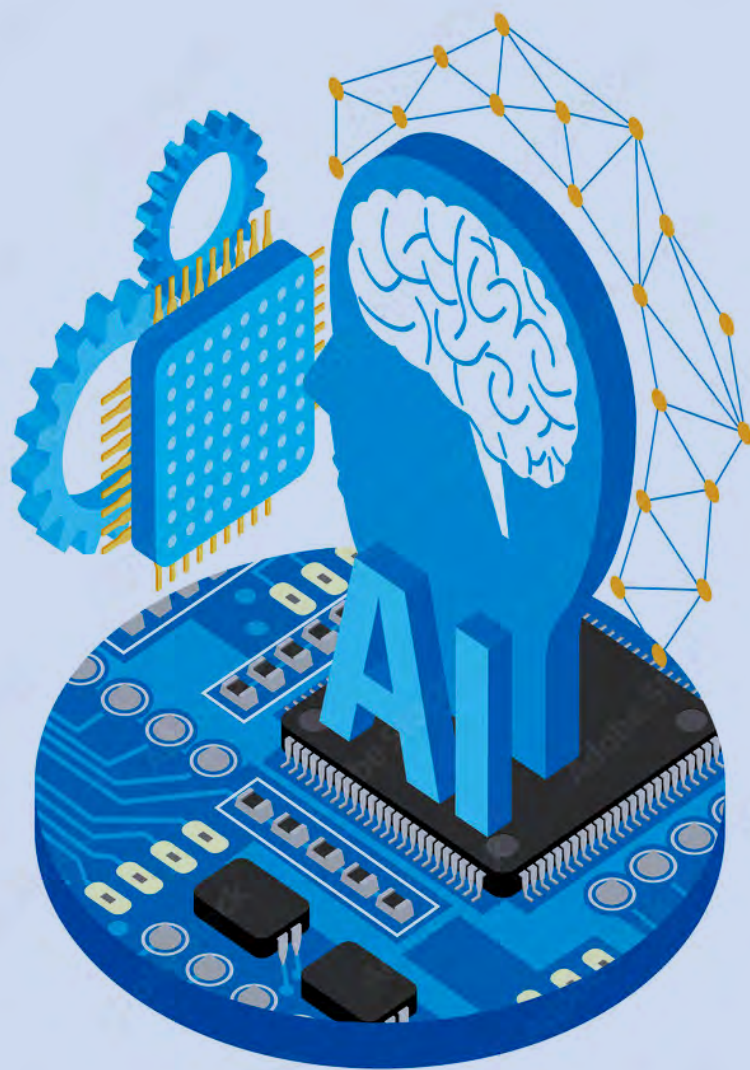


CCN-CERT
BP/30



Approach to Artificial Intelligence and Cybersecurity

BEST PRACTICE REPORT

OCTOBER 2023

ccn-cert
centro criptológico nacional

CCN
centro criptológico nacional

Edited by:



Paseo de la Castellana 109, 28046 Madrid

© National Cryptologic Centre, 2023

Authors: Carlos Galán Cordero and Carlos M. Galán Pascual

Date of issue: October 2023

LIMITATION OF LIABILITY

This document is provided in accordance with the terms contained herein, expressly rejecting any type of implicit guarantee that may be related to it. Under no circumstances can the National Cryptologic Centre be held responsible for direct, indirect, fortuitous or extraordinary damage derived from the use of the information and software indicated, even when warned of such a possibility.

LEGAL NOTICE

The reproduction of all or part of this document by any means or process, including reprography and computer processing, and the distribution of copies by public rental or loan, is strictly prohibited without the written authorisation of the National Cryptologic Centre, subject to the penalties established by law.

Content

Purpose of the document	5
1. Introduction	6
1.1 Definition of artificial intelligence (AI) and cybersecurity	6
1.2 Brief history of AI in cybersecurity	7
1.3 Current importance and relevance of the issue	9
2. Fundamentals of Artificial Intelligence	13
2.1 <i>Machine Learning (ML)</i>	14
2.2 <i>Deep Learning (DL)</i>	16
2.3 Classification algorithms	18
2.4 Generative AI	20
3. AI applications in Cybersecurity	23
3.1 Threat detection and behavioural analysis	24
3.2 Automatic response and orchestration	33
3.3 Threat prediction	35
3.4 Biometric identification and authentication	37
3.5 Vulnerability scanning and automated pentesting	40
3.6 Defence against automated adversaries	42
Defence with Artificial Intelligence	43
3.7 Generative AI and Cybersecurity	46
4. Study scenarios	52
4.1 Modern threat detection and response systems	53
Challenges	55
Lessons learned	55
Adoption and adaptation	56
Evolving threats in response to modern systems	57
4.2 Successful implementations of AI in cybersecurity	59
4.3 Failures and lessons learned	62

Content

5. Challenges and limitations of AI in Cybersecurity	63
5.1 Adversarial attacks against AI models	64
5.2 Overreliance on automated solutions	67
5.3 False positives and false negatives	69
5.4 Privacy and ethics in the application of AI	71
6. Future of AI in Cybersecurity	77
6.1 Emerging trends	78
Autonomous cyber self-defence	78
Federated learning	79
Explainable AI systems (XAI)	80
Adoption of Blockchain for security	82
AI models based on user behaviour	83
Quantum AI	84
Human-machine collaboration	86
AI at the edge (Edge AI)	88
6.2 Ongoing investigations	90
6.3 Potential impact on industry and society	92
7. Recommendations and best practices	94
7.1 Integration of cybersecurity teams and AI teams	95
7.2 Further training	97
7.3 Designing robust and resilient systems	100
8. Conclusion	102
8.1 Final reflections on the current state and future of AI in cybersecurity	102
8.2 Subsequent actions and recommendations for future research	104

Purpose of the document

The purpose of this document is, as its title suggests, to take a closer look at the area of work of two disciplines, Artificial Intelligence and Cybersecurity, which, with origins clearly separated in time, have seen how their areas of expertise have been coming closer over the last few years to what currently constitutes a new practical activity, which brings together the knowledge and previous experience of both: **Artificial Intelligence applied to Cybersecurity**; what we could call **Artificial Intelligence CyberSecurity, AICS**.

Its introductory nature, which makes it more of a survey than a scientific treatise, is intended to make it easier for this document to find its target readers: professionals or scholars of information systems, their applications and challenges, with a special focus on the managers of organisations (public or private), their management departments, including technical and legal areas and, of course, cybersecurity teams.

With this document, we hope to provide a first reading — which can be complemented with other more specific readings such as those referenced here — and with those with which the coming years will bring news of the new and undoubtedly surprising realities that will materialise, together, in cybersecurity and artificial intelligence.

DISCLAIMER:

In this text, for ease of understanding, equipment, instruments or commercial material from different entities are identified. Such identification does not imply the recommendation or approval by the National Cryptologic Centre, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose indicated in each case.

1. Introduction

1.1 Definition of artificial intelligence (AI) and cybersecurity

Although there is no consensus that can provide a universal definition, we can say that **Artificial Intelligence (AI)** is a subfield of computer science that aims to develop systems capable of performing tasks that, so far, require human intelligence; tasks that may include **learning** (acquiring information and rules for using the information), **reasoning** (using rules to reach approximate or definitive conclusions) and **self-correction**¹.

For example, the current Proposal for a European Regulation on Artificial Intelligence², in trilogues³ at the time of writing, states that a definition of AI should be based on the main functional characteristics of the software, and in particular its ability to generate, in relation to a specific set of human-defined objectives, content, predictions, recommendations, decisions or other output information that influences the

1 ENISA, in its document ARTIFICIAL INTELLIGENCE AND CYBERSECURITY RESEARCH (June, 2023) points out that: "There is no common definition of AI (European Commission. Joint Research Centre. AI watch: defining Artificial Intelligence: towards an operational definition and taxonomy of artificial intelligence. Publications Office, 2020. doi:10.2760/382730. (<https://data.europa.eu/doi/10.2760/382730>)).

Although there is no common definition, the definitions reviewed show some commonalities (cf. CCI5) which can be considered as the main characteristics of AI: (i) perception of the environment, including consideration of the complexity of the real world; (ii) information processing (collecting and interpreting inputs (in the form of data); (iii) decision-making (including reasoning and learning): taking actions, performing tasks (including adapting and reacting to changes in the environment) with a certain level of autonomy; (iv) achieving specific goals".

2 Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL ESTABLISHING HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE LAW) AND AMENDING CERTAIN LEGISLATIVE ACTS OF THE UNION (Brussels, 21.4.2021).

3 The so-called *trilogues* are informal groups that are set up for each legislative proposal and are composed of three members: one from the Commission, one from the Parliament and one from the Council presidency (<https://spanish-presidency.consilium.europa.eu/es/noticias/los-trilogos/>)

1. Introducción

environment with which the system interacts, whether in a physical or digital dimension; adding that a definition of "AI system" should be complemented by a list of the specific techniques and strategies used in its development.

The term **cyber security** essentially refers to the practice of protecting systems, networks and programmes from cyber-attacks. These cyber-attacks are often aimed at accessing, altering or destroying valuable or confidential information, extorting money from users, or disrupting processes and services.

