TRAFICOM
Finnish Transport and Communications Agency

Huoltovarmuuskeskus
Försörjningsberedskapscentralen
National Emergency Supply Agency

Applying artificial intelligence in Cybersecurity

Traficom Research Reports
07/2024

TRAFICOM
Finnish Transport and Communications Agency

Huoltovarmuuskeskus
Försörjningsberedskapscentralen
National Emergency Supply Agency

Date of publication
13.3.2024

Title of publication
Applying artificial intelligence in cybersecurity

Author(s)
Samuel Marchal, Bartosz Nawrotek, WithSecure

Commissioned by
Finnish Transport and Communications Agency Traficom

Publication series and number Traficom Research Reports 07/2024	ISSN (e-publication) 2669-8781 ISBN (e-publication) 978-952-311-917-8

Keywords
Artificial Intelligence, Machine Learning, Cybersecurity, Large Language Models, Threat prevention, Threat detection, Risk management.

Abstract

The cybersecurity industry has utilized Artificial Intelligence (AI) for over two decades, applying it across various domains such as spam filtering, malware detection, and intrusion detection to enhance performance through automation, speed, scalability, and adaptability. While AI's impact has been predominantly in reactive cybersecurity measures, new AI technologies are promising for proactive security efforts, including advanced threat intelligence, security risk management, and heightened security awareness.

Nevertheless, the path to successfully using AI for cybersecurity is paved with many pitfalls, and successful applications have been developed at the cost of many failures. This success requires advanced knowledge, skills and experience in both AI and cybersecurity, which a few specialized organizations have been able to gather and reap the benefits from. In the context of the current AI hype, a widespread interest has risen into exploring the possible applications of AI, including those to cybersecurity. Many organizations seek to know if and how they could use AI to improve their security posture. This turns out to be challenging without field experience, considering the shortage of skilled AI and security experts.

This report aims to provide organizations with insights into AI's capabilities and potential benefits for cybersecurity, detailing existing applications, their maturity levels, and associated challenges. AI's effectiveness in improving threat detection and endpoint security is noted, though its applications in threat intelligence and vulnerability management remain nascent. The report emphasizes the importance of a meticulous development process and the integration of AI into security measures, underscoring the necessity of aligning AI solutions with business objectives, thorough understanding of data, and testing early prototypes in real-world settings. Developing cross-competency among experts in AI and cybersecurity is crucial for success.

Looking ahead, emerging AI technologies like Large Language Models (LLMs) are set to revolutionize cybersecurity applications by supporting security education, analytics, threat intelligence, and vulnerability management. These models promise to enhance the processing and correlation of information from vast amounts of unstructured data, potentially leading to more autonomous and complex task management. Nonetheless, as AI applications evolve, they will encounter new ethical, technical, and regulatory challenges that could impede progress.

The report was conducted in collaboration with the National Emergency Supply Agency.

Contact person Aleksi Blomqvist, Markus Mettälä	**Language** English	**Confidence status** Public	**Pages, total** 39

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

TRAFICOM
Liikenne- ja viestintävirasto

Huoltovarmuuskeskus
Försörjningsberedskapscentralen
National Emergency Supply Agency

Julkaisupäivämäärä
13.3.2024

Julkaisun nimi
Tekoälypohjaiset kyberturvallisuusratkaisut

Tekijät
Samuel Marchal, Bartosz Nawrotek, WithSecure

Toimeksiantaja
Liikenne- ja viestintävirasto Traficom

Julkaisusarjan nimi ja numero
Traficomin tutkimuksia ja selvityksiä 07/2024

ISSN (verkkojulkaisu) 2669-8781
ISBN (verkkojulkaisu) 978-952-311-917-8

Asiasanat
Tekoäly, koneoppiminen, kyberturvallisuus, tietoturva, suuret kielimallit, uhkien ennaltaehkäisy, uhkien tunnistaminen, riskienhallinta.

Tiivistelmä

Kyberturvallisuusalalla on jo yli kahden vuosikymmenen ajan hyödynnetty tekoälyteknologioita, mikä on merkittävästi tehostanut kyberuhkien torjuntaa. Tekoälyä on käytetty laajasti muun muassa roskapostin suodattamisessa, haittaohjelmien tunnistamisessa ja tunkeutumisen estämisessä, mikä on lisännyt kyberturvallisuusratkaisujen automaatiota, nopeutta, skaalautuvuutta ja sopeutumiskykyä. Vaikka suurin osa tekoälyn hyödyistä on nähty reaktiivisissa toimenpiteissä, on kiinnostusta herättänyt myös tekoälyn potentiaali myös ennaltaehkäisevissä toimissa, kuten uhkatiedustelussa ja turvallisuusriskien hallinnassa.

Tekoälyn soveltaminen kyberturvallisuusratkaisuissa on kuitenkin osoittautunut haastavaksi, ja monet onnistuneet sovellukset ovat syntyneet pitkällisten kokeilujen ja epäonnistumisten jälkeen. Onnistunut soveltaminen vaatii sekä tekoälyn että kyberturvallisuuden syvällistä ymmärrystä ja osaamista. Nykyinen tekoälybuumi on saanut monet organisaatiot pohtimaan, miten ne voisivat hyödyntää tekoälyä parantaakseen kyberturvallisuuttaan.

Selvitys tarjoaa kattavan katsauksen tekoälyn soveltamisesta kyberturvallisuuden parantamisessa ja paneutuu tekoälyn hyötyihin, haasteisiin ja tulevaisuuden kehitysmahdollisuuksiin. Tekoälyä hyödynnetään kyberturvallisuudessa laajaan sovellusten kirjoon, alkaen perinteisistä suodatus- ja tunnistusmekanismeista kehittyneempiin ennaltaehkäiseviin toimenpiteisiin. Vaikka tekoäly on edistynyt uhkien havaitsemisessa ja päätelaitteiden suojauksessa, sen soveltaminen esimerkiksi uhkatiedusteluun ja haavoittuvuuksien hallintaan on vielä kehitysvaiheessa. Selvityksessä muodostetaan kuvaa tulevaisuuden kehityskuluista ja potentiaalisista käyttötapauksista.

Selvityksen tulokset osoittavat, että tekoälyn onnistunut hyödyntäminen kyberturvallisuudessa vaatii monialaista osaamista, ja että organisaatioiden tulisi arvioida huolellisesti tekoälyyn pohjautuvien ratkaisujen soveltumista kriittisiin käyttötapauksiin. Lisäksi selvitys painottaa datan ymmärtämisen, saatavuuden ja laadun merkitystä, sekä prototyyppien testaamista todellisissa ympäristöissä ennen laajamittaisen käyttöönoton harkitsemista.

Uusien tekoälyteknologioiden, kuten suurten kielimallien, yleistyminen tarjoaa merkittäviä mahdollisuuksia kyberturvallisuussovellusten kehittämiselle. Näitä teknologioita voidaan hyödyntää muun muassa turvallisuuskoulutuksessa, analytiikassa ja uhkatiedustelussa, mahdollistaen laajojen datamäärien tehokkaan käsittelyn ja asiayhteyksien tunnistamisen. Tekoäly tuo mukanaan myös haasteita, kuten eettisiä kysymyksiä, teknisiä riskejä ja sääntelyyn liittyviä seikkoja, jotka kaikki vaativat huolellista harkintaa ja strategista lähestymistä.

Selvitys on toteutettu yhteistyössä Huoltovarmuuskeskuksen kanssa.

Yhteyshenkilö	Raportin kieli	Luottamuksellisuus	Kokonaissivumäärä
Aleksi Blomqvist, Markus Mettälä	Englanti	Julkinen	39

Jakaja
Liikenne- ja viestintävirasto Traficom, Kyberturvallisuuskeskus

Kustantaja
Liikenne- ja viestintävirasto Traficom, Kyberturvallisuuskeskus

TRAFICOM
Transport- och kommunikationsverket

Huoltovarmuuskeskus
Försörjningsberedskapscentralen
National Emergency Supply Agency

Utgivningsdatum
13.3.2024

Publikation
Cybersäkerhetslösningar baserade på artificiell intelligens

Författare
Samuel Marchal, Bartosz Nawrotek, WithSecure

Tillsatt av och datum
Transport- och kommunikationsverket Traficom

Publikationsseriens namn och nummer	
Traficoms forskningsrapporter och utredningar 07/2024	ISSN (elektronisk publikation) 2669-8781 ISBN (elektronisk publikation) 978-952-311-917-8

Ämnesord
Artificiell intelligens, maskininlärning, cybersäkerhet, informationssäkerhet, stora språkmodeller, förebyggande av hot, identifiering av hot, riskhantering.

Sammandrag

Inom cybersäkerhetsområdet har man redan i över två årtionden använt sig av tekniker med artificiell intelligens, vilket i hög grad har effektiviserat bekämpningen av cyberhot. Artificiell intelligens har använts i stor utsträckning vid bland annat filtrering av skräppost, identifiering av skadliga program och förhindrande av intrång, vilket har ökat automatiseringen av samt snabbheten, skalbarheten och anpassningsförmågan hos cybersäkerhetslösningar. Även om största delen av fördelarna med artificiell intelligens har setts vid reaktiva åtgärder, har den artificiella intelligensens potential väckt intresse även vid förebyggande åtgärder, såsom underrättelser om hot och hantering av säkerhetsrisker.

Tillämpning av artificiell intelligens i cybersäkerhetslösningar har dock visat sig vara utmanande, och många framgångsrika applikationer har uppkommit efter långvariga försök och misslyckanden. En lyckad tillämpning kräver en djup förståelse av och kompetens i såväl artificiell intelligens som cybersäkerhet. Det nuvarande uppsvinget för artificiell intelligens har fått många organisationer att fundera över hur de skulle kunna utnyttja artificiell intelligens för att förbättra sin cybersäkerhet.

Utredningen ger en heltäckande översikt över tillämpningen av artificiell intelligens vid förbättring av cybersäkerheten och fördjupar sig i fördelarna med samt utmaningarna och de framtida utvecklingsmöjligheterna för artificiell intelligens. Artificiell intelligens utnyttjas inom cybersäkerhet för ett brett spektrum av applikationer, allt från traditionella filtrerings- och identifieringsmekanismer till mer avancerade förebyggande åtgärder. Även om den artificiella intelligensen har gjort framsteg i att upptäcka hot och skydda terminalutrustning, är tillämpningen av den för exempelvis underrättelse om hot och hantering av sårbarheter ännu under utveckling. I utredningen skapas en bild av utvecklingsgången och potentiella användningsfall i framtiden.

Resultaten av utredningen visar att ett framgångsrikt utnyttjande av artificiell intelligens inom cybersäkerhet kräver tvärvetenskapligt kunnande och att organisationerna noggrant borde överväga tillämpning av lösningar baserade på artificiell intelligens för kritiska användningsfall. Dessutom understryker utredningen vikten av att förstå data, tillgången till och kvaliteten på data samt testning av prototyper i verkliga miljöer innan ett storskaligt ibruktagande övervägs.

