completely new defence techniques will be required to detect and counter AI-based phishing techniques that utilise synthesised content. Although research is starting to address such attacks, there is no effective solution to counter them yet. There are also no solutions available to prevent side-channel credential theft attacks that can learn and reproduce human behaviours used in implicit key logging. Research is needed to develop novel defences against these and many other potentially new AI-enabled attack techniques.

Overall, many security processes will need to be modified and made a lot faster – not just those designed to detect and respond to cyberattacks. Passwords may need to be updated more frequently and identified vulnerabilities will need to be fixed in a timelier manner – AI-enabled attacks will be able to exploit known vulnerabilities with a much shorter lead time and on a potentially massive scale. Certain security processes will become deprecated when they are found to be insecure in the face of AI-enabled attacks. This will likely be the case for voice authentication methods over phone calls as well as many other biometric or behaviour-based authentication methods, which, while being convenient, can be easily spoofed by AI generation techniques. Finally, users will have to change their habits and be trained to cope with the new kinds of deception that AI enables. The notion of what can be used for authentication and establishing trust will have to be revisited. A familiar voice on the phone or a familiar face in a video chat will no longer be sufficient grounds to prove the identity of an individual and therefore should not be trusted anymore.

Solutions to cope with AI-enabled cyberattacks

Mitigating AI-enabled attacks will require organisations to first deploy technical solutions to detect them. This is a complicated task, especially when AI is used offsite by the attacker, and the only means to detect this fact is through the data artifacts input into machine learning models or output from generative models. Generative models typically leave a signature in the content they generate which can be identified using classification techniques. However, identification of this signature requires information about the specific model used by the attacker, which is often unknown. This information may be available in some cases, though. Natural language generation models used to create phishing content will likely be based on pre-trained weights from established projects, such as GPT-3. Large language models are expensive to train, and we'd not expect any group, aside from perhaps a nation state, to train their own model.

To ease the identification of machine learning-generated content, the data used to train attackers' models can be marked or tainted such that it will also taint the model and its resulting predictions or generated content. Public data often used by attackers, such as social media accounts, audio and video from potential target employees, can be watermarked to strengthen the signature left in the generated content. This watermarking approach eases the identification of machine learning-generated content and makes it independent from the type of model used by the attacker. Alternatively, public content likely to be used for attacks can be modified in a way that makes it unlearnable or unusable by any machine learning technique. Another approach for the detection of AI-enabled cyberattacks is to create fake data that is very

sought-after by attackers, in the fashion of a canary or a honeypot. For instance, fake user accounts for high-profile targets can be created on social media in the hope that a machine learning model for target identification will select them. Since these profiles are fake, they can be monitored, and any contact made with them could be used to identify and potentially track reconnaissance activities. The modification of content to prevent or track its usage in machine learning models will require collaboration from content publishers, such as social media platforms, and involve their responsibility to ensure that the content they publish is only used for legitimate purposes.

Once technical solutions for detecting AI-enabled cyberattacks are deployed, we will have the means to collect related threat intelligence. Reports of AI-enabled attacks could be gathered in a common repository and catalogued in a similar way to how information about vulnerabilities is stored in CVE databases. Such a knowledge base would be a valuable resource for organisations to assess their security posture with respect to AI-enabled attacks and for security experts to keep up to date with new threats.

Coping with AI-enabled attacks will require the use of more AI systems for defence. Similar technologies and advances in AI will be required to enable future attacks and to defend against them. Winning this new arms race will boil down to attackers and defenders vying to adopt new AI advances first. New security solutions will have to leverage AI advances before attackers do. Cybersecurity practitioners must continue to invest in AI expertise, which may prove challenging considering the current shortage in AI talent. An additional challenge comes from the asymmetric nature of the attacker-defender dilemma. Attackers will be free to use AI techniques in the manner of their choosing, whilst defenders will be bound by emerging regulations on the usage of AI, such as the European Commission's AI Ac[18]. In this scenario, it is possible that attackers will eventually gain more benefit from AI than defenders. On the other hand, AI regulation could enforce careful consideration on the repurposing of any newly developed AI solution with respect to its use for malicious purposes. This could slow down the emergence of AI-enabled cyberattacks in the future.

[18]EC Artificial Intelligence Act - https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

Finnish Transport and Communications Agency Traficom
PO Box 320, FI-00059 TRAFICOM
Switchboard: +358 29 534 5000

traficom.fi

ISBN 978-952-311-828-7
ISSN 2669–8757 (e-publication)

TRAFICOM
Finnish Transport and Communications Agency