1.2 Brief history of AI in cyber-security

The relationship between AI and cybersecurity has been consolidated over the years. Initially, cybersecurity systems relied mainly on predefined signatures and rules to detect threats. However, with the rise and evolution of cyber threats, the need for more advanced and adaptive systems has become evident.

The first attempts to use **machine learning techniques for intrusion detection** were made in the 1990s⁴, but it was not until the 2010s, thanks to advances in **deep learning** technologies and the availability of large datasets, that AI began to play a significant role in cyber security, offering more efficient and accurate solutions to constantly evolving threats.

⁴ Indeed, during the 1990s, there was a growing interest in the use of machine learning techniques for intrusion detection, recognising that traditional signature-based techniques would not be sufficient to detect new or modified attacks, known as zero-day attacks. In response to this, behavioural and machine learning based techniques were explored, such as: **IDES (Intrusion Detection Expert System)**: developed in the late 1980s and early 1990s by SRI International, IDES was one of the first behavioural-based intrusion detection systems. It used statistical techniques to establish a profile of a user's or system's "normal" activity and then alerted to significant deviations from this behaviour. **ADAM system**: In 1995, a system called ADAM (Automated Detection, Analysis, and Measurement) was proposed by Lee and Stolfo. This system used clustering algorithms to detect anomalous activity in auditing systems. **Neural Networks**: During the 1990s, neural networks were also explored as a tool for intrusion detection. For example, in 1998, Ghosh, Schwartzbard and Schatz proposed using neural networks to detect anomalous behaviour in network connections. **LERAD system**: In the late 1990s, Barbara and Wu developed LERAD (Learning Rules for Anomaly Detection), which was a machine learning-based technique for detecting anomalous activity in audit datasets. **KDD Cup 1999 Dataset**: Perhaps one of the most influential efforts in the area of machine learning-based intrusion detection was the 1999 KDD Cup. It provided a dataset containing a variety of simulated intrusions in a network environment. This dataset has been widely used by the research community to evaluate and compare different intrusion detection methods.

1. Introduction

PERIODS	ACTIVITIES
Beginnings (1960s and 1970s):	- During the early stages of computer development, the idea of automating security was not a priority. Computer systems were not as widely interconnected as they are today, and the very concept of cyber security was in its infancy. - Early approaches to AI during this period focused on topics such as natural language processing and so-called expert systems, but not on cybersecurity.
Birth of cybersecurity (1980s):	- With the rise of personal computing and the development of the first networks, the first cyber threats emerged. - Security tools relied on known signatures and patterns to detect threats, which could be considered a primitive form of machine learning, although AI as such had not yet been significantly integrated into cyber security.
Early approaches to AI in cybersecurity (1990s):	- Intrusion detection systems (IDS) started to incorporate basic machine learning techniques to identify anomalous traffic patterns. - Research and academic work began to emerge exploring the use of classification algorithms to improve malware and attack detection.
Big Data explosion and AI breakthrough (2000s):	- With the proliferation of the Internet and the emergence of more advanced threats, large data sets (logs, network traffic, etc.) became a crucial source for cyber security. - AI techniques started to be used to analyse these large volumes of data for suspicious patterns or anomalous behaviour.
Deep learning and cybersecurity (2010s):	- The rise of deep learning (especially convolutional and recurrent neural networks) found applications in cybersecurity, such as advanced malware detection based on features and behaviours rather than signatures. - Automatic response and orchestration systems were introduced⁵, using AI to make real-time decisions in the face of identified threats. - However, the concept of "adversarial attacks" also emerged⁶ against AI models, in which attackers aim to trick or confuse machine learning models.
Present and future (2020s onwards):	- Today, AI is an essential tool in cybersecurity, not only for detection and response, but also for threat prediction. - As cyber threats become more sophisticated, so does the need for more advanced and robust AI solutions, including generative AI applied to cyber security. - Concerns about ethics, privacy and liability in AI applied to cybersecurity are also coming to the fore, and these areas are likely to see significant development in the coming years, as evidenced by the European Regulation on AI discussed above.

⁵ Security Orchestration, Automation, and Response (SOAR) systems.

⁶ Adversarial attacks.

1.3 Current importance and relevance of the issue

Today, cybersecurity is not only a technical issue, but a global concern that affects public institutions, businesses and individuals. With the digitisation of many services and the creation of connected critical infrastructures, the need to secure these systems is paramount; all due to the reality posed by the following characteristics of the **digital transformation** of society:

- ▶ **Exponential data growth:** we can say that we live in the era of *Big Data*. Petabytes of data are generated every day, and in this vast ocean of information, detecting malicious patterns or anomalous behaviour is an extremely complicated task. AI, through advanced algorithms, can analyse large volumes of data in real time, identifying potential threats that would be virtually impossible to detect by manual or traditional methods.
- ▶ **Evolving threats:** Cyber threats are not static, but constantly evolving. Threat actors are constantly developing new techniques and tactics to circumvent security systems. AI enables adaptability and continuous learning, which means it can "learn" from new threats and adapt accordingly, providing an additional layer of protection.
- ▶ **Automation and rapid response:** In the face of a cyberattack, the response must be immediate. AI can automate the actions to be taken, such as isolating a compromised device or blocking suspicious access, much faster than a human could. This reduces exposure time and potentially minimises damage.
- ▶ **Complex pattern recognition:** AI is exceptionally useful in identifying patterns in large data sets. In the context of cybersecurity, this means it can identify malicious behaviour based on subtle patterns that might go unnoticed by traditional systems.

Today, cybersecurity is not only a technical issue, but a global concern that affects public institutions, businesses and individuals

1. Introduction

- ▶ **Shortage of cybersecurity professionals:** There is a growing demand for cybersecurity professionals. AI can help fill this gap, taking over tasks that require real-time analysis and response and allowing human experts to focus on more strategic tasks.
- ▶ **Economic and societal cost:** Security breaches can result in huge economic losses, reputational damage and, in the case of critical infrastructure, can even endanger human lives. AI applied to cybersecurity not only protects an organisation's assets and data, but can also play a crucial role in protecting society as a whole.
- ▶ **Ethical and regulatory challenges:** As AI becomes more deeply embedded in cybersecurity (and life in general), new ethical and regulatory questions arise. Who is responsible if an AI makes a wrong decision? How do we ensure that AI acts in a fair and non-discriminatory way? These are crucial questions that highlight the importance of considering AI not only from a technical perspective, but also from an ethical and social one⁷.

All this without forgetting that, as ENISA⁸, points out, there are multiple **actors and threat actors** already using AI techniques to develop their actions, among them:

- ▶ **Cybercriminals**, whose primary motivation is profit, will tend to use AI as a tool for attacks, but also to exploit vulnerabilities in existing AI systems. For example, they could try to attack AI chatbots to steal credit card information or other data. They could also launch a ransomware attack against AI-based systems being used for supply chain management and warehousing.

AI applied to cybersecurity not only protects an organisation's assets and data, but can also play a crucial role in protecting society as a whole

⁷ For more information on this, see "Certification as a control mechanism for artificial intelligence in Europe". (C. Galán. Spanish Institute for Strategic Studies. 2019). https://www.ieee.es/Galerias/fichero/docs_opinion/2019/DIEEO46_2019CARGAL-InteligenciaArtificial.pdf

⁸ ENISA- AI Cybersecurity Challenges (2021).

1. Introduction

- ▶ **Insiders**, including employees and contractors who have access to an organisation's networks, can perform malicious actions, whether malicious or unintentional. Malicious intruders could, for example, attempt to subtract or sabotage the training dataset used by the company's AI systems. Conversely, others might inadvertently corrupt the training dataset accidentally.
- ▶ **Nation-states** or **state-sponsored actors** who, in addition to developing ways to exploit AI systems to attack other countries (including industries and critical infrastructure), as well as using AI systems to defend their own networks, will actively seek out vulnerabilities in AI systems that they can exploit. This may be a means to cause harm to another country or to gather information.
- ▶ **Terrorists**, who seek to cause physical harm or even loss of life, e.g. by cyber-attacking driverless vehicles to use them as a weapon.
- ▶ **Hacktivists**, most of whom tend to be ideologically motivated, may also try to hack into AI systems to demonstrate that it is something that can be done. More and more groups are concerned about the potential dangers of AI, and it is not inconceivable that they could hack into an AI system to gain publicity.
- ▶ There are also **unsophisticated actors**, such as script kiddies, who may be criminally or ideologically motivated. These are usually unskilled individuals who use pre-written scripts or programmes to attack systems, as they lack the knowledge to code them themselves.
- ▶ In addition to these traditional threat actors, it seems increasingly necessary to include **competitors** as threat actors as well, as some companies may be beginning to make clear their intention to attack rivals in order to gain market share.

Malicious intruders could, for example, attempt to subtract or sabotage the training dataset used by the company's AI systems

This shapes a broad and extremely sensitive threat landscape⁹.

⁹ ENISA (2021), op. cit.

1. Introduction



AI THREAT TAXONOMY

Nefarious activity/abuse (NAA): "intentional actions directed at ICT systems, infrastructures and networks through harmful acts with the aim of subtracting, altering or destroying a specific target".

Eavesdropping/Interception/Hijacking (EIH): "actions aimed at eavesdropping, interrupting or gaining control of a third-party communication without consent".

Physical Attacks (PA): "actions aimed at destroying, exposing, altering, disabling, subtracting or gaining unauthorised access to physical assets such as infrastructure, hardware or interconnections".