Det faktum att nya tekniker med artificiell intelligens, såsom stora språkmodeller, blir allt vanligare innebär stora möjligheter att utveckla cybersäkerhetsapplikationer. Dessa tekniker kan utnyttjas inom bland annat säkerhetsutbildning, analys och underrättelse om hot, vilket möjliggör effektiv behandling av stora datamängder och identifiering av kontexter. Artificiell intelligens innebär även utmaningar, såsom etiska frågor, tekniska risker och frågor om reglering, vilka alla kräver noggrant övervägande och ett strategiskt angreppssätt.

Utredningen har gjorts i samarbete med Försörjningsberedskapscentralen.

Kontaktperson	**Språk**	**Sekretessgrad**	**Sidoantal**
Aleksi Blomqvist, Markus Mettälä	Engelsk	Offentlig	39

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

Contents

Figures

Tables

List of abbreviations

AGI	Artificial General Intelligence
AI	Artificial Intelligence
C&C	Command-and-Control
CAPTCHA	Completely Automated Public Turing test to tell Computers and Humans Apart
CloudDR	Cloud Detection & Response
CVE	Common Vulnerabilities and Exposure
CWE	Common Weakness Enumeration
DL	Deep Learning
DNN	Deep Neural Networks
DoS	Denial-of-Service
EDR	Endpoint Detection & Response
GDPR	General Data Protection Regulation
GenAI	Generative Artificial Intelligence
ICS	Industrial Control Systems
IDS	Intrusion Detection System
IoT	Internet of Things
IPS	Intrusion Prevention System
LIME	Local Interpretable Model-agnostic Explanations
LLM	Large Language Model
ML	Machine Learning
NLP	Natural Language Processing
OSTI	Open-Source Threat Intelligence
PE	Portable Executable
PII	Personally Identifiable Information
RNN	Recurrent Neural Network
SHAP	SHapley Additive exPlanations
SIEM	Security Information and Event Management
SOC	Security Operation Centre
UEBA	User and Entity Behaviour Analytics

1 AI for cybersecurity

Artificial intelligence, machine learning, deep learning and generative AI are ambiguous overlapping fields that need to be demystified and understood. They are trendy and have benefited from numerous recent advances that have provided new capabilities and enabled new applications. Many of these capabilities are useful for cybersecurity applications They can automate and improve intelligent processes as well as support the work of cybersecurity experts.

1.1 What is AI?

Artificial Intelligence (AI) refers to the ability for 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, generalize, 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 many tasks generally connected to human intelligence, 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 shortcut term to describe machine learning.

Machine learning (ML) 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 in three prominent types, namely

1) *supervised learning* which is designed to perform or replicate known tasks (task-driven),

2) *unsupervised learning* which is designed to extract hidden information from data (data-driven),

3) *reinforcement learning* which is designed to learn new tasks via trial-and-error while trying to maximize a defined reward (trial and error-driven).

Deep Learning (DL) 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 current hype around AI and machine learning. Deep learning techniques achieve unmatched, often super-human performance in complex tasks such as image classification, text translation or playing complex games. Machine learning and deep learning 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, the availability of large data sets, and cheap processing power have been the key factors in driving recent progress in deep learning.

Generative AI (GenAI) refers to a type of machine learning algorithm that creates or generates new content, whether it is images, text, music, or even videos. It does not just retrieve information or make decisions based on predefined choices but instead produces entirely new data based on patterns it has learned from a given dataset.

Large Language Models (LLMs) fall under the joint umbrella of GenAI, Deep Learning and Natural Language Processing (NLP), with a focus on text generation. These models are trained on vast amounts of text data to understand language patterns and generate human-like text. They work by predicting the next word in a sequence based on the words that came before it. They can complete sentences, generate stories, answer questions, and even code based on the patterns they have learned. These models are the closest example we have developed so far to Artificial General Intelligence (AGI), a hypothetical form of AI that possesses the ability to understand, learn, adapt, and apply its intelligence across a wide range of tasks or domains much like a human being.

In this report, while we primarily use the term AI, we will consistently refer to its specific subfield, machine learning (ML), aligning with its prevalent usage in mainstream communication.

1.2 AI capabilities relevant to cybersecurity

AI and ML provide many capabilities that can be applied to a wide array of tasks across various domains. Depending on the specific problem and the availability of data, machine learning techniques can be tailored and applied to achieve accurate predictions, classifications, or insights. The following capabilities demonstrate the versatility of machine learning in solving problems in the cybersecurity industry.

Some ML techniques can be used in capabilities that automate cybersecurity processes. These capabilities concentrate on identifying malicious content, events, or behaviour, and distinguishing them from benign elements.

- **Classification** is the task of assigning an object (or input data) into predefined categories or classes. This class assignment is learned from previously observed data, and it can be applied to new unknown data, which can be classified. Classification is widely used in cybersecurity for malicious content detection for malware as well as phishing and spam email. For these types of applications, classification requires examples of both benign and malicious contents that one wants to differentiate.
- **Anomaly detection** is the task of identifying patterns or instances that deviate significantly from what is considered normal or expected within a system or a dataset. It operates by establishing a baseline of normal behaviour and then flagging any data points or events that fall outside this established norm. Anomaly detection is commonly used in cybersecurity to establish a baseline of "normal" behaviour within a network or system. Deviations from this baseline are then continuously monitored and any anomalies are flagged as indications of potential security threats or a system compromise.
- **Behavioural analysis** is a task close to anomaly detection. It involves studying and understanding typical patterns of behaviour within a system, network, or user activity. In contrast to anomaly detection, behavioural analysis does not focus solely on outliers or anomalies. It aims instead to comprehend usual or expected behaviour of entities within a system. In cybersecurity, behavioural

analysis often involves creating profiles of normal behaviour for users, devices, or systems based on historical data, user actions, and system activities. The identification of deviations or changes from the established profiles might indicate potential security risks such as insider threats or unauthorized accesses.

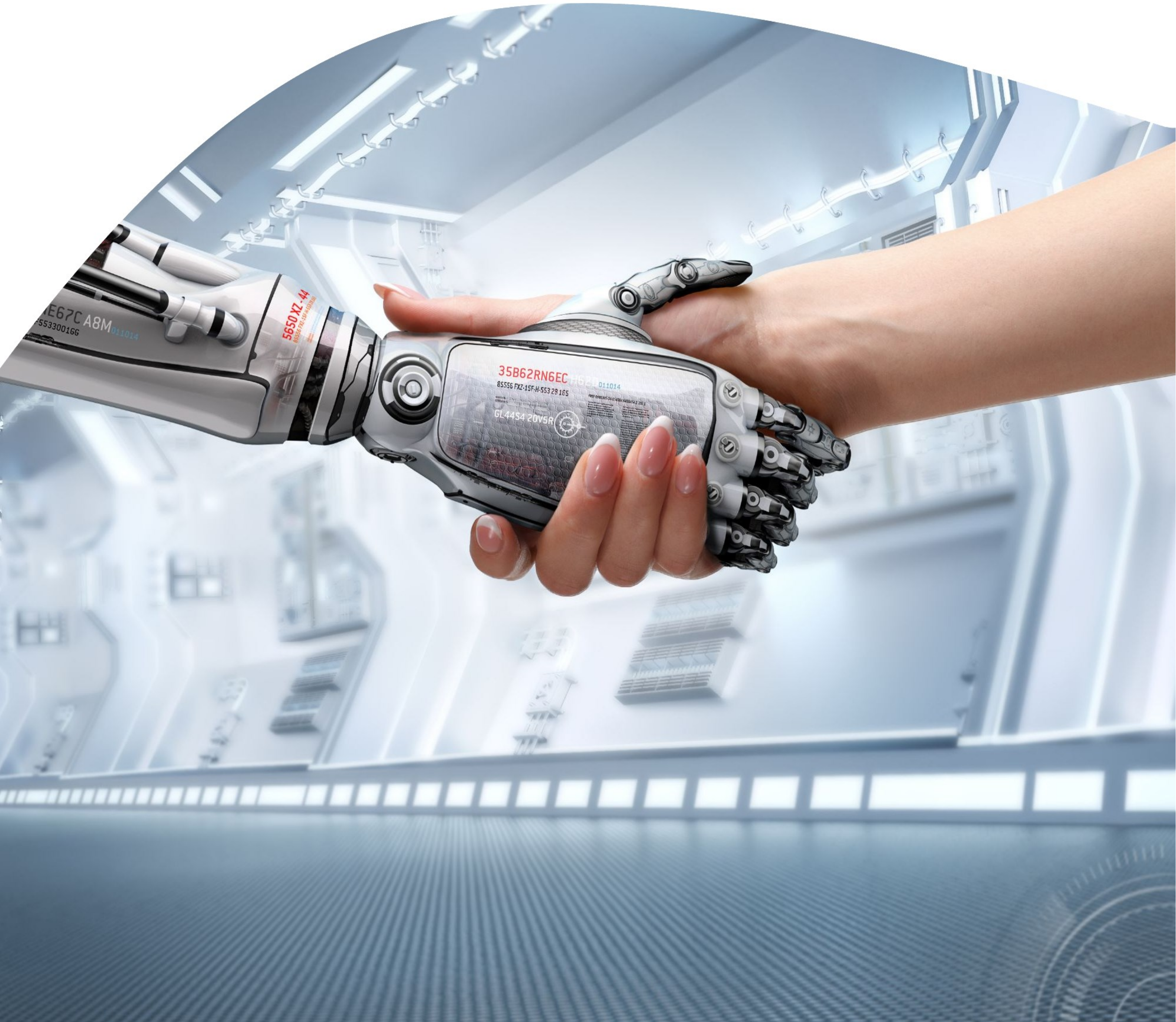
- **Pattern recognition** involves the automatic discovery of meaningful patterns in datasets. It enables systems to recognize and interpret information, attributes or characteristics that repeat or exhibit similarities, which are indicative of the patterns being sought. Pattern recognition is mostly used in cybersecurity in the form of feature extraction, the process of extracting relevant features from data to represent patterns of interest. In cybersecurity, patterns of interests are typically malicious contents and behaviours, from which a signature is extracted. Signatures are further used for pattern matching and identification of know phishing attempts, malware infection, malicious user behaviour, etc. The applications of pattern recognition are typically opposite from anomaly detection and behavioural analysis, as it aims to model specific malicious contents and behaviours rather than generic benign behaviours.

AI can also be used to assist human cybersecurity experts. The goal of these capabilities is to streamline their decision-making processes and enable them to concentrate their expertise where it is most impactful.

- **Clustering** is a machine learning technique used to group similar data points together based on their inherent characteristics or features. It aims to find natural groupings or clusters within a dataset without any predefined labels or classes. Clustering can be used to aid the analysis and categorizations of large volume of threat intelligence data or to identify and group different malware variants. It can also be combined with behavioural analysis to group and identify different user or system profiles.
- **Information retrieval** is the task of mining and extracting useful insights from content, which matches or is semantically similar to a given query. Information retrieval can be used for threat intelligence. AI algorithms can process, analyse, summarize, and categorize large volumes of threat intelligence data from various sources, helping security teams identify and respond to emerging threats more effectively.
- **Ranking** refers to the task of arranging a set of items in a specific order based on their relevance, importance, or likelihood of meeting a user's needs or preferences. ML performs ranking by learning from labelled or implicitly ranked data to create a ranking model. Ranking is typically used in cybersecurity to prioritize the efforts of security experts. It can be used to prioritize patching vulnerabilities in vulnerability management, and to triage incidents in incident response processes or events in security information and event management (SIEM) systems.
- **Generation** is the task of creating new content that fits a previously determined target distribution. Generation can be used in cybersecurity for vulnerability assessment and penetration testing. AI-based generation tools can be used simulate attacks and identify vulnerabilities within systems or networks. This can help organizations find vulnerabilities before they are exploited to proactively strengthen their defences.

2 Considerations for using AI in cybersecurity

The integration of AI in cybersecurity presents both promises and challenges for organizations aiming to bolster their security frameworks. This transition is not seamless, primarily due to the limited familiarity cybersecurity experts have with AI techniques and their advantages. However, the adoption of AI in cybersecurity endeavours often yields several sought-after benefits, which are only possible to reach by acknowledging the peculiarities and challenges of applying AI in cybersecurity. Real-time response, environment dynamicity, adversarial nature, usability/security trade-offs, explainability or privacy are examples of concerns that must be addressed.



2.1 Drivers & benefits from AI

Adopting AI-based solutions to address a particular cybersecurity issue is not a seamless transition. Generally, cybersecurity experts lack familiarity with AI techniques and their benefits. As a result, the conventional approach relies on human expertise and manual tasks until they prove inadequate. However, when organizations do opt for AI, they often seek and frequently attain the following benefits depicted in Figure 1.

Speed and Automation

Scale and Complexity

Adaptability

Efficient resource utilization

Discovery of new attacks/threats

Figure 1: Benefits from AI in cybersecurity

Speed and automation. Cyberattacks can unfold in a matter of seconds and a quick response is required to mitigate them and minimize their impact. Digital environments generate a vast amount of data that is relevant for security monitoring, e.g., files, objects, security events, security alerts, etc. It is challenging for security operators to analyse and respond to threats effectively. AI enables the processing and analysis of large-scale, diverse data sources in real-time. It can summarize, observe and extract patterns from large amounts of complex data in a very short time. AI-driven automation can streamline incident response processes, enabling faster and more effective actions to contain, mitigate, and recover from security breaches.

Scale and complexity. Modern digital environments are increasingly complex and exposed to numerous threats. The sheer amount of security monitoring relevant data they produce needs to be analysed at scale in a timely manner. AI enables scalability in cybersecurity operations, allowing organizations to handle and process massive amounts of data efficiently, regardless of scale or complexity. While most threats could be detected using manually defined rules, their large and increasing number discourage this costly approach anymore. AI automates and abstract the process of rule creation by including in a single AI model the knowledge from potentially 1000s of rules, which are learned implicitly and automatically. AI can also bring customization at scale to address problems similar in nature but depicting variability. For instance, AI enables tailored behavioural analysis for users, devices, or networks, detecting anomalies and unusual patterns that signal potential threats aligned with their respective expected behaviours. This customization extends to groups based on roles, organizations, or location.

Adaptability. Cybersecurity is characterized by an evolving threat landscape. Threats, vulnerabilities, and attacks change quickly over time, requiring defences to evolve at the same pace. Detection rules need to be redefined or adapted, signatures from new malware need to be extracted, patterns of new attacks need to be identified, etc. AI models can quickly learn from updated data using automated retraining, which does not require to change the design or features of AI models. They can adapt to evolving attack patterns, enhancing their ability to detect and prevent emerging threats. Regular training cycles also enable models to continuously learn from their evolving sample population, which includes analyst-labelled detections or analyst-reviewed alerts. This prevents recurring errors and enables models to learn and enforce expert-generated ground truth.

Efficient resource utilization. Security expertise is scarce and automation through AI augments human capabilities, automating mundane repetitive tasks, and freeing up expert time for more strategic and complex security operations. For instance, AI can quickly analyse, synthesize, and correlate large volumes of historical and dynamic threat intelligence, enabling teams to operationalize data from various sources in near real-time. It enables security analysts to effectively prioritize resources to address their organization's critical vulnerabilities and investigate time-sensitive security alerts and detections. It can also reduce alert fatigue in detection & response scenarios by correcting mistakes and ranking security events by order of importance.

Discovery of new attacks/threats. Through capabilities like anomaly detection and behavioural analysis, AI techniques have the potential to discover new and emerging threats. By modelling normal behaviours and recognizing deviations as potential new attacks, these AI techniques do not need to define a threat or attack to detect it. They can assist security teams in proactively identifying and mitigating new threats and attacks. However, this potential is tempered by the prevalence of noise and disturbances, given that most anomalies are not necessarily tied to malicious behaviours.

2.2 Particularities and challenges of AI for cybersecurity

AI models used in cybersecurity must be interpretable and explainable to ensure that security professionals can trust them, understand their decisions, and take appropriate actions to enhance security postures effectively.

Applying AI for any application comes with a set of challenges. Cybersecurity is a particular domain of AI applications, which comes with its own characteristics, considerations, and challenges. Acknowledging these peculiarities helps in planning and implementing AI solutions effectively while mitigating potential risks and limitations to ensure trust and reliability in cybersecurity applications.

Real-time response. Cyberattacks can cause significant damage within minutes or even seconds. Real-time response enables swift action upon detecting threats, helping to prevent them or minimize their impact. Quick automation is necessary to perform tasks such as threat mitigation or detection & response. AI models for cybersecurity must be efficient, scalable, and able to operate in real-time. This involves optimizing models for fast inference, reducing computational overhead, and ensuring accurate and timely predictions without compromising security.

Dynamic environments. Cybersecurity operates in dynamic environments. The threat landscape evolves continuously with threat actors developing new attack techniques and malware variants to exploit vulnerabilities. Benign behaviours are also dynamic with changing systems configurations and variability in user behaviour. AI models for cybersecurity need to be agile, robust, and capable of learning from and adapting to the ever-changing cybersecurity landscape. This involves continuous monitoring, updating, and retraining to ensure their effectiveness in detecting and responding to emerging threats and changes in the system landscape.

Adversarial nature. The presence of an adversary is a given in cybersecurity applications. Attackers constantly evolve their techniques to bypass defences and deceive security measures. When targeting AI-based defences, attackers can use adversarial attacks specially designed to fool AI models, such as model evasion or poisoning. Addressing this adversarial nature requires a comprehensive understanding of potential threats, constant vigilance, and the implementation of proactive defence strategies to fortify AI models against attacks. AI models for cybersecurity need to be robust and resilient to withstand various forms of attacks, where attackers deliberately manipulate data to deceive AI systems.

Cost of errors. The effectiveness of security solutions is often in tension with the usability of systems they protect. Errors like false positives (flagging normal behaviour as malicious) and false negatives (missing actual threats) have significant consequences on usability and security respectively. The impact and cost of these different errors is context specific, and it often depends on the criticality of the systems to protect. Balancing between minimizing both types of errors is crucial in cybersecurity to provide a sensible security/usability trade-off in a specific context. AI models for cybersecurity must have high accuracy and their accuracy performance (e.g., precision, recall, false positive rate, etc.) must be parametrizable to reach a targeted trade-off between security and usability.

Data challenges. In cybersecurity, datasets often suffer from class imbalance, where instances of normal behaviour or benign data significantly outnumber instances of malicious activity. Moreover, anomaly detection and threat identification often suffer from a lack of labelled data due to the scarcity of known attacks. AI models for cybersecurity need to generalize well with limited labelled examples. They also need to handle class imbalance to prevent biases toward the majority class. This can be partly addressed through the choice of AI model types (e.g., ensemble methods) and training approaches (e.g., semi-supervised or active learning) combined with data augmentation.

Interpretability and explainability. Understanding why an AI model made a decision is often necessary in cybersecurity. It increases trust and transparency for security analysts having to act based on AI decisions. It helps in identifying the root cause of an issue, understanding the attack vectors, and devising effective countermeasures. AI models used in cybersecurity must be interpretable and explainable to ensure that security professionals can trust them, understand their decisions, and take appropriate actions to enhance security postures effectively. This involves using interpretable AI models or explainable methods (e.g., SHA, LIME) together with producing documentation and reporting for AI systems.

Privacy concerns. Cybersecurity often involves analysing sensitive data, such as personally identifiable information (PII), which raises concerns about data exposure or misuse when used in AI models for cybersecurity. Addressing these privacy concerns involves a combination of technical measures, ethical considerations, regulatory compliance, and transparent communication. Balancing the need for effective cybersecurity measures with protecting individuals' privacy rights is essential for responsible and trustworthy use of AI in cybersecurity. AI models for cybersecurity must maintain confidentiality while analysing potentially sensitive information.

Addressing these domain specificities often involves developing specialized algorithms, employing anomaly detection techniques, using ensemble models for improved accuracy, incorporating domain knowledge, and emphasizing model explainability.

3 AI applications for cybersecurity

For 25 years, AI has been integral to the cybersecurity industry, initially with the inception of spam and phishing email filters. Over time, its applications have expanded across numerous cybersecurity domains. While it has matured in reactive security measures such as threat detection or endpoint security, employing this technology for proactive security applications like threat intelligence or vulnerability management has presented some challenges*.*

This section outlines the historical and potential applications of AI in eight primary cybersecurity areas. It offers examples of AI utilization and specifies the AI tasks or capabilities involved (refer to section 2.2). Additionally, it highlights challenges associated with each application and assesses their respective maturity levels. Figure 2 and table 1 summarize these findings. This section is intended to be read in a selective rather than a comprehensive manner, based on interests of the reader for specific AI applications. Different applications use similar AI techniques and face similar challenges; thus, the content of the following sections may overlap. For the interested reader, more comprehensive and technical details about these applications can also be found in academic survey papers[1][2][3].

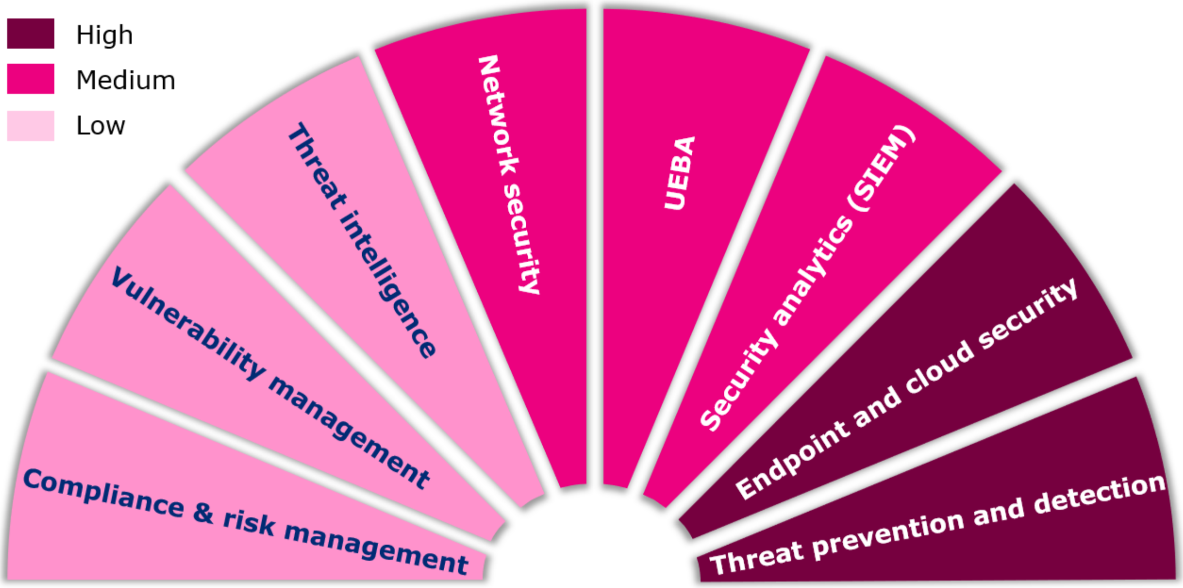


Figure 2: Maturity level of AI cybersecurity applications

3.1 Threat Prevention and Detection

Threat prevention and detection involves implementing strategies and tools to safeguard an organization's assets from potential cybersecurity threats by detecting malicious contents that can compromise their confidentiality, integrity, or

[1] Apruzzese, G. et al. "The role of machine learning in cybersecurity." *Digital Threats: Research and Practice*. 2023

[2] Dasgupta, D. et al. "Machine learning in cybersecurity: a comprehensive survey." *The Journal of Defense Modeling and Simulation*. 2022

[3] Sarker, I.H. et al. "Ai-driven cybersecurity: an overview, security intelligence modeling and research directions." *SN Computer Science.* 2021

availability. In the context of an increasingly interconnected world, the application of AI in threat prevention and detection is a cornerstone of robust cybersecurity strategies and AI has been used for this purpose for over 20 years. An early and successful application of AI for malicious content detection has been in the form of *pattern recognition*, to extract compact signatures from malicious contents such as malware files and using these signatures for further matching and identification of unknown software files.