Unintentional Damages (UD): unintentional actions that cause "destruction, damage or injury to property or persons and result in a failure or reduction of utility".

Failures/Malfunctions (FM): means "partial or total malfunctioning of an asset (hardware or software)".

Outages (OUT): "unexpected interruptions of service or decrease in quality below a required level".

Disasters (DIS): "sudden accident or natural catastrophe causing great damage or loss of life".

Legal (LEG): "legal actions by third parties (contracting or non-contracting), for the purpose of prohibiting actions or compensating losses on the basis of applicable law".

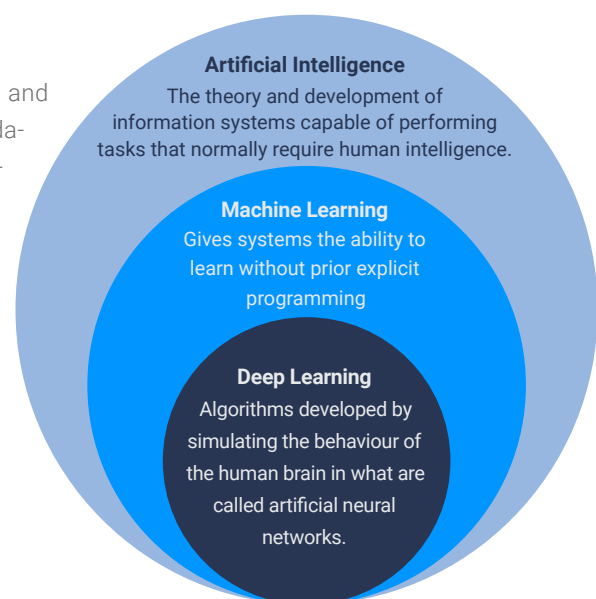
2. Fundamentals of Artificial Intelligence

Artificial intelligence (AI) is a broad field of study that encompasses diverse techniques and technologies. From the early days of computer science to the present day, AI has evolved from a theoretical concept to a practical tool that has applications in countless domains, including cybersecurity. In this cybersecurity context, AI acts as a **force multiplier**, offering advanced capabilities that go beyond what is possible with traditional methods.

To understand how AI benefits cybersecurity, it is essential to become familiar with the specific techniques and technologies that are being applied. These techniques range from machine learning and its subdomains to fuzzy logic, neural networks or the more recent generative AI. Each of these techniques has its own characteristics, advantages, challenges and applications within cyber security, and together they make up an arsenal that organisations can use to defend against growing and evolving cyber threats.

In this section, we will explore some of the main AI techniques and technologies being used in cybersecurity, providing a solid foundation for understanding how AI is revolutionising the way we protect our systems and data.

From a descriptive point of view, it is useful to place the different **AI models** in a context that allows a better understanding of their techniques and characteristics. The following figure develops this idea.



2.1 Machine Learning (ML)

Machine learning is a method of data analysis that automates the building of analytical models. Instead of being explicitly programmed to perform a task, machines are "trained" by using large data sets and running algorithms that give them the ability to learn how to perform the task.

Machine learning techniques can be classified into the following **types**:

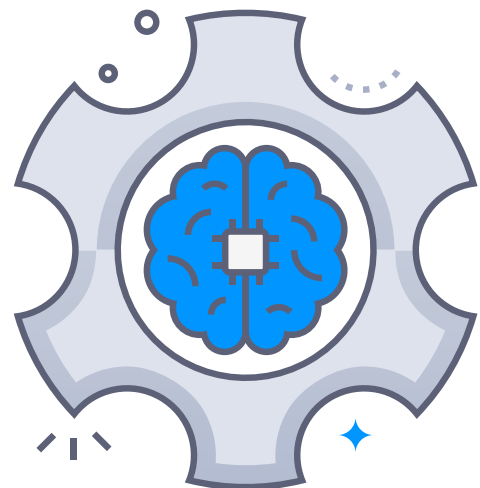
Supervised learning	This is the most common technique. In it, the model is trained using a labelled dataset, which means that each example in the dataset is accompanied by the "correct answer". Once trained, the model can start making predictions or decisions without human intervention. Examples of applications are the classification of e-mails as "spam" or "non-spam" or the prediction of house prices based on characteristics such as size and location.
Unsupervised learning	In this case, the model is trained on an unlabelled dataset, and its goal is to discover hidden structures in the data. Common techniques include clustering and dimensionality reduction. An example could be segmenting customers into different groups, based on their buying behaviour.
Reinforcement learning	It is a type of learning where an agent learns how to behave in a given environment by performing certain actions and receiving rewards or penalties in response. It is often used in robotics, gaming and navigation.

2. Fundamentals of Artificial Intelligence

Within these types, specific techniques have been developed, such as **decision trees**, *support vector machines*, the *Naive Bayes'* classifier, the so-called **K-means clusterig**, the **Hidden Markov Model** or **genetic algorithms**, which we will examine briefly below.

In cybersecurity, machine learning (ML) can be useful in applications such as threat detection (as it can analyse large volumes of data to identify patterns of anomalous or suspicious behaviour, enabling faster and more accurate threat detection), **code analysis** (since by training AI models on malware datasets, ML can help identify and classify new variants, even if they have not been observed previously), **phishing and fraud detection** (by having ML models analyse the characteristics of websites and emails, to determine whether they are malicious or legitimate).

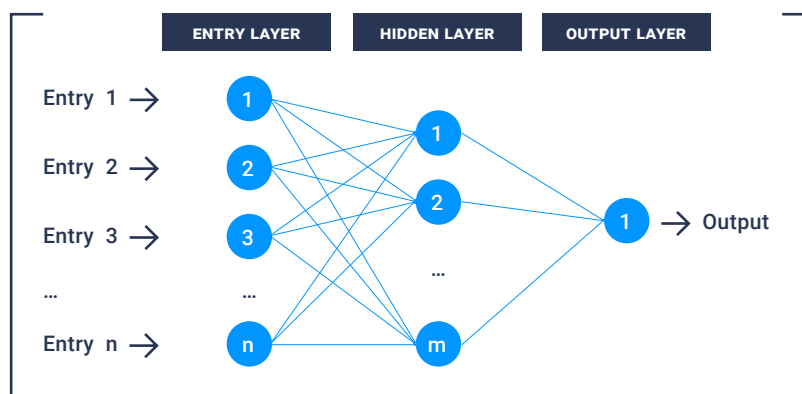
However, building ML-based cybersecurity systems has demands and challenges, such as **having good quality data**, as machine learning is only as good as the data it is trained on, so if the data is biased or of poor quality, the resulting models will also be poor; or so-called **over-fitting**, as a model may be too complex and "memorise" the training dataset, rather than generalising on new and unseen data; or so-called **adversarial attacks**, whereby an attacker may try to fool an ML model by presenting data specifically designed to confuse it and generate erroneous decisions.



2.2 Deep Learning (DL)

So-called **deep learning** is a machine learning technique that uses **neural networks**¹⁰ with three or more layers.

Artificial neural networks (ANNs) are computational models inspired by the functioning of neurons in the human brain. They are composed of units or nodes called "neurons" that are organised in layers: input, hidden(s) and output. Each connection between neurons has an associated weight, which is adjusted during the training process.



These networks can learn patterns and data representations at increasing levels of complexity, enabling them to perform tasks that were considered too complex for traditional machine learning algorithms.

We can divide these technologies into the following **types**:

¹⁰ Several training datasets are currently available. The most widely used are: KDD Cup99; DEFCON; CTU-13; Spam Base; SMS Spam Collection; CICIDS2017; CICAndMal2017; Android Validation; IoT-23 data set; each with special characteristics to address specific issues.

2. Fundamentals of Artificial Intelligence

Convolutional Neural Networks (CNNs)	Especially useful for image and video related tasks, as they can efficiently identify and extract image features.
Recurrent Neural Networks (RNNs)	They are particularly effective for working with sequences of data, such as time series or text, because of their ability to "remember" previous information from the sequence.
Long Short-Term Memory Neural Networks (LSTM)	It is a variant of RNNs, designed to address the problem of the fading gradient¹¹ and retain information in the long term. Like RNNs, they are used for sequence analysis, although with higher accuracy on longer sequences.
Generative Adversarial Networks (GANs)	This is a type of model that uses two networks (a generative and a discriminative network) that work together to generate authentic-looking data.

In relation to cybersecurity, such DL techniques can be used to:

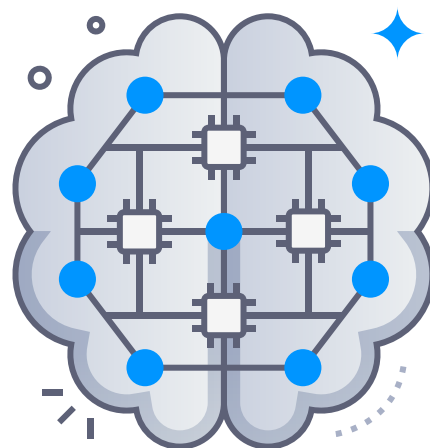
- ▶ **Malicious code detection:** since neural networks can be trained to identify malicious software based on patterns and features extracted from files. For example, a CNN could analyse the content of a binary file and determine whether or not it has malware characteristics.
- ▶ **Network traffic analysis:** RNNs, given their ability to analyse sequences, can be useful for inspecting network traffic for anomalous or malicious patterns.
- ▶ **Phishing detection:** CNNs can be trained to analyse the visual content of websites and determine whether they mimic or replicate legitimate sites, with the purpose of deceiving users.