Malware detection is a primary application for AI in threat detection where the main approach relies on binary classification to distinguish between benign and malware files. The goal of AI models is to identify malicious code hidden within legitimate executables, documents, or URLs. Features used for this identification originate for file analysers either static, reviewing source/binary code for quick responses, or dynamic, executing code in a sandbox environment for accurate but slower results. AI models designed for malware detection possess the capability to acclimate to the ever-evolving landscape of malware. They excel in extrapolating knowledge, thereby assisting in the identification of new malicious samples that employ attack techniques akin to those utilized by known malware.

AI has historically proven its efficacy in phishing and spam detection, but recent strides in Deep Learning and Natural Language Processing (NLP) have propelled substantial advancements. Traditional binary classification methods have been and remain the primary approach for detecting spam, phishing attempts, and analysing email and website content. However, the emergence of Deep Learning techniques has reduced the need necessity for feature engineering. This evolution allows for the direct processing of email or SMS messages as raw text input to AI models, streamlining the analysis process and enhancing the adaptability of AI-based detection techniques. Similar success has been met by classification methods for identifying and blocking malicious domains using features extracted from URLs and scraped websites[4]. The best results involve a multi-layered approach, leveraging diverse types of features to ensure robust and comprehensive protection against malicious URLs[5].

While being an effective approach, using AI for threat prevention and detection also encounters several challenges. One of the primary hurdles involves the ever-evolving threat landscape, where new malware types and attack methods emerge regularly. This necessitates constant adaptation and re-learning for AI models. The handling of incorrect decisions such as the misidentification of benign activities as threats or missing actual threats can lead to unnecessary alarms or undetected breaches. The lack of interpretability or explainability in AI models is also problematic, as understanding the rationale behind malware detection is critical. In the context of malicious domain detection, an extra challenge emerges when attackers conceal site content behind CAPTCHAs, a scenario currently unaddressed by available methods.

The maturity level in using AI in threat prevention and detection is high. It is a critical component with applications spanning software exploitation identification, malware identification, phishing and spam detection, and malicious domain blocking. Machine Learning and Deep Learning have been effectively utilized in

[4] Marchal, S. et al. "Know your phish: Novel techniques for detecting phishing sites and their targets." *36th IEEE International Conference on Distributed Computing Systems (ICDCS).* 2016

[5] Cohen, D. et al. "Website categorization via design attribute learning." *Computers & Security.* 2021

these domains, enabling proactive identification and mitigation of threats. This has significantly enhanced the effectiveness and efficiency of cybersecurity strategies, making them better equipped to handle the increasing complexity of modern IT systems.

3.2 Endpoint and Cloud Security

Endpoint and Cloud Detection & Response services (EDR and CloudDR) are essential for providing advanced security solutions for local devices and cloud-based services. These services monitor and analyse the events that occur on devices and servers, such as file operations, network connections, and user actions, to detect and mitigate potential threats in real-time. The main objective of these services is to generate a comprehensive and accurate list of incident information and respond to incidents to mitigate cyberattacks and to mitigate their impact.

AI plays a pivotal role in bolstering EDR and CloudDR services by efficiently handling the vast and intricate data stemming from devices and cloud services. A first successful application of AI in detection & response is to filter out harmless and irrelevant security events based on their prevalence. EDR and CloudDR systems generate many security events that could be relevant to detecting cyberattacks when correlated. However, many generated events are also common and unrelated to attacks. Statistical modelling techniques can easily identify and filtered out such events from further analysis based on their frequency, duration, and similarity. This reduces the processing load of event correlation systems and reduces the false alarms that would be potentially triggered if these harmless events were processed. AI can also aid in detecting anomalous chains and sequences of events that deviate from typical behaviour. These chains can be related to malicious activities like malware execution, data exfiltration, or privilege escalation. Leveraging anomaly detection techniques based on sequence mining and graph analysis, AI can identify these chains and generate system alerts to cybersecurity experts for further investigation and response.

AI also improves the prioritization and analysis of aggregated and contextualized information about detected suspicious events. Clustering techniques can be used for grouping similar incidents together and extracting a prototype, which is a unique representative element of the whole cluster. This helps saving time for cybersecurity analysts who only need to analyse a single incident, instead of several similar ones, to provide a verdict and a response for incidents within the cluster. AI can also help in ranking incidents according to their severity, impact, and urgency, and provide guidance for the best course of action. By improving the prioritization and analysis of incidents, the response process of cybersecurity experts mitigates threats more efficiently and effectively.

The use of AI in detection & response requires frequent adaptation of the models to the dynamic nature of events characteristics. It is also important that AI models are tailored to the specific context and purpose of devices and services, such as operating systems, applications, and functions. The performance evaluation of AI models is challenging, as the data is usually highly imbalanced, with only a small fraction of the events being related to cyberattacks. AI models used for detection & response must be accurate and robust, as false positive incidents can reduce the usability and trustworthiness of the solution. Models must also be fast and responsive, as any delay in detection and response can increase the impact caused by successful attacks.

The maturity level of AI applications for EDR and CloudDR is high, indicating that AI has become an integral part of these services, enhancing their efficiency and effectiveness in detecting and responding to potential threats. AI can provide valuable insights and support for cybersecurity experts, enabling them to protect devices and services from cyberattacks.

3.3 Network Security

AI offers advanced capabilities for the monitoring and analysis of large amounts of network traffic across complex network infrastructures. The main application of AI for network security relates to *Intrusion Detection* and *Intrusion Prevention Systems* (IDS and IPS), where network traffic patterns are analysed to detect anomalies and suspicious activities that may indicate unauthorized access, compromised assets or potential attacks. Another popular network security application relates to the analysis of unknown network traffic to identify its source, being it a user, device, or application, potentially malicious communication, e.g., *Command and Control* (C&C) traffic or worm spreading in the network.

IDS and IPS typically leverage anomaly detection capabilities to model normal network traffic behavior and detect deviations as potential signs of attacks. This approach is particularly effective to detect obvious attacks generating significant change in network traffic such as port and service scanning or *Denial-of-Service* (DoS) attempts. Anomaly detection is much less effective at detecting subtle attacks and malware communications which are stealthier and resemble normal communications. Flagging them would require tuning anomaly detection to be more sensitive, which would lead to many false alarms for anomalies that are not malicious. To address this type of threat, IDS and IPS rely on classification methods, which provide better accuracy. Binary classification is used to distinguish benign from malicious communication while multi-class classification is used to identify different types of attacks and malicious behaviours in a fine-grained manner. User, device, and application identification rely instead on pattern recognition where the network communication generated by the entity to identify is tightly modelled and further used in pattern matching. For instance, signal processing techniques coupled with simple classification methods can be used to identify IoT devices in home networks[6].

Features used as inputs to these ML algorithms are extracted from the network traffic with the aim of representing the type of communication occurring in the network. They are extracted at different granularity levels depending on a set of requirements and constraints stemming from the activity to detect, analysis latency, available processing power as well as availability of protocol information. Network traffic encryption limits the information available for feature extraction. Features can be computed from single network packets, groups of network packets, communication flows, or from all network traffic captured from a specific vantage point.

[6] Marchal, S. et al. "AUDI: Toward autonomous iot device-type identification using periodic communication." *IEEE Journal on Selected Areas in Communications*. 2019

A main challenge in network security applications relates to the diversity of network traffic to analyse. The abundance of communication protocols makes it difficult to model complex environments and behaviors. Network traffic is also increasingly encrypted, preventing access to fine-grained packet and payload features that contain relevant information. This jeopardizes the identification of stealthy attacks, which can be more easily cloaked and adapted to match benign network communication when the granularity of captured network traffic is low. AI applications to network security are the most successful in monitoring simple, stable, and predictable network environments composed of low-end devices with simple behaviours, such as Internet-of-Things (IoT) networks or Industrial Control Systems (ICS). An example of a successful network anomaly detection system for IoT network relies on low granularity network data and builds device-specific AI models of their behaviour using Recurrent Neural Network (RNN), a type of DNN for modelling sequences[7].

Despite its long history, applying AI to network security remains at a moderate maturity level. This is mostly due to the constant evolution of networks and their complexity, which regularly raises new challenges for applying AI techniques to improve their security.

3.4 User and Entity Behaviour Analytics (UEBA)

A key to success in UEBA lies in understanding and evaluating the relevance and accuracy of detected abnormal event.

User and Entity Behaviour Analytics (UEBA) is a cybersecurity approach that focuses on analysing and understanding the behaviour of users, devices, applications, and entities within an organization to detect potential security threats and anomalous activities[8]. UEBA aims for a proactive and adaptive security approach, leveraging machine learning to detect subtle and evolving threats, reducing response times, and enabling security teams to identify and address potential security incidents more effectively.

UEBA applies behavioral analysis capabilities on user and entity event data, which is by its nature typically both voluminous and heterogeneous. Typical events to model and monitor include login attempts, downloads/uploads of documents, and the usage of certain applications and processes. Event characteristics that are considered for modelling include the type of the event, its time of occurrence, location, duration, author, and similar factors. They can be modelled using unsupervised clustering techniques and time series analysis to capture the location, temporality, seasonality, and similarity in the occurrence of these events. These AI models or *profiles* establish a baseline of typical activities, access patterns, and interactions for a given user or entity. New events initiated by these entities and users can then be compared against the learned profiles to identify if they seem normal or anomalous. The flagged events can then be investigated for signs of potential insider threats, unauthorized access, or other abnormal behaviour. AI-

[7] Nguyen, T.D. et al. "DÏoT: A federated self-learning anomaly detection system for IoT." *39th IEEE International conference on distributed computing systems (ICDCS).* 2019

[8] Shashanka, M. et al. "User and entity behavior analytics for enterprise security." *IEEE International Conference on Big Data.* 2016

based UEBA services are already provided to mitigate identity-based risks in organizations[9].

Building an accurate behavioural model requires a large amount of historical event data about the profiled entities and users. This means that UEBA needs an extended learning period, during which it can not provide protection for new entities. A further challenge is that entities should not depict abnormal behaviors during this learning period to ensure that profiles do not include and learn anomalies. The behavior of entities targeted by UEBA is typically complex and dynamic, which presents an inherent variability that is exacerbated over time. This makes UEBA highly prone to false alerts for abnormal events which do not represent any threat. It also means that behaviour profiles need to be regularly retrained and updated to capture evolutions in behaviour. A key to success in UEBA lies in understanding and evaluating the relevance and accuracy of detected abnormal events. Based on this understanding, responses to detections need to be adapted to strike a tradeoff between a high level in mitigating and preventing risky events and low user disturbance.

The responses must be adapted to the security risk level of an event and the potential consequences. The responses must be easily parametrizable and modifiable over time. The parameters are typically manually defined and can be specific to different user groups for the same type of event. Let’s consider as en example event the download of a confidential file in the middle of the night from an abnormal location. Example responses might include

a) blocking the action,

b) asking for additional authentication (e.g., multi-factor authentication),

or c) raising a silent alert having no direct impact to the user.

Training a UEBA system and tuning its responses to achieve a sensible compromise of usable security can be a long process. The implementation of a feedback loop for users to comment on and correct inappropriate responses is key to improving UEBA performance and reaching this compromise.

UEBA is a rather recent application of AI in cybersecurity and its maturity level is medium. The challenges in modelling the diversity and dynamicity of user and complex entity behaviour result in a comparatively low accuracy, which requires a lot of detection fine-tuning in the response phase. This prevents the usage of UEBA in critical automated decision making.

3.5 Security Analytics

Security Information and Event Management (SIEM) systems integrate and analyse data from various sources, such as event data analysis (EDR and ClouDR), user and entity behaviour analytics (UEBA), network security, and other security systems. This integration enables a holistic view of the security posture of the entire corporate network and facilitates the identification and mitigation of emerging threats. Traditional SIEM solutions face several limitations, such as producing many

[9] *What is Entra ID protection?*. https://learn.microsoft.com/en-us/entra/id-protection/overview-identity-protection

false positive alerts, relying on a limited amount of data for analysis, and lacking the ability to respond quickly and effectively to incidents.

AI can address these challenges and enhance SIEM solutions with various capabilities. AI can automate the complex processes of data aggregation, normalization, enrichment and correlation through information retrieval and pattern recognition techniques. AI can correlate past security events and use threat intelligence to detect and prevent advanced persistent threats (APTs) that evade traditional security measures. It can potentially automate alert generation and implement predefined response actions or partially orchestrate complex response workflows using generation methods. However, full automation is complex to achieve and prone to errors given the current capabilities of generative AI methods and their limitations, like hallucinations. It would require a careful training of AI models, giving them knowledge of team organization, individuals capabilities, existing workflows, etc. More promising approaches to using AI in SIEM is to prioritize and support the investigation of security events for security analysts, much like in endpoint detection & response systems as discussed above. Clustering techniques can be used to group security events and decrease the workload of manual analysis. Ranking can also be used to prioritize the manual analysis of security events by SIEM operators.

The integration of AI into SIEM systems holds considerable promise, but it also comes with a set of challenges. Data acquisition can reveal difficult since it is often unavailable, of low quality or the collection can be restricted due to privacy issues and to its sensitive nature. Moreover, the normalization and correlation of data of different nature from different systems is an additional challenge. There are instances where AI models may incorrectly classify harmless activities as threats, known as false positives, or fail to detect actual threats, known as false negatives. Both situations can lead to serious consequences, such as unnecessary alarm or unnoticed security breaches.

Although AI has been applied for SIEM applications, the maturity level of this technology remains medium, meaning that there is room for further improvement and innovation. AI can use natural language processing and reasoning to assist cybersecurity experts in querying databases and conducting initial cybersecurity evaluations. This would enable experts to focus on more complex and critical cybersecurity tasks and provide them with more context. Additionally, AI can use generative models to create realistic and diverse scenarios and cyberattack simulations. This would help security teams to test and improve their SIEM solutions, as well as train and evaluate their skills and readiness.

3.6 Threat Intelligence

Cybersecurity threat intelligence revolves around collecting, processing, and analysing data to grasp threat actors' goals, targets, and attack strategies. It is a security measure meant to anticipate and prevent cyberattacks before they occur. The integration of AI into threat intelligence platforms holds substantial promise. AI can automate mundane tasks, enabling cybersecurity analysts to concentrate on intricate threat analysis. Its capacity to handle large datasets and spot patterns offers crucial insights that supports threat prevention and detection. This is particularly valuable in today's interconnected landscape, where the sheer volume and intricacy of threats continue to escalate.

AI can enrich threat intelligence platforms by automating the search and integration of relevant data from cyber intelligence reports and open-source threat intelligence (OSTI) feeds using *information retrieval* capabilities, ensuring continuous updates with the latest threat information[10]. LLMs can also readily be used in a simple manner to summarize and synthesize cyber intelligence reports, saving time for threat analysts. Aside from known sources of threat intelligence, information retrieval techniques can also be used to find threat intelligence from uncommon sources, such as social media where vulnerability information could be found using Deep Learning techniques[11]. Additionally, AI can track the evolution of crucial themes in cyber threats over time, enabling the identification of patterns and prediction of future threats, which fosters a proactive approach to threat management[12]. Considering the diverse global sources of intelligence often presented in multiple languages, AI can also be used to translate these multilingual threat sources, preventing critical information from being overlooked due to language barriers. Classification abilities and explainable AI methods can also guide cybersecurity experts through threat analysis dashboards, directing their attention to emerging attack trends.

While the integration of AI into threat intelligence platforms has significant potential, it also presents several challenges. It needs volumes of high-quality data, and it can potentially generate incorrect outputs, which can significantly harm the quality of information available in threat intelligence portal. The processing of large amount of heterogeneous input information sources represents a challenge in ensuring that relevant data is not missed, and all data of different nature is properly correlated to generate sensible outputs.

3.7 Vulnerability Management

The potential benefits of AI for vulnerability management are significant. It already brings benefits for supporting cybersecurity analysts in penetration testing activities and reduces the costs and efforts of manual vulnerability analysis and mitigation. As generative AI methods improve, they are expected to play a more significant role in vulnerability management in the future.

Vulnerability management is the process of identifying, assessing, and mitigating the risks posed by security flaws in IT systems. Vulnerability management is supported by a collective industry effort to record and maintain public databases of known vulnerabilities, such as the Common Vulnerabilities and Exposures (CVE) list. While comprehensive information about vulnerabilities in common libraries and software is available, keeping an up-to-date knowledge of newly discovered vulnerabilities and how they affect an organization, as well as finding new

[10] Pantelis, G. et al. "On Strengthening SMEs and MEs Threat Intelligence and Awareness by Identifying Data Breaches, Stolen Credentials and Illegal Activities on the Dark Web." *16th International Conference on Availability, Reliability and Security.* 2021

[11] Iorga, D. et al. "Yggdrasil—early detection of cybernetic vulnerabilities from Twitter." *23rd International Conference on Control Systems and Computer Science.* 2021

[12] Kim, G. et al. "Automatic extraction of named entities of cyber threats using a deep Bi-LSTM-CRF network." *International journal of machine learning and cybernetics*. 2020

vulnerabilities in original source code are challenging. AI can help in addressing these challenges.

AI can be used to easily keep track of new vulnerabilities in databases and test if they apply to the software or code of an organization. For instance, natural language processing (NLP) combined with ranking techniques can be used to list and prioritize the known vulnerabilities that most likely affect a given software program[13]. Deep learning used in combination with pattern recognition can be used to identify know vulnerabilities in source code[14]. Automated vulnerability categorization is another domain where AI capabilities like classification can be used to divide the detected vulnerabilities in an organization into predefined groups from the Common Weakness Enumeration (CWE) categories[15]. This procedure helps to organize and prioritize the vulnerabilities according to their severity, impact, and exploitability according to CVE information. AI methods have also been used to improve the discovery of new unknown vulnerabilities by supporting software fuzz testing. Generative deep learning methods can be used to guide fuzz test case generation towards improving coverage of vulnerabilities, rather than global code coverage, thereby optimizing and improving vulnerability discovery[16]. There is some hope that generative AI methods could also be used directly in an offensive manner to fuzz software program on their own, but there has not been a very successful example of such an application yet. Similarly, new tools based on generative AI methods and LLMs show promises in assisting red teaming and penetration testing exercises[17]. These solutions can effectively support the choice of offensive tools, the selection of next steps for tests and the interpretation of test results, while still presenting limitations.

A key challenge for using AI in vulnerability management is the lack of sufficient and reliable data for training and testing the AI models, especially for the purpose of detecting unknown and zero-day vulnerabilities. Generative AI methods still struggle at generating code or executable commands that need follow a strict syntax to be compiled and executed. A single wrong character in text generated for human reading is acceptable, while in executable commands, source code or network packets, it can make their interpretation impossible. Finally, the ethical and legal implications of using AI for vulnerability discovery and exploitation may pose a threat to the security and privacy of IT systems, since these AI systems can potentially be repurposed by attackers to perform actual cyberattacks.

The maturity level for using AI in vulnerability management is low and still in the pilot stages, as most of the techniques are still in development or testing and have not been widely adopted or deployed in real-world scenarios. However, the potential benefits of AI for vulnerability management are significant. It already brings benefits for supporting cybersecurity analysts in penetration testing activities and reduces the costs and efforts of manual vulnerability analysis and mitigation. As

[13] Huff, P. et al. "A recommender system for tracking vulnerabilities." *16th International Conference on Availability, Reliability and Security* (ARES). 2021

[14] Jeon, S., and Kim, H.K. "AutoVAS: An automated vulnerability analysis system with a deep learning approach." *Computers & Security.* 2021

[15] Saha, T. et al. "SHARKS: Smart hacking approaches for risk scanning in Internet-of-Things and cyber-physical systems based on machine learning." *IEEE Transactions on Emerging Topics in Computing*. 2021

[16] Wang, J. et al. "Skyfire: Data-driven seed generation for fuzzing." *IEEE Symposium on Security and Privacy.* 2017

[17] Deng, G. et al. *PentestGPT: An LLM-empowered Automatic Penetration Testing Tool.* arXiv. 2023

generative AI methods improve, they are expected to play a more significant role in vulnerability management in the future.

3.8 Compliance and Risk Management

Security risk management is the process of identifying, estimating and prioritizing cybersecurity risks associated with operations, assets, and individuals within an organization. The Cybermeter (Kybermittari) introduced by Traficom[18] is an example tool for security risk management and evaluation of the maturity level of an organization's cybersecurity. AI can support and reshape the complex, costly and time-consuming process of security risk management, which typically requires active human involvement and expertise. The automation of risk analysis, cybersecurity capability evaluation, and impact assessment fortifies the capabilities of risk management teams by harnessing both internal and external risk data, thereby facilitating real-time evaluation of risk factors and associated metrics. AI stands as a catalyst to automate multifaceted tasks such as calculating risk scores, inferring the probability of potential security incidents, identifying critical vulnerability risk indicators, and conducting comprehensive risk assessments and decision analysis.