¹¹ The *vanishing gradient* problem is an obstacle that arises in the training of traditional artificial neural networks, especially recurrent neural networks (RNNs). It refers to the tendency of gradients to decrease exponentially as they back-propagate across layers and over time in RNNs. When gradients approach zero, it implies that the weights of the neurons are not effectively updated during the training process, leading to inefficient or stagnant training. Long Short-Term Memory (LSTM) networks were specifically designed to address this problem, as well as the related problem of the exploding gradient, where gradients can grow exponentially. LSTMs achieve this through their cellular structure, which includes entry, forgetting and exit gates. These gates, combined with a state cell, allow LSTMs to retain or discard information over long sequences, ensuring that the gradient is maintained and propagates properly through the network without fading or exploding. This special design allows LSTMs to learn long-term dependencies in the data, making them particularly useful for tasks such as machine translation, natural language processing, time series prediction and more, where it is crucial to remember information from earlier parts of a sequence.

2. Fundamentals of Artificial Intelligence

- ▶ **Generating malware samples for testing:** GANs can be used to create malware samples. Thus, the generative component would create samples while the discriminative component would evaluate their authenticity, which can help improve the robustness of certain tools.
- ▶ **Behavioural analysis:** Neural networks can learn patterns of user or system behaviour and detect deviations from these patterns, which could indicate malicious activity or compromise.

However, as in the previous case, the use of such techniques requires consideration of certain issues, such as the **need for large datasets**, since DL often requires large amounts of labelled data for training; or **training time**, since training deep learning models can be computationally intensive; or **interpretability**, since, unlike other algorithms, deep neural networks often act as "black boxes", meaning that their decisions are not easily interpretable by humans.



2.3 Classification algorithms

Classification algorithms are a branch of supervised machine learning. Their main goal is to take an input (or instance) and assign it to one of predefined classes. In the context of cybersecurity, these classes could be, for example, "malicious" or "benign".

Thus, a classification algorithm aims to learn, from a training data set, how to categorise unseen entries into one or more categories or classes.

We can classify sorting algorithms into the following **types**, noting their applications in cybersecurity.

2. Fundamentals of Artificial Intelligence

TYPE	FEATURES	APPLICATION IN CYBER SECURITY
Logistic regression	A statistical method for analysing data sets in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (yes/no, 1/0, true/false).	Determine whether an activity is malicious or not based on several characteristics.
Support Vector Machines (SVM)	These algorithms seek to find the hyperplane that best divides a dataset into classes.	Classification of emails as spam or non-spam, feature-based malware detection.
Decision trees and random forests	<i>Decision trees</i> divide the data set into subsets based on the value of the input features. <i>Random forests</i> are a collection of decision trees that participate together to provide a final prediction.	Intrusion detection based on characteristics such as protocol type, IP address, duration, etc.
Neural networks	Already studied before. These are structures inspired by the human brain that can learn complex patterns.	Malware detection, user behaviour analysis, anomaly detection, among others.
K-Nearest Neighbours (K-NN)	This is an algorithm that ranks an entry based on how its k nearest neighbours are ranked.	Detection of malicious activity based on its similarity to known behaviour.
Naive Bayes	It is based on Bayes' theorem and assumes independence between features. It is especially useful when the dimension of the data is high.	Classification of emails as spam or legitimate, text analysis to identify malicious communications.

As always, the practical application of these techniques requires taking into consideration certain issues, such as **data quality** (since, as we have pointed out before, a model is only as good as the data it is trained on), **class imbalance** (since in many cybersecurity scenarios, such as malware detection, the majority of instances may be benign, and only a small percentage malicious, which can lead to model bias if not handled properly), or interpretability (since it is important to understand the reasons behind a model's decisions, especially in a critical context such as cybersecurity), or interpretability (since it is important to understand the reasons behind a model's decisions, especially in a critical context such as cybersecurity, which can lead to model bias if not handled properly), or **interpretability** (since it is important to understand the reasons behind a model's decisions, especially in a critical context such as cybersecurity), or **adaptability** (since if threat actors are constantly evolving their tactics, techniques and procedures, it is essential that classification models can quickly adapt to new threats).

2.4 Generative AI

Generative artificial intelligence (AI) refers to a subset of machine learning techniques that aim to generate new data that is similar, but not identical, to the data it was trained on. Unlike discriminative machine learning techniques, which learn to differentiate between different types of data (e.g., classifying emails as "spam" or "non-spam"), generative techniques aim to produce data that resembles the input data.

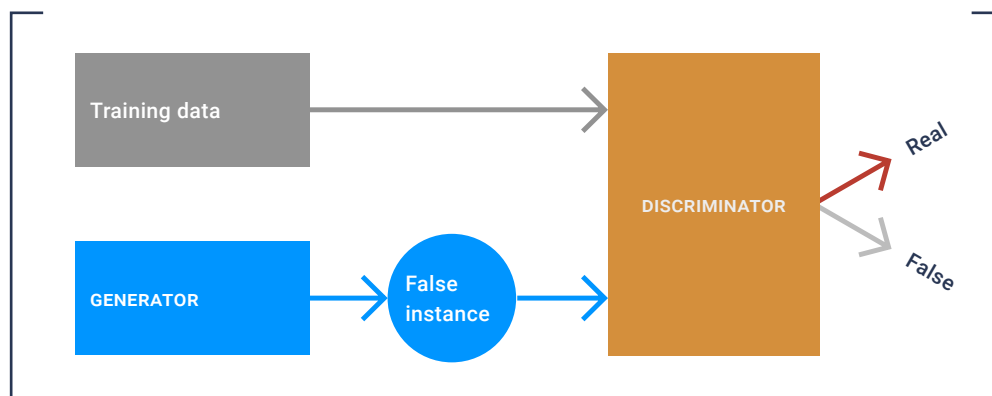
One of the most popular and effective approaches within **generative AI are Generative Adversarial Networks (GANs)**, which we have already mentioned. These networks are based on two models that work together:

1. A **Generator** element, whose purpose is to generate data. Initially, it produces data randomly, but with time and feedback from the discriminator, it improves its ability to generate data that resembles real data.
2. A **Discriminator** element, which examines the data and tries to distinguish between real data and data generated by the generator element. It provides feedback to the generator on how well (or poorly) it is doing.

Thus, the generator element tries to produce increasingly convincing false data, while the discriminator element constantly improves its ability to detect this false data. With sufficient training, the generator can eventually produce data that is almost indistinguishable from real data for humans and machines.

Generative artificial intelligence (AI) refers to a subset of machine learning techniques that aim to generate new data that is similar, but not identical, to the data it was trained on

2. Fundamentals of Artificial Intelligence



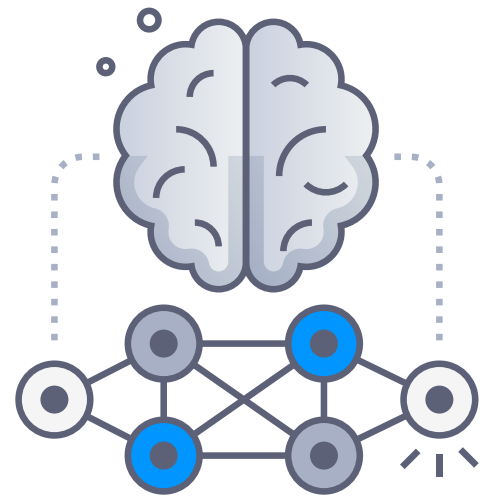
The applications of generative AI are multiple: Image Processing (creation of artistic images, improvement of image resolution, generation of images of objects or scenes that do not exist in reality, etc.); **Audio** (creation of music, sound effects or synthetic voices); **Text** (generation of coherent texts, stories, poetry, etc.); **Video** (creation of synthetic videos or modification of existing videos, such as deepfakes, etc.); Design and modelling (generation of data sets for training when real data is not available or insufficient); Design and modelling (generation of data sets for training when real data is not available or insufficient).); Video (creation of synthetic videos or modification of existing videos, such as deepfakes); **Synthetic data** (generation of datasets for training when no or insufficient real data is available); **Design and modelling** (generation of designs for products, architecture or 3D modelling, etc.).

In cybersecurity, generative AI also finds several applications, such as the following:

- ▶ **Creating malware samples:** GANs can be trained to generate malware variants that evade detection by traditional security solutions. While this may seem dangerous, security researchers can use this technique in controlled environments to improve the robustness of detection systems.
- ▶ **Strengthening detection systems:** By generating malware or malicious network traffic, security teams can use these samples to train and improve their detection systems. In essence, it pits AI against itself to improve detection.