AI can ease asset management, a foundation for risk management, by automatically identifying, categorizing, and keeping track of information, people, equipment, and systems that support an organization to accomplish its goals. The automation of asset management can be supported by classification and pattern recognition techniques applied to each asset of an organization. Combined with scanning approaches that would regularly identify the appearance of new assets or the removal of old assets, AI can help keep an up-to-date inventory of assets. Features representing the identified assets can further be used as input to ranking algorithms to automatically identify their relative importance and their security risk. These results can in turn be used to prioritize the deployment of security measures to protect and reduce the risk associated with high value assets. Similarly to asset management, AI can be used to make an inventory of deployed security measures within an organization and help automatically computing how these defenses mitigate the security risk in correlation with the identified threats and vulnerabilities. Such an application helps assess the security posture of an organization by automating these typically manual inventory and rating tasks, like the ones required by the Cybermeter.

AI techniques can also support supply chain risk management by automating threat analysis and prediction, optimizing cybersecurity investment, and assessing the cyber resilience of supply chains. Resilience assessment can be generalized to evaluate the protection, detection, response, and recovery mechanisms related to any asset and processes within an organization. This automation is enabled by information retrieval and correlation techniques applied on systems and security documentation, asset inventory, historical incident data, network and system logs, vulnerability data and threat intelligence from both within and outside the organization.

AI can also play a significant role in bolstering security compliance and governance within organizations, thereby decreasing their security risk. For instance, AI has

[18] Traficom. *Kybermittari - Cybermeter.* https://www.kyberturvallisuuskeskus.fi/en/our-services/situation-awareness-and-network-management/kybermittari-cybermeter

already been used for policy enforcement in communication networks[19]. Policy proxies can be deployed in routers to identify the network traffic subject to policies using classification methods and assisting in applying these policies to ensure adherence to regulations. Information retrieval and generation capabilities also have a potential for automated extraction of risk indicators and converting them into actionable insights to proactively prevent cybersecurity breaches by swiftly addressing emerging risks.

The maturity of AI applications for security compliance and risk management remains low. It relies on analyzing, synthetizing, and extracting highly targeted information from large amounts of heterogeneous data. There is no single definition for risk or unique process for risk assessment and risk management. Thus, it is difficult to train AI systems for performing a task which has no clear definition. Nevertheless, the recent advances in information retrieval and generation provided by LLMs hold great promises for this type of application.

[19] Odegbile, O. et al. "Dependable Policy Enforcement in Traditional Non-SDN Networks." *39th IEEE International Conference on Distributed Computing Systems (ICDCS).* 2019

Application	AI Capabilities	Maturity
Threat prevention and detection	- Classification - Pattern recognition - Clustering	High
Endpoint and cloud security	- Classification - Anomaly detection - Pattern recognition - Clustering - Ranking	High
Network security	- Classification - Anomaly detection - Pattern recognition	Medium
UEBA	- Behavioural analysis - Clustering	Medium
Security analytics (SIEM)	- Pattern recognition - Anomaly detection - Information retrieval - Ranking - Generation	Medium
Threat intelligence	- Classification - Clustering - Information retrieval - Generation	Low
Vulnerability management	- Classification - Pattern recognition - Ranking - Generation	Low
Compliance and risk management	- Classification - Information retrieval - Ranking - Generation	Low

Table 1: AI cybersecurity applications, their leveraged capabilities and maturity level.

4 Recommendation and good practices: How to use AI?

Developing AI models for cybersecurity presents a complex undertaking fraught with numerous pitfalls, the avoidance of which usually comes from experience. Drawing insights from a successful AI development instance and input from data scientists and cybersecurity experts, a set of recommendations is exemplified and provided here to guide the inception and enhancement of AI systems tailored for cybersecurity applications.

4.1 Story of a successful ML application development

In AI projects, the likelihood of failure is high due to the complexity of implementation and deployment, and the uncertainty of outcomes. Both these aspects must be addressed as early as possible through the definition of business objectives and feasibility studies.

Several considerations come into play when developing AI applications for cybersecurity. We exemplify how these considerations should be applied by describing the development of an AI system for cybersecurity. From this example, a generic methodology for successful AI development can be drawn and generalized to other cybersecurity use-cases. The example application is a detection system for malicious portable executables (PE), which is a file format for executable applications.

A good AI development process starts with the definition of business objectives for the application to develop. Considerations regarding the need for an AI-based solution and possible non-AI alternatives should be discussed then: benefits and challenges of AI-based security solutions must be weighed against each other to identify potential gains. For the PE detection system, the volume of files to be analyzed, the diversity of malicious PEs and the complexity in identifying them called for an AI-based solution. The detection of malicious PE files was a feature already provided by a product from an external vendor. The business objectives for this new solution were two-fold 1) gaining independence from the external vendor and 2) improving the performance and accuracy of malicious PE detection. Based on this, concrete metrics were defined to evaluate the target performance of the system, which set criteria for success against which the system should be regularly evaluated. The evaluation metrics for the PE detection model were the cost of operation, overall accuracy of detection, false positive rate, and coverage. Threshold values defining success were defined based on the performance of the currently used product made by the external vendor. This definition of business objectives and their translation into key performance indicators (KPIs) with clear target values is paramount to ensure the AI system brings gains from a business perspective.

After the definition of business objectives comes the definition of requirements for deployment in production. A feasibility study needs to be performed to ensure the availability of data, the integration of the solution in the existing process(es), the need and possibility to modify existing system components, the acceptable cost of operation, and other factors. The several stages of the AI pipeline must be defined, such as data collection and processing, sanitization, feature extraction, training, validation, operation, monitoring, and improvement. Each stage should be prioritized according to its importance in achieving a deployable ML system, i.e., while data collection and training are mandatory steps; sanitization, monitoring and improvement may not be. For each stage, a theoretical but feasible solution must be proposed. Blockers, challenges, and issues must be identified. For the PE detection system, data collection, data processing, feature extraction, training and validation were the minimum required steps to get a viable AI system. Each of these stages needed a realistic and deployable solution. If the implementation of a critical stage seems not feasible during this design phase, the project should be shut down to cut down on investment for a solution that has no future. For a given organization, the same issues are often recurring in different AI projects. It is important to identify these recurrent issues, keep track of them and, if needed, develop processes and tools to address them consistently.

In AI projects, the likelihood of failure is high due to the complexity of implementation and deployment, and the uncertainty of outcomes. Both these aspects must be addressed as early as possible through the definition of business objectives and feasibility studies. Only after the potential for gains and feasibility are confirmed can the development of the system start.

The next step is to assess if the AI solution has potential towards the business objectives. A prototype must be built quickly, but a lot of care must be put into gathering relevant data to train and test this prototype. It is important to collect a small experimental data set that is representative of the global data distribution that the production system will need to analyze. This data set will be used for training and validation of the ML model according to the KPIs previously defined. For the PE detection system, data from different organizations, having different activities, and from different geographical locations was collected over a continuous period. The data used for training and validation was then split according to collection time: older data to train the model and newer data to validate it. Data collection and preparation during the development phase must be representative of the real production environment to obtain sensible performance results, close to what is expected to be obtained when the real system is deployed. Even when this step is carefully implemented, it is likely that discrepancies in performance results will remain, but it is important to minimize them.

After data collection is addressed and a first prototype AI system is developed, it is important to get it to production as soon as possible by developing the minimal functionalities required for the mandatory stages of the AI pipeline. The AI system should ideally produce all the results it is supposed to, but these results should not be used yet for decision making. These results should be used to compute the KPIs initially defined (cost, accuracy, etc.) and used in an iterative process of training, validation, and improvement of the ML model until target KPI values are reached. It is also sensible to deploy the initial ML model with a limited load of data, e.g., 1-10% of the normal data stream, and to progressively increase its load to eventually reach full load. This enables cost saving during the improvement phase and it removes the scaling factor from the initial performance evaluation. An early release to production enables the quick identification of problems that may not occur in development or staging environments. This presents two advantages. First, it provides extra opportunities to identify real issues and limitations that would make the solution unfeasible, making it possible to shut down an unfeasible project early on. Second, by implementing and testing mandatory functionalities, one can identify how solutions for issues in each AI stage (e.g., data collection) may impact and generate constraints on other AI stages (e.g., sanitization or training). This leads to re-evaluating the initial feasibility requirements, to likely increase them, but also to bring them closer to reality.

For the PE detection system, this exact test deployment process was applied. The monitoring and visualization of performance results was also implemented early in this test production AI pipeline. It allowed the easy computation of accuracy and coverage KPIs and comparison between the new AI-based PE detection system and the solution from the external vendor when applied to the same production data. This results comparison led to identify the strength of each system, their scopes and, if the systems should be combined in series or in parallel, considering accuracy, running cost and latency of each system. The visualization of results was additionally used for security experts to manually analyze when the decisions between the two systems disagreed. This feedback was used to improve the AI system in a human assisted reinforcement learning fashion. Additional performance comparison was also performed using standardized tests performed by an independent testing organization. These tests are a common practice to compare and rate the performance of different products in the cybersecurity industry.

Once it was confirmed the AI-based PE detection system met its KPI and it could run with full load at an acceptable cost, its results started to be used for decision making in the detection of malicious PE. The development for optional stages of the AI pipeline, like data drift detection and improvement through feedback, were started to further improve the performance and maintainability of the system. After an AI project is deployed it is important to focus on easy and low-cost maintainability. Performance of the system can quickly decrease, and one should be able to address issues quickly in such a situation. As an organization wants to develop, test, and deploy more ML-based systems it is important to consider developing processes and libraries to assist and uniformize development and maintenance. For instance, creating common libraries for quick deployment of models and a shared AI performance monitoring function is very useful. Also, the collection and aggregation of data in a single and uniform storage system helps in accessing data for AI development and operation purposes. These initiatives greatly improve the efficiency of AI system development, testing, deployment and maintenance, thereby reducing their cost.

4.2 Keys for success in AI for cybersecurity

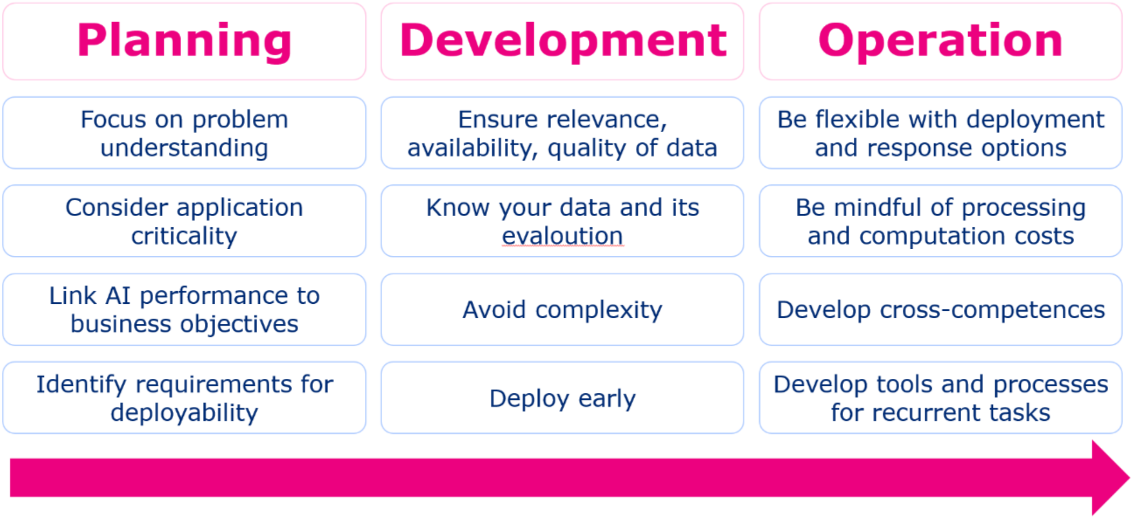
Based on the previous example of successful AI development and on feedback from data scientists and cybersecurity experts developing AI systems, we draw the following recommendations for starting or improving the development of AI systems for cybersecurity applications. These recommendations are summarized in figure 3.