2. Fundamentals of Artificial Intelligence

- ▶ **Network traffic simulation:** Generative AI can simulate normal network traffic or attack traffic to test the robustness of a network or system. This is especially useful in training cyber defenders and testing security systems.
- ▶ **Creation of fake domains:** In the field of zero-day threat protection, GANs can be used to generate fake domains that look like real malicious domains. This helps security systems predict and block domains that could be used in future attacks.
- ▶ **Adversarial attacks:** As mentioned above, adversarial attacks involve the introduction of small perturbations in data to fool machine learning models. GANs can be used to generate these perturbations efficiently, which can help defenders understand and mitigate these attacks.
- ▶ **Phishing and malicious content generation:** GANs can be trained to generate emails or web pages that mimic legitimate ones, making them potentially useful tools for phishing attacks. However, they can also be used in defence, generating phishing samples to train detection systems.



3. AI applications in Cybersecurity

As is well known, cyber security is a never-ending struggle between attackers and defenders. While attackers seek new vulnerabilities and ways to compromise systems, defenders seek to anticipate (prevention), detect (detection) and respond to these attacks (response).

AI, with its ability to process large amounts of data at extraordinary speed and learn from it, offers significant solutions to (current and emerging) cybersecurity challenges.

Traditionally, some of the key applications of AI in cyber security are shown in the table below:

Threat detection and response	AI-based systems can analyse patterns in network traffic or user behaviour to identify anomalies or suspicious activity. Once detected, AI can act quickly, often faster than a human team, to mitigate or neutralise the threat.
Predictive analytics	AI can use historical data to predict future threats or vulnerabilities, allowing organisations to proactively prepare and protect themselves ¹² .
Authentication and identity management	AI can employ advanced biometrics, user behaviour and other factors to authenticate individuals with high accuracy, reducing the risk of unauthorised access.
Phishing protection	By analysing the content, images and patterns of texts or documents (e.g. emails), AI can identify phishing attempts with high accuracy, protecting users from potential scams.
Optimisation of security settings	AI can evaluate security configurations and policies to identify possible weaknesses and propose improvements.

¹² More information on what has been given of proactive analysis, in what is called the Diamond Model, can be found in CCN-STIC Guide 425 Intelligence Cycle and Intrusion Analysis. <https://www.ccn-cert.cni.es/series-ccn-stic/guias-de-acceso-publico-ccn-stic/1093-ccn-stic-425-ciclo-de-inteligencia-y-analisis-de-intrusiones/file.html>

3. AI applications in Cybersecurity

As we move into this section, we will explore in detail how AI fits into these and other cybersecurity domains, its potential and, of course, the ethical and privacy considerations associated with its use.

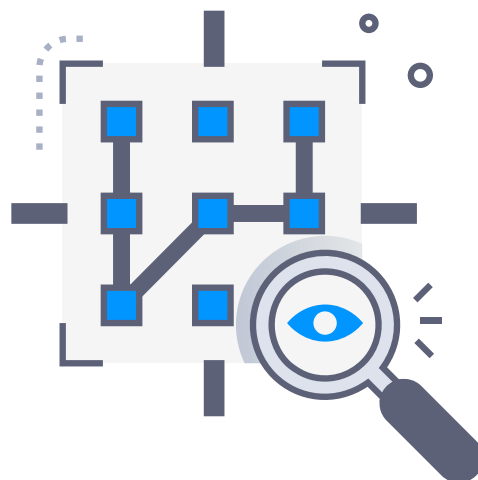
3.1 Threat detection and behavioural analysis

Threat detection and behavioural analysis are essential to identify and respond to cyberattacks in real time. With the implementation of AI in these fields, cyber security has seen a marked improvement in detection efficiency and accuracy.

The amount of data that organisations (public or private) process on a daily basis is immense. Manually detecting threats in such volumes is virtually impossible. Modern cyber-attacks often employ stealthy tactics, such as lateral movement and low-profile persistence, making them difficult to detect with traditional methods.

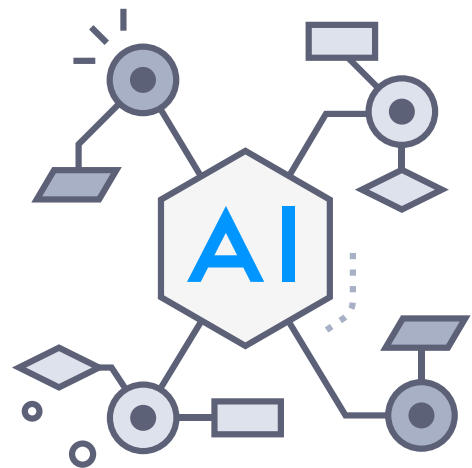
Thus, instead of relying only on known malware signatures, the AI focuses on **patterns of anomalous behaviour**. This makes it possible to detect previously unknown threats or malware variants that have been slightly modified. By analysing user and system behaviour, AI can identify unusual activity, such as accessing files at odd hours or the unusual transfer of large amounts of data.

Behavioural threat detection has experienced a rapid growth in popularity and adoption, and a number of tools and systems, both commercial and open source, specialised in this approach have been emerging. Some of the most popular tools are listed below:



3. AI applications in Cybersecurity

- ▶ **Darktrace**¹³: Darktrace uses machine learning and AI algorithms to detect, respond to and mitigate threats in real-time based on patterns of anomalous behaviour. The tool is known for its "*Enterprise Immune System*", which learns and establishes what can be understood as a "business as usual" situation in the network and then identifies deviations from this norm.
- ▶ **Vectra**¹⁴: Vectra offers real-time threat detection using AI. It focuses on detecting malicious behaviour within network traffic and provides a detailed view of the ongoing attack chain, enabling security teams to respond quickly.
- ▶ **CrowdStrike Falcon**¹⁵: CrowdStrike is well known for its endpoint protection solutions. Its Falcon platform uses behavioural-based techniques to detect and prevent threats that other signature-based systems might miss.
- ▶ **Cylance**¹⁶: CylancePROTECT is an endpoint protection solution that uses AI models to identify and block malware based on its characteristics and behaviours, rather than known signatures.
- ▶ **Gurukul**¹⁷: Provides user and entity behavioural analytics (UEBA) solutions that use machine learning algorithms to detect insider threats, fraud and unauthorised access.
- ▶ **Wazuh**¹⁸: This is an open-source platform for threat detection, vulnerability management and integrity monitoring. It uses rules and decoders to analyse security events and detect anomalous behaviour.
- ▶ **Snort**¹⁹: Although best known as an intrusion detection and prevention system (IDPS), Snort has evolved to incorporate behaviour-based capabilities. The Snort community develops and shares new rules that can detect anomalous behaviour.



13 <https://es.darktrace.com/>

14 www.vectra.ai

15 www.crowdstrike.com

16 www.cylance.com

17 www.gurukul.com

18 www.wazuh.com

19 www.snort.org

3. AI applications in Cybersecurity

- ▶ **ELK Stack (Elasticsearch, Logstash, Kibana)**²⁰: Although ELK itself is not a behavioural detection tool, it can be configured with specific plug-ins and rules to perform behavioural analysis of logs and events.

AI systems operating under the **Machine Learning for Behavioural Analysis model** are trained using large datasets of both legitimate and malicious behaviour. Through supervised learning, AI can learn to classify and detect anomalous activity. Thus, over time and as more data is processed, these systems can improve their accuracy through unsupervised learning and reinforcement learning.

Many modern cybersecurity tools have incorporated machine learning (ML) into their capabilities to improve threat detection and response. These tools use ML to learn and adapt to new threats by studying patterns and behaviours in data. In addition to the tools listed above, some of the most popular solutions are listed below:

- ▶ **Endgame**²¹: This platform uses ML for endpoint protection, threat detection and response. Its ML capability focuses on detecting attack techniques and tactics without relying solely on signatures.
- ▶ **PatternEx**²²: is a user and entity behavioural analysis (UEBA) solution that uses machine learning. It analyses large volumes of data to identify patterns that suggest malicious activity.
- ▶ **SentinelOne**²³: is an endpoint protection solution that uses machine learning to detect, classify and respond to malicious and anomalous behaviour.
- ▶ **Kaspersky Machine Learning for Anomaly Detection (MLAD)**²⁴: designed for industrial systems, Kaspersky's MLAD uses machine learning to detect deviations in the operation of industrial machines.
- ▶ **Splunk**²⁵: While Splunk is primarily a data analytics and SIEM (security information and event management) tool, it has capabilities that allow users to implement machine learning models to identify patterns and anomalies in large volumes of data.

AI systems operating under the Machine Learning for Behavioural Analysis model are trained using large datasets of both legitimate and malicious behaviour

²⁰ www.elastic.co

²¹ (Acquired by Elastic): <https://www.elastic.co>

²² <https://www.patternex.com>

²³ <https://www.sentinelone.com>

²⁴ <https://www.kaspersky.com>

²⁵ <https://www.splunk.com>

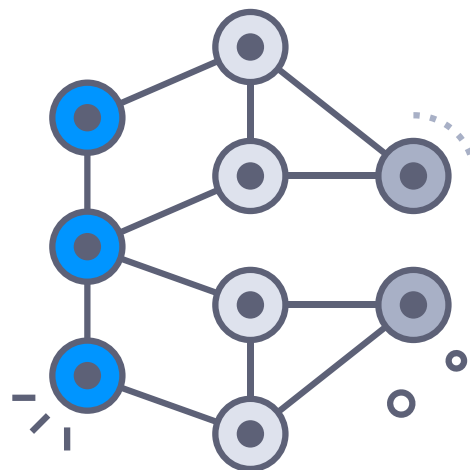
3. AI applications in Cybersecurity

Neural networks, especially deep learning neural networks, have been shown to be effective in detecting patterns in large datasets and can be used to identify malware in files based on their characteristics, detect DDoS attacks based on traffic patterns or identify phishing attempts through text and content analysis.