- **Focus on understanding the problem**: Clearly define the problem you want to address, the need for an AI-based solution and the expected gains it will provide. Do not choose AI to solve a problem based on motivations such as AI hype, marketing value, or "having a lot of unused or under-used data". One must clearly know the challenges to solving a problem, understand the capabilities provided by AI, and evaluate if the solution fits the problem. A typical requirement in cybersecurity problems is extremely high accuracy or incredibly low false positives. One must understand if such requirements can be met by using an AI-based solution.
- **Consider application criticality:** AI systems will always make mistakes. It is important to consider how critical the decisions of the AI system (and responses to them) are. It is better to use AI solutions for applications with low criticality or make sure that there are humans in the loop when increasing the criticality. Using this process will prevent dramatic failures. Also, when starting to use AI for cybersecurity, only consider non-critical applications during the first project(s).

- **Link AI performance to business objectives:** From the start of the project, business objectives must be identified and linked to quantifiable AI performance metrics, e.g., accuracy, precision, false positive rate, latency, and cost. These performance metrics must be evaluated as early as possible and recomputed regularly throughout the project development.
- **Identify requirements for deployability**: Before going into project development, study the feasibility of deployment, the dependencies of the AI system, its integration with existing systems, the need and possibility to modify existing components. Ensure that deploying the AI system is feasible and all dependencies can be realistically addressed within budget. Many security-related AI projects end up never being used in production due to integration and deployment issues.
- **Ensure relevance, availability, and quality of data:** Avoid the common mistake of starting on AI project "from data", i.e., this data is available, let us see if we can solve a problem with it. Start from the problem, define the data needed to solve it and develop the capabilities to collect it. Then, ensuring data quality, consistency, completeness and representativity is paramount for development, validation, and production. An AI project can only be successful with high quality data.
- **Know your data and its evolution:** ML system performance can change quickly over time due to changes in the distribution of data in production environments. The data input to security processes is highly subject to these changes commonly known as *data drift*. It is paramount to understand how quickly data distribution changes and to monitor it. Solutions must be developed to cope with these changes and to prevent performance degradation. A periodic retraining of the ML model with fresh data is often the best solution. The periodicity of retraining must be chosen based on the speed of data drift and the cost of obtaining new data (which often also needs to be labelled).
- **Avoid complexity:** Do not fall into the trap of hype for deep learning and large language models. Always choose the simplest AI solution meeting the requirements of the problem being solved. Complex AI solutions always come with drawbacks such as higher running costs, higher needs for data, difficulties in understanding their decisions, and potential generalizability issues.
- **Deploy early:** The transition from test to production environment often reveals issues that were thought to be addressed when they are not. Production environments bring new constraints and challenges. For instance, data available in the production environment will almost always be different than that available during development and validation. In addition to scalability and latency issues, it can also lead to discrepancies in accuracy expectations. Developing a minimum viable product and deploying it early makes it possible to anticipate these issues with solutions that address them in production and not only in test environment.
- **Be flexible with deployment and response options:** When looking at using the results of an AI system, one must have an open mind and be flexible. The outputs of an AI project remain uncertain until the end, and different options should be considered to reap its benefits. AI systems are rarely accurate enough to be used as standalone systems in critical decision making. Different deployment options and combinations with other systems should be considered, in parallel or in a pipeline, based on the strength and scope of each system. For instance, a fast and low-cost AI system that has suboptimal accuracy could be used as first step in a pipeline. The first system only acts in cases of high

certainty and sends other cases to a second higher accuracy component, which can then afford to be costlier and slower. In addition, responses to decisions of AI systems should be adaptive and defined according to accuracy and criticality of the resulting actions. For instance, AI systems providing suboptimal accuracy should not be used to make critical decisions or be associated with intrusive responses degrading user experience and usability.

- **Be mindful of processing and computation costs:** AI used in cybersecurity often requires the parallel processing of large amounts of data, both for training and prediction. The training of complex ML models like DNNs and LLMs can be very computation intensive, and one must consider the likely need to perform this task on a regular basis because of data drift. The cost for data processing and computation can be very high in cloud services considering the price of the used hardware (e.g., GPU). As this cost can be a showstopper, it should be estimated early on, and optimization strategies should be developed to keep it acceptable.
- **Develop cross-competences**: Effective problem solving requires both technical competence in the AI solution and in the domain in which it is applied. The most effective way to develop effective AI solutions that meet their expected outcomes is by having data scientists with deep understanding of the cybersecurity domain or cybersecurity experts with advanced training and theoretical foundation in data science. This cross-competence is usually not readily available in the job market, and it must develop in-house through a lengthy training process. A starting point is to give generic cybersecurity training to data scientists AND data science training to cybersecurity experts such that both these experts can have a common language and collaborate more effectively in problem solving.
- **Develop tools and processes for recurrent tasks:** AI project time spans from conception to deployment are quite extensive. If an organization plans to develop several AI systems, it must systematize processes and develop libraries to ease and automate repetitive common tasks. For instance, libraries and systems for data loading, libraries for deployment to production, common API for AI systems, libraries for monitoring and platforms for visualizing performance should be developed to optimize AI projects.



Planning	Development	Operation
Focus on problem understanding	Ensure relevance, availability, quality of data	Be flexible with deployment and response options
Consider application criticality	Know your data and its evaloution	Be mindful of processing and computation costs
Link AI performance to business objectives	Avoid complexity	Develop cross-competences
Identify requirements for deployability	Deploy early	Develop tools and processes for recurrent tasks

Figure 3: Keys for successful AI applications in cybersecurity, organized by project stages.

5 The future of AI for cybersecurity

The constant increase in complexity, speed, and scale of cyberattacks has been a catalyst to the continuous adoption of AI in cybersecurity applications. This trend is bound to continue in the future, considering the growing AI hype and the availability of new AI technologies, like LLMs. Long-lasting benefits from this trend will only be reached through sensible applications of AI to relevant security problems. Nevertheless, current and future AI applications will face ethical, technical and regulatory challenges, which are new, not considered thus far, and may hinder their development.

5.1 Applications of LLMs to cybersecurity

Large Language Models (LLMs) offer an interesting potential to enhance numerous cybersecurity applications where AI has previously encountered challenges. Their capacity to comprehend tasks in natural language and yield outcomes interpretable by humans significantly lowers the entry barrier for non-AI cybersecurity experts to explore AI technologies. This widening scope for experimentation amplifies opportunities in cybersecurity applications, paving the way for increased creativity and innovation. Nonetheless, this initial exploration will necessitate subsequent expert involvement to transform experimental ideas into production-level solutions. For organizations having yet a low maturity in using AI for cybersecurity, LLMs are a game changer presenting an opportunity to catch-up with their delay.

LLMs may not yield substantial enhancements in mature applications such as threat detection or endpoint security, particularly in terms of detection capabilities. However, their proficiency in synthesizing intricate inputs and generating easily comprehensible outputs holds the potential to introduce greater transparency and provide explainability to opaque decision-making processes. This potential augmentation could elevate system usability by providing users with comprehensive explanations and better context regarding actions like blocking the opening of a file or the visit of a website. Additionally, they can support the enhancement of detection systems by offering insights into the reasons behind errors, including false positives and false negatives, thereby helping to fix these errors. LLMs can also potentially support more conventional security solutions used in threat detection and endpoint security, like the definition of detection rules for rule engines. Based on attack investigation, detection rules could be formulated by security analysts using natural language and fed to LLMs which would translate and generate formatted rules applicable by a rule engine. Overall, LLMs can increase the automation and the usability of already mature AI applications for cybersecurity.

Beyond these applications, LLMs have a great potential for improving security tasks involving human experts. Their ability to process and synthesize large amounts of heterogeneous data can be handy in threat intelligence and vulnerability management where the wealth of information to collect and analyse is humongous. LLMs can pull data from both internal and external information sources, combining them together and extracting relevant information about new threats affecting the organization and new vulnerabilities affecting its systems. In security analytics and security operation centres (SOC), LLMs can support security alert investigation, pulling information from different systems to contextualize security events. They not only streamline initial contextualization tasks but also facilitate deeper investigations through iterative interactions, allowing analysts to delve deeper and seek additional information as they get a better understanding of an attack. Acting as a bridge across various information sources and providing reasoned insights, LLMs can offer substantial assistance in human-involved security operations. This transformation is evident in new products aiding security teams[20] and showcasing the evolving integration of LLMs in enhancing security processes. LLMs also hold promise in suggesting responses and mitigation actions for the investigated threats. As the reliability of LLMs advances, we anticipate increased automation in