In addition to the commercial firms mentioned above, which are also working on neural networks, some of the best-known tools that make use of these techniques are listed below::

- ▶ **Deep Instinct**²⁶: this company uses deep learning neural networks to prevent and detect malware in real time, offering solutions for both endpoints and mobile devices.
- ▶ **SparkCognition**²⁷: this company offers DeepArmor, a solution that uses neural networks to provide real-time threat protection.
- ▶ **NVIDIA**²⁸: Although not a security tool per se, NVIDIA offers platforms and libraries such as CUDA and cuDNN that accelerate neural network computing.

On the other hand, there are many tools that have been able to incorporate AI mechanisms into their technologies, based on more traditional concepts, in order to make them more efficient.



25 <https://www.splunk.com>

26 <https://www.deepinstinct.com>

27 <https://www.sparkcognition.com>

28 <https://www.nvidia.com>

3. AI applications in Cybersecurity

Some of these **traditional practical applications include** the following:

Intrusion Detection and Prevention Systems (IDPS)

Using AI, these systems can detect and block malicious traffic in real time with greater accuracy.

An intrusion detection and prevention system (IDPS) is essential for detecting and responding to malicious activity on a network or system. The integration of artificial intelligence (AI) into these systems has significantly improved their ability to identify and react to threats in real time.

Examples include:

- **Darktrace** (<https://www.darktrace.com>): As mentioned above, Darktrace is known for its AI-based approach to threat detection and prevention. Its Enterprise Immune System technology uses machine learning to detect anomalous behaviour in real time.
- **Vectra** (<https://www.vectra.ai>): Vectra Cognito uses AI to automatically detect and prioritise anomalous behaviour in real-time to discover active attacks and insider threats.
- **Cisco Stealthwatch** (<https://www.cisco.com>): although not an IDPS in the traditional sense, Stealthwatch uses machine learning to detect anomalous behaviour in the network and integrates with other Cisco solutions to provide prevention capabilities.
- **Lastline** (<https://www.lastline.com>): offers solutions that use AI techniques, such as machine learning, to detect and respond to advanced, evasive and zero-day threats.
- **Awake Security** (<https://www.awakesecurity.com>): its platform uses AI to analyse network traffic and detect threats. It can identify malicious and risky behaviour without relying on signatures or prior knowledge.
- **Fortinet** (<https://www.fortinet.com>): While Fortinet offers a variety of security solutions, its FortiGate with built-in IDPS functionality has also included AI to improve threat detection.

3. AI applications in Cybersecurity

Forensic analysis tools

AI can speed up investigations after a security incident by quickly identifying indicators of compromise and mapping an attacker's path.

Digital forensics, especially when applied to security incidents (see the discipline DEFIR: Digital Forensics and Incident Response), can generate large amounts of data to investigate. Artificial intelligence and, in particular, machine learning (ML) is playing a significant role in this area, helping to identify patterns, perform faster analysis and obtain more accurate insights.

Some of the most popular tools that integrate AI into their forensic analysis capabilities are listed:

- **Autopsy** (<https://www.sleuthkit.org/autopsy/>): although primarily a digital forensic analysis tool, it has modules and plugins that can leverage AI-based capabilities to analyse data and look for specific patterns.
- **Cellebrite** (<https://www.cellebrite.com/>): Known for its mobile device forensics solutions, Cellebrite uses AI to assist in the identification and categorisation of relevant data on mobile devices.
- **Brainspace** (<https://www.brainspace.com/>): this is an analytics and visualisation platform that uses machine learning to assist in investigations, document reviews and data analysis. It is used in legal investigations but can also be applied in digital forensics.
- **Cyber Triage** (<https://www.cybertriage.com/>): Partnered with Autopsy, this tool uses AI techniques for rapid assessment of compromised systems, looking for evidence of malicious activity.
- **Endgame** (part of Elastic; <https://www.elastic.co/>): its platform provides incident and threat response capabilities and uses ML techniques to analyse data and detect malicious activity.
- **ReversingLabs** (<https://www.reversinglabs.com/>): provides solutions for the analysis of malicious files and artefacts with AI-based capabilities to identify, classify and disaggregate threats.

These tools, combined with human expertise, can provide faster and more accurate forensic analysis, which is crucial in incident response and investigations.

3. AI applications in Cybersecurity

Automated response systems

Once a threat is detected, the AI can initiate predefined actions to contain or mitigate the attack, such as isolating a compromised system or blocking a suspicious IP address.

Automated response, often combined with threat detection, is a crucial component of modern security. By using artificial intelligence (AI), these systems can make real-time decisions to contain, mitigate or neutralise threats without immediate human intervention. In addition to the commercial firms already mentioned, some of the most popular tools and solutions that integrate AI to provide automated response capabilities are listed below:

- **Darktrace Antigena** (<https://www.darktrace.com>): Antigena is an extension of Darktrace's AI-based detection system, which has the ability to take automatic actions in response to detected threats, such as blocking connections or quarantining devices.
- **Palo Alto Networks - Cortex XDR** (<https://www.paloaltonetworks.com>): This platform detects threats and automates the response. It uses machine learning techniques to identify threats and can perform actions such as blocking malicious processes or updating firewall rules automatically.
- **FireEye Helix** (<https://www.fireeye.com>): is a security platform that uses AI to detect threats and automate responses. It can integrate with a variety of tools and systems to execute response actions.
- **IBM Resilient** (<https://www.ibm.com>): is an incident response platform that, combined with Watson, IBM's AI system, can provide recommendations and automate actions in response to security incidents.
- **Fortinet FortiResponder** (<https://www.fortinet.com>): is an incident response solution that integrates with other Fortinet products to provide automated, rules-based capabilities. While response is primarily based on defined rules, detection and insights can be powered by AI techniques.

As it seems logical to assume, automation must be used with care. Incorrect configuration or lack of proper oversight can lead to unwanted responses that negatively affect operations. AI and automation should be seen as tools that complement, but do not replace, human security experts.

The use of these techniques also poses challenges. While AI can improve accuracy, there is still a risk of **false positives**, which in high numbers can lead to security team fatigue and possible omissions.

3. AI applications in Cybersecurity

As with signature-based systems, attackers are developing techniques to **evade AI-based detection**, such as so-called *data poisoning* or model manipulation.

Indeed, evasion of AI-based systems is a tactic employed by threat actors to avoid detection by security systems using artificial intelligence or machine learning techniques. These methods are based on understanding and exploiting the inherent weaknesses or biases of machine learning models. Some tools and techniques are designed specifically for this purpose, while others have been adapted to evade AI.

The following are some of the concepts or tools and resources that have been used or studied in relation to AI evasion:

ADVERSARIAL MACHINE LEARNING	It is a category rather than a specific tool. Adversarial attacks seek to introduce small perturbations in the input data to fool machine learning models.
CLEVERHANS²⁹	It is a software library that provides tools for testing the robustness of machine learning models against adversarial attacks.
DEEP-PWNING³⁰	It is a security assessment tool to assist in the analysis of systems using deep learning. It can be used to assess the resilience of models against adversarial modifications.
GAN (GENERATIVE ADVERSARIAL NETWORKS)	Although they are not evasion tools per se, GANs can be used to generate data to fool AI systems. As noted above, these networks are composed of two elements: a generator that creates images and a discriminator that attempts to distinguish between real and generated images.
FGSM (FAST GRADIENT SIGN METHOD)	It is an adversarial attack technique that introduces perturbations in the input data to confuse the machine learning model.

These security challenges include, as we have seen, the enormous potential for adversarial manipulation of training data and adversarial exploitation of model sensitivities to disrupt ML classification and regression performance.

Thus, AML (*Adversarial Machine Learning*) refers to the design of ML algorithms that can withstand security challenges, the study of attackers' capabilities and the understanding of the consequences of attacks.