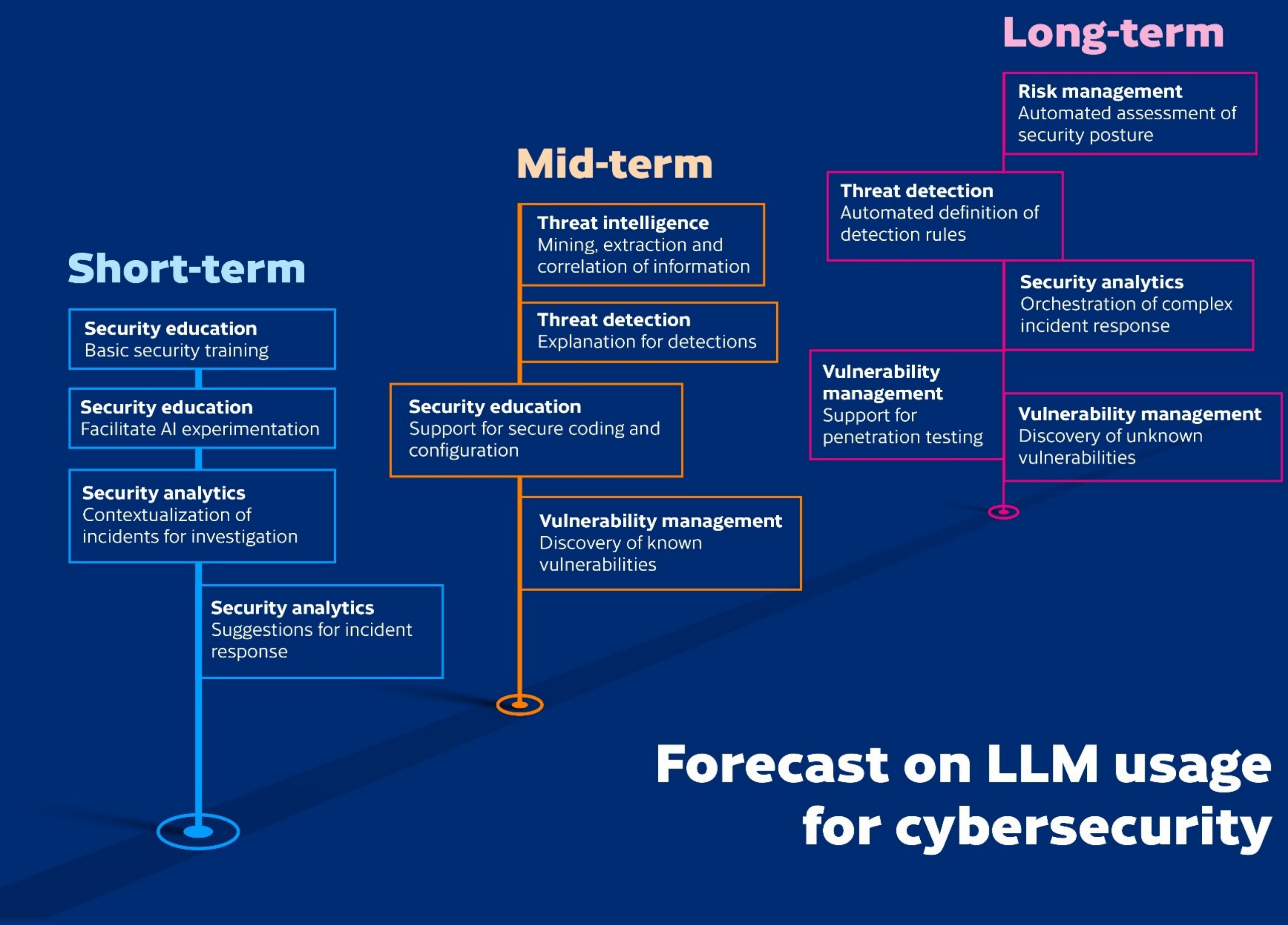
[20] *Microsoft Secutity Copilot.* https://www.microsoft.com/en-us/security/business/ai-machine-learning/microsoft-security-copilot

investigations and subsequent responses, reducing human intervention in these processes. This shift has the potential to mitigate the scarcity of security experts, empowering tier-1 security analysts to tackle only intricate tasks demanding their expertise.

A last LLM application holding great potential is toward security education. In a similar way as the SOC assistant previously discussed, LLMs can also be more active to train junior security analysts at alert investigation. Given a security event, they can suggest the information to pull and correlate, indicators to look at, and the response actions to perform. LLMs can also teach and enhance the application of security practices for regular developers and system administrators. Coding assistants can teach secure coding and ensure its principles are consistently applied during software development. For instance, some coding assistants include a security scanning feature to find and pre-emptively resolve potential security vulnerabilities in code[21]. LLMs can also be setup as assistants to configure complex systems, such as cloud services, in a secure manner. Even though cloud service providers implement high grade security in their infrastructure, a significant security threat comes from incorrect deployment configuration from their users, due to the complexity of defining proper parameters. LLMs can guide system administrators through this configuration process, interacting and providing information parameters as well as suggesting values based on the needs expressed to the LLM. Finally, LLMs can also help in training non-expert users of IT systems, teaching them about security threats and helping them to detect them better. For instance, LLMs are already used to simulate advanced spear-phishing attacks that would be launched by skilled attackers, increasing the ability of end users to detect and avoid falling victim for them.

Error! Reference source not found. depicts a tentative timeline for the adoption of LLMs in cybersecurity applications. Their first main application will likely be educational, enabling experimentation and teaching basic security practices at scale. Early applications will also encompass support for security analytics, enabling contextualisation of security events and recommending responses to them. As time goes by, LLMs will support more complex tasks related to threat intelligence and vulnerability management, like the discovery of know vulnerabilities in systems and codes the support to penetration testing exercise. In the long term, we can forecast that LLMs will hypothetically be able to deal with tasks requiring a higher level of expertise such as the discovery of new vulnerabilities (zero-day), the automation of security posture assessment and management, or the orchestration of responses to complex incidents.

[21] *Amazon CodeWhisperer.* https://aws.amazon.com/codewhisperer/



Short-term

Security education
Basic security training

Security education
Facilitate AI experimentation

Security analytics
Contextualization of incidents for investigation

Security analytics
Suggestions for incident response

Mid-term

Threat intelligence
Mining, extraction and correlation of information

Threat detection
Explanation for detections

Security education
Support for secure coding and configuration

Vulnerability management
Discovery of known vulnerabilities

Long-term

Risk management
Automated assessment of security posture

Threat detection
Automated definition of detection rules

Security analytics
Orchestration of complex incident response

Vulnerability management
Support for penetration testing

Vulnerability management
Discovery of unknown vulnerabilities

Forecast on LLM usage for cybersecurity

Figure 4

5.2 Risks, threats, and upcoming challenges

The increased adoption of AI for many applications has come with revelations regarding limitations about this technology. Issues related to bias (unfairness) of AI systems, lack of explainability but also security and privacy issues have attracted significant attention lately. AI systems are vulnerable to new security threats called adversarial ML attacks, which only affect these systems. Among these attacks, model poisoning and model evasion are concerning threats since they compromise the integrity of AI systems and the reliability of their predictions. Model poisoning is an attack whereby an adversary maliciously injects or modifies the training data or the training logic of an AI model to reduce the correctness and/or confidence of its predictions. Model evasion is an attack whereby an adversary maliciously constructs inputs to be sent to an AI system at inference time to receive incorrect predictions. Considering AI systems used for detection and prevention of security threats, cyberattacks are likely to leverage these attacks to fool AI systems and circumvent defences. Consequently, AI systems used for cybersecurity must consider these threats and implement defences against them to ensure they are as secure and resilient against attacks as non-AI based systems. Even though real-world adversarial ML attacks have been seldom observed yet, they are projected to become prevalent in the future.

Privacy issues are also an increasing concern since it has been shown that AI models, especially GenAI and LLM models, are prone to leak information about the data they are trained with. This issue either put constraints on the data that can be used to train AI models, or it requires additional rail guards to ensure no data leakage is possible. New AI technologies also have their issues. While conventional AI models tend be highly reliable at certain tasks, LLMs have more issues with the quality and reliability of their outputs. Hallucination is specific issue of LLMs where they generate incorrect or nonsensical information that seems plausible or coherent but is inaccurate or fabricated. This lack of reliability limits applications for automated decisions like threat detection and prevention, where high accuracy is required. Hallucination also raises concerns for applications such as support in security incident investigation or security training where users would have to rely on accurate information. While this issue remains, sensible applications of LLMs would only be limited to those involving human experts being able to quickly check the validity of the produced outputs.

The widespread adoption of AI in cybersecurity will come through the development of trustworthy AI systems, namely reliable, secure, privacy-preserving, fair and transparent. Reaching maturity in all these aspects will take a long time though. In the meantime, a sensible AI risk management process must be implemented together with AI governance to identify, understand, and mitigate the risk associated with using AI in cybersecurity. Navigating this path will be challenging yet inevitable, considering AI is likely to be the sole solution to counter AI-based cyberattacks, as this threat is growing[22]. AI is likely to be the new arms race between attacker and defender in the cybersecurity battle.

[22] Traficom. *The security threat of AI-enabled cyberattacks.* 2022

5.3 AI requirements, regulation, and standardization

Organizations leveraging AI in cybersecurity should establish robust AI governance, usage guidelines and a code of conduct. This will ensure ethical, safe, and efficient integration of AI within their cybersecurity systems, prioritizing benefits while mitigating potential harm.

The application of AI in cybersecurity has been open and subject to little constraints so far. AI can be used for virtually any cybersecurity application, provided that the data input to AI systems is collected and processed in compliance with data regulation, e.g., the EU GDPR[23]. Data regulation sets constraints on the use of personally identifiable information (PII) for AI. It also raises challenges in the development and experimentation of AI systems by limiting data access for experimentation. Customer data collection is authorized only if it serves a functionality in an existing product or service. However, data is required for the development and testing of AI systems before validating their utility and being able to offer a resulting service. Consequently, data intended to be used for AI, and which does not serve another purpose yet, requires explicit consent from customers to be collected for experimentation.

Regulation on the ethical and safe usage of AI is now being written and adopted. Even though some industrial sectors like finance and automobile have their own regulation on the use of AI, the cybersecurity industry falls under generic AI legislation. Some examples of AI regulation are the EU AI act[24], under provisional agreement and applicable within the European Union, or the AI bill of rights[25] being proposed in the USA. The European Union is the most advanced into regulating the use of AI, including in cybersecurity applications. The AI act will most likely ban some AI applications and set strong requirements for high-risk applications such as infrastructure management, law enforcement, health care or migration management. Under the provisional classification, most cybersecurity applications would fall under limited or low risk AI applications, which would only be required to transparency and provision of information to users. Nevertheless, it remains unclear if AI-based security solutions used to secure high-risk applications would be subject to the same strong requirements as the applications they secure. These requirements provisionally include risk management, human oversight, robustness, and security, among others. Security would be a sensible requirement for AI systems used in cybersecurity applications since the existence of an attacker attempting to circumvent and compromise defence systems is a given in this industry. It might be thus desirable to have an industry specific regulation setting the sensible requirements for the use of AI in cybersecurity.

The upcoming AI regulation is yet to be linked to comprehensive technical standards, which would equip AI practitioners with clear guidelines to meet the established legal requirements. While some initiatives exist to define technical

[23] Council of the European Union. *General Data Protection Regulation (GDPR).* 2016

[24] Council of the European Union. *Artificial Intelligence Act (AI act).* 2021

[25] US Office of Science and Technology Policy. *AI bill of right.* 2023

standards for AI, e.g., the ISO/IEC JTC 1/SC 42[26], these are still early work showing little progress. This is understandable since, taking the example of security requirements, there is no foolproof defence that exist yet against certain adversarial ML attacks. Nevertheless, guidelines and recommendations are available already to manage and mitigate the security risk of AI-based systems like the NIST AI Risk Management Framework[27], the MITRE ATLAS framework[28] or the guidelines for secure AI development[29] from the UK NCSC.

The further development of AI certification, based on technical standards, could streamline the compliance process for AI-based cybersecurity systems. However, this prospect appears distant for now. In the meantime, organizations leveraging AI in cybersecurity should establish robust AI governance, usage guidelines and a code of conduct. This will ensure ethical, safe, and efficient integration of AI within their cybersecurity systems, prioritizing benefits while mitigating potential harm.

[26] ISO/IEC. *ISO/IEC JTC 1/SC 42.* https://www.iso.org/committee/6794475.html

[27] NIST. *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* 2023

[28] MITRE *ATLAS.* https://atlas.mitre.org/

[29] UK National Cyber Security Center. *Guidelines for secure AI system development. 2023*

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

traficom.fi

ISSN (e-publication) 2669-8781
ISBN (e-publication) 978-952-311-917-8

TRAFICOM
Finnish Transport and Communications Agency