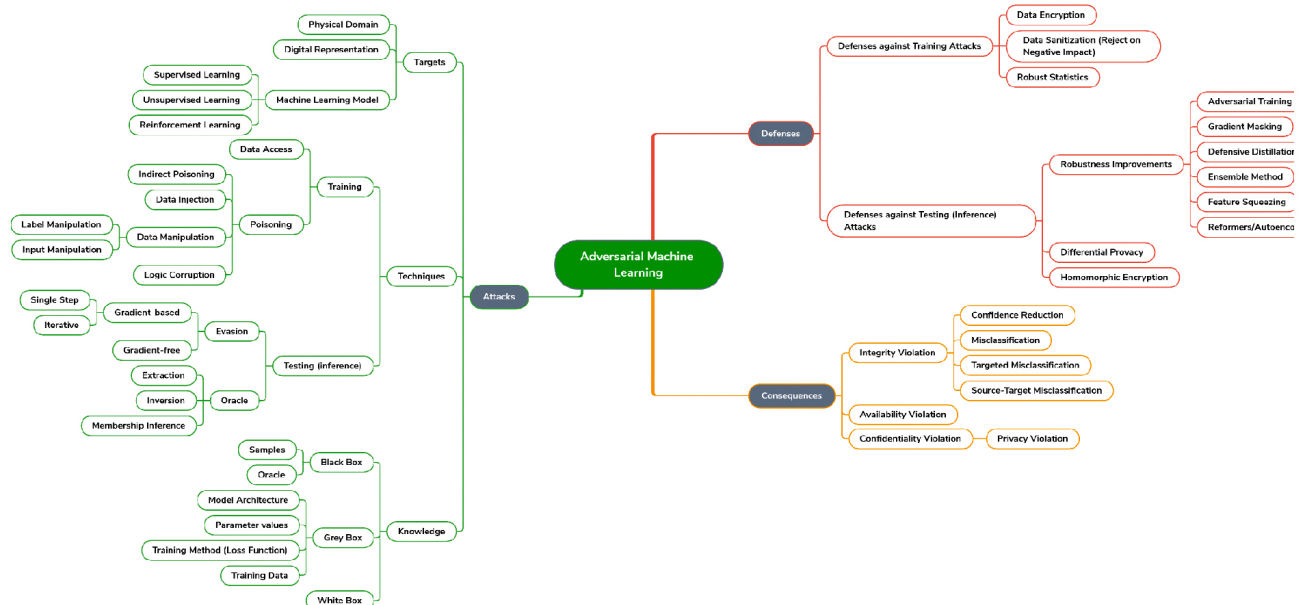
²⁹ <https://github.com/tensorflow/cleverhans>

³⁰ <https://github.com/cchio/deep-pwning>

3. AI applications in Cybersecurity

Since attacks are launched by adversaries with malicious intent, ML security must address defences aimed at preventing or mitigating the consequences of such attacks. While ML components can also be adversely affected by various unintended factors, such as design flaws or data biases, these factors are not intentional adversarial attacks, and do not fall within the scope of security addressed by the AML literature.

For its undoubted interest, we reproduce the **Taxonomy of Attacks, Defences and Consequences in AML**, from the National Institute Of Standards and Technology (NIST)³¹.



Finally, it seems clear that the adoption of AI for threat detection and behavioural analysis is on track to transform organisations' ability to defend against cyberattacks. However, as with any tool, it is essential to use it in conjunction with other cyber security techniques and approaches to ensure a comprehensive defence.

31 NIST - Draft NISTIR 8269 - A Taxonomy and Terminology of Adversarial Machine Learning (2019).

3.2 Automatic response and orchestration

Security Orchestration, Automation and Response (SOAR) refers to the ability of a security system to automatically detect and respond to a threat or vulnerability without human intervention, often coordinating multiple systems and tools in the process.

This model therefore requires three components: **orchestration**, understood as the coordination and integrated management of security tools and systems, allowing them to work together harmoniously³²; **automation**, understood as the ability to perform specific tasks without human intervention³³; and **response**, referring to actions taken in response to a security event, which may be automatic (e.g. blocking an IP) or may require human intervention (e.g. investigating a possible intrusion).

The use of tools configured under the SOAR model is particularly useful since, given the speed of threat propagation, the ability to automatically respond to threats can be crucial to minimise damage.

In addition, automation allows security teams to focus on higher value tasks or more sophisticated threats, leaving repetitive or routine tasks to automated solutions, removing the human factor from the handling of the latter and, consequently, reducing the risk of errors or inconsistencies.

Finally, SOAR systems can handle a large volume of alerts and events, many of which would be overwhelming for a human team.

Security Orchestration, Automation and Response (SOAR) refers to the ability of a security system to automatically detect and respond to a threat or vulnerability without human intervention, often coordinating multiple systems and tools in the process

³² For example, if a system detects malicious code (malware), orchestration could ensure that the information is shared with all relevant tools for subsequent analysis and response.

³³ These could include blocking a malicious IP, disabling compromised user accounts or even patching vulnerable software.

3. AI applications in Cybersecurity

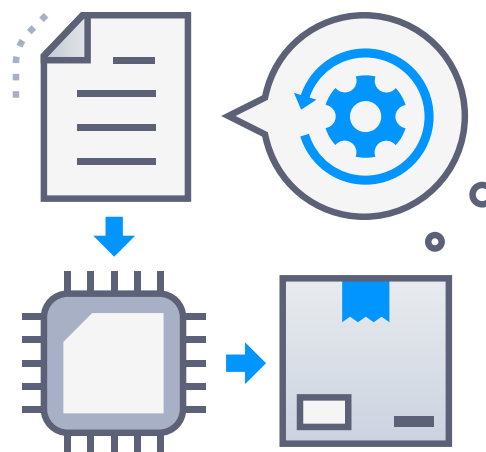
Despite their advantages, the use of SOAR tools also poses challenges: for example, poorly configured automatic response can cause more problems than it solves, such as blocking legitimate traffic or ignoring genuine threats; or the over-reliance on such systems without human review or oversight, which can lead to a lack of detection of more sophisticated or complex threats.

SOAR tools are typically used in the following **scenarios**:

- ▶ **Incident Response:** If a system detects anomalous behaviour, such as an unusual increase in traffic to a specific destination, it can automatically block that activity and alert the security team.
- ▶ **Tool Integration:** By integrating multiple tools (such as intrusion detection systems, firewalls and endpoint solutions), orchestration enables a more holistic and coordinated response to threats.
- ▶ **Workflow automation:** For example, upon detection of vulnerable software, a SOAR system could automatically initiate a patching or update process.

Some of the most popular SOAR tools are listed below:

- ▶ **Splunk Phantom**
(https://www.splunk.com/en_us/software/splunk-security-orchestration.html)
- ▶ **Siemplify**
(<https://www.siemplify.co/>)
- ▶ **Palo Alto Networks - Cortex XSOAR**
(<https://www.paloaltonetworks.com/cortex/xsoar>)
- ▶ **IBM Resilient**
(<https://www.ibm.com/security/incident-response/resilient-soar-platform>)
- ▶ **CyberSponse**
(<https://www.cybersponse.com/>)



3.3 Threat prediction

The concept of 'threat prediction' is an evolution of traditional threat detection and represents a shift in the approach to cyber security from a reactive to a proactive stance. Therefore, threat prediction refers to the process of anticipating and detecting potential cyber-attacks or vulnerabilities before they materialise, using advanced analytics and, for the part that interests us now, artificial intelligence techniques to identify patterns and signals that suggest an imminent attack or the emergence of new vulnerabilities.

The advantages of this model are clear: it seems clear that by predicting a threat before it occurs, organisations have time to prepare, strengthen their defences and reduce the potential impact of the attack. In addition, prediction allows organisations to focus their efforts and resources on those threats that are most likely to occur, rather than spreading them over a wide range of possible scenarios and, finally, by anticipating threats and acting on them proactively, organisations can improve their overall security posture.

Essentially, the **prediction** methods used by the tools of this model are the following three:

The concept of 'threat prediction' is an evolution of traditional threat detection and represents a shift in the approach to cyber security from a reactive to a proactive stance

As with previous models, this concept also presents significant challenges.

Trend analysis	By analysing past trends, organisations can anticipate types of threats that could emerge in the future.
Threat intelligence	It involves collecting and analysing information on existing and emerging threats from a variety of sources, such as intelligence feeds, investigator reports and security event data.
Predictive models	It uses algorithms and mathematical models, often powered by AI and machine learning, to analyse large datasets and identify patterns that suggest an imminent threat.

3. AI applications in Cybersecurity

Indeed, threat prediction, especially when based on predictive models, can lead to **false positives**, which can divert resources and attention from other critical areas. On the other hand, building and maintaining predictive models, especially those using advanced AI techniques, can be **complex** and require specialised staff. Finally, as always, the accuracy of predictions is highly dependent on the **quality, relevance and timeliness of the input data**.

Applications of this model in cybersecurity have focused on **predicting malicious code** (based on the characteristics of known malware, new variants or evolutions of malware can be predicted), **predicting phishing attacks** (by analysing patterns in previous phishing campaigns, future attacks can be anticipated or suspicious domains identified) and **predicting DDoS attacks** (by observing traffic patterns and other signals, it is possible to anticipate a DDoS attack before it occurs).

The following are some of the best-known examples of tools that have used this model, in its different variants, from traditional statistical analysis to machine learning and artificial intelligence, to anticipate threats before they occur. Some of these have already been mentioned above

The accuracy of predictions is highly dependent on the quality, relevance and timeliness of the input data

- ▶ **Darktrace Antigena**
(<https://www.darktrace.com/en/products/antigena/>)
- ▶ **Recorded Future**
(<https://www.recordedfuture.com/>)
- ▶ **Palo Alto Networks – AutoFocus**
(<https://www.paloaltonetworks.com/cortex/autofocus>)
- ▶ **CyberInt**
(<https://www.cyberint.com/>)
- ▶ **Lookout**
(<https://www.lookout.com/>)
- ▶ **SparkCognition DeepArmor** (<https://www.sparkcognition.com/deeparmor-endpoint-protection/>)
- ▶ **CrowdStrike Falcon**
(<https://www.crowdstrike.com/es-es/products/falcon-platform/>)
- ▶ **CylancePROTECT**
(<https://www.blackberry.com/us/en/products/blackberry-protect>)
- ▶ **Kenna Security Platform**
(<https://www.kennasecurity.com/platform/>)

3.4 Biometric identification and authentication

Biometric identification and authentication refer to the use of unique physical or behavioural characteristics of an individual to verify or confirm their identity. These characteristics may include, but are not limited to, fingerprints or fingerprinting, facial recognition, voice recognition, iris patterns, among others.

The following **types of biometrics** can be distinguished:

Physical Biometry	It is based on physical characteristics of the body, such as: • Fingerprints: The unique ridges and valleys on the fingertips. • Facial Recognition: The structure and characteristics of the face. • Iris recognition: The unique patterns in the iris of the eye. • Geometry of the Hand: The shape and size of the hand.
Behavioural Biometrics	It is based on the actions performed by the individual, such as: • Keystroke Dynamics: The way an individual presses the keys on a keyboard. • Voice Recognition: The unique characteristics of a person's voice. • Walking pattern: The way a person walks.

The use of biometric methods has several advantages, such as **unique-ness** (biometric characteristics are unique to each individual, reducing the likelihood of duplication or spoofing), **convenience** (users do not need to remember passwords or PIN codes) and **difficulty of forgery** (since it is difficult to replicate or forge biometric data, especially when compared to passwords).

3. AI applications in Cybersecurity

However, the use of biometric mechanisms poses certain **challenges and limitations**, for example:

- ▶ **Recognition errors:** no biometric system is 100% accurate. There can be false positives (recognising someone who is not the user) or false negatives (not recognising the legitimate user).
- ▶ **Privacy concerns:** the collection, storage and use of biometric data raises concerns about privacy, consent and legal compliance.
- ▶ **Irrevocability:** Unlike passwords, which can be changed, biometric characteristics are permanent. If biometric data is compromised, it cannot be replaced or altered.
- ▶ **Cost:** The implementation of biometric systems may require specialised hardware and software, which may incur additional costs.

Nevertheless, the use of biometrics has found various **applications in cybersecurity**, such as secure logical access (many devices and applications offer biometric authentication options as an additional layer of security), **online transactions** (biometric authentication can be used in online banking transactions and mobile payments to verify the user's identity), **physical access control** (biometric systems can be used to control access to buildings, rooms or other restricted areas, etc.).

As mentioned, biometric identification and authentication have become popular in many devices due to their ability to provide an additional layer of security. Some of the most well-known examples of tools and systems using biometrics are shown below:

- ▶ **Apple Face ID and Touch ID:** Face ID allows unlocking iPhone, iPad and some Macs using facial recognition, while Touch ID uses fingerprint (**Face ID** and **Touch ID**).
- ▶ **Windows Hello:** is a feature from Windows 10 that allows users to access their devices using facial or fingerprint recognition (<https://www.microsoft.com/es-es/windows/windows-hello>)
- ▶ **Samsung Pass:** is a biometric authentication tool that allows Samsung device users to unlock their smartphones and access applications and websites using iris, facial or fingerprint recognition (<https://www.samsung.com/global/galaxy/apps/samsung-pass/>)

Nevertheless, the use of biometrics has found various applications in cybersecurity, such as secure logical access, online transactions, physical access control

3. AI applications in Cybersecurity

- ▶ **BioID:** is a cloud-based facial authentication platform that can be integrated into various applications to provide biometric authentication (<https://www.bioid.com/>)
- ▶ **AuthenTrend:** offers fingerprint-based authentication solutions for different applications, from USB drives to enterprise solutions (<https://www.authentrend.com/>)
- ▶ **Nuance VocalPassword:** is a voice recognition solution that verifies the user's identity based on the unique characteristics of the user's voice (<https://www.nuance.com/omni-channel-customer-engagement/security/vocalpassword.html>)

These are just a few traditional examples of the many biometric authentication solutions available on the market. It is essential to bear in mind that when considering any biometric solution, it is crucial to assess security, privacy and usability to ensure that it meets specific organisational, user or regulatory requirements³⁴.

In addition to the above, biometric identification and authentication solutions have started to integrate **advanced AI capabilities**, especially in areas such as facial recognition and behavioural analysis, to improve accuracy and reduce false positives. Some of the most well-known examples are listed below:

- ▶ **Trueface:** uses AI to provide facial recognition solutions. Its algorithms learn and improve over time, which increases identification accuracy (<https://www.trueface.ai/>)
- ▶ **Kairos:** is a cloud-based platform that uses AI to analyse faces in videos and photos, offering facial recognition solutions (<https://www.kairos.com/>)
- ▶ **BehavioSec:** this platform uses AI to analyse behavioural patterns in real time, such as typing dynamics and mouse handling, to authenticate users (<https://www.behaviosec.com/>)



³⁴ This would be the case of compliance with the National Security Framework (Royal Decree 311/2022 of 3 May) for the entities within its scope of application.

3. AI applications in Cybersecurity

- ▶ **ID R&D:** uses AI in its voice and behavioural biometric solutions to provide more secure and efficient authentication (<https://www.idrnd.net/>)
- ▶ **Deepware Scanner:** is a fingerprint scanner based on deep neural networks. It uses AI to analyse and verify fingerprints with high accuracy.
- ▶ **Affectiva:** although primarily focused on emotional interpretation through facial analysis, Affectiva uses AI for real-time analysis of facial expressions, which has potential applications in areas of behavioural authentication or emotional responses (<https://www.affectiva.com/>)

3.5 Vulnerability scanning and automated pentesting

As is well known, **vulnerability analysis** is a systematic process to assess, identify and classify security vulnerabilities in information systems. These vulnerabilities can be caused by software bugs, inadequate configurations, hardware failures, or even poor security management practices.

This scanning process typically involves **identification** (tools scan systems, networks and applications for known vulnerabilities), **classification** (once detected, vulnerabilities are ranked according to their severity and risk), **remediation** (solutions are proposed to mitigate or fix the detected vulnerabilities) and **verification** (after remediation, a further verification is performed to confirm that the vulnerabilities have been adequately addressed).



3. AI applications in Cybersecurity

Penetration testing, commonly known as **pentesting**, is a simulated attack on a system with the goal of discovering vulnerabilities before the real attackers do so. Unlike vulnerability analysis, which typically uses automated scans to identify known vulnerabilities, pentesting often involves experts actively attempting to exploit vulnerabilities and penetrate systems, simulating the tactics, techniques and procedures (TTPs) of real adversaries.

The process generally involves **reconnaissance** (gathering information about the target), **scanning** (identifying possible entry points), **penetration** (exploiting vulnerabilities), **access maintenance** (simulating an attacker's movements after gaining access) and **analysis** (containing the report of findings and recommendations for fortifying the system).

As you might expect, **Artificial Intelligence** has also been incorporated into vulnerability analysis and penetration testing, with the following procedures:

Improved automation	With AI, tools can scan networks and systems faster and more accurately, identifying vulnerabilities that traditional tools might miss.
Continuous learning	AI-based tools can learn from each scan, adapting to new vulnerabilities and attack techniques.
Advanced simulation	In pentesting, AI can simulate more complex attacker behaviour, testing systems against emerging and advanced threats.
Prioritisation of risks	AI can help prioritise vulnerabilities based on context and historical data, allowing security teams to focus on the most imminent or damaging threats.
Integration and correlation	AI-based solutions can correlate data from multiple sources, offering a more holistic view of an organisation's security posture.

Tools such as **Tenable.io**, **Qualys Cloud Platform** or **Checkmarx** are already using AI capabilities to improve their scanning and analysis. In addition, pentesting platforms such as Cobalt are incorporating AI to automate and improve parts of the process.

The integration of AI in these areas is promising, but it is essential to remember that, for now, the combination of human experts with these advanced tools provides the most robust and comprehensive approach to cyber security.

3.6 Defence against automated adversaries

As technology advances, not only defenders improve their tools, but also attackers. **Automated adversaries** are those programs, bots and scripts designed to carry out attacks without direct human intervention. These attacks can range from simple brute force attacks to more sophisticated models that can adapt and change tactics on the fly.

The following table shows a **typology** of the most common automated adversaries and their general characteristics.

TYPES		FEATURES
Bots and Scrapers	They can be used for many tasks, such as scraping websites, but they can also be used for attacks, such as login attempts or exploiting vulnerabilities in a website.	1. Speed: They can launch attacks at a speed that is practically impossible for a human. 2. Adaptability: Some advanced automated systems can change tactics if they detect that a particular approach is not working. 3. Scale: They are capable of addressing thousands or even millions of targets simultaneously. 4. Persistence: They can continue their attacks for long periods without fatigue or distraction.
Worms	These are malicious programs that automatically replicate themselves to spread to other computers, often by exploiting vulnerabilities in software.	
DDoS bots	Some botnets are used to launch coordinated DDoS attacks, flooding targets with traffic to bring down services or infrastructure.	
Automated phishing systems	They can quickly generate fraudulent websites or send mass emails with malicious links.	

3. AI applications in Cybersecurity

Defence with Artificial Intelligence

To protect against these automated adversaries, defence must also be agile, adaptive and, in many cases, also automated. This is where artificial intelligence comes in. Let's look at the most common scenarios:

1. **Anomaly detection:** AI can analyse large data sets to detect anomalous patterns that may indicate an automated attack.

Anomaly detection is one of the most common applications of artificial intelligence in cybersecurity. The idea is to identify 'normal' patterns of behaviour and then detect deviations or 'anomalies' from those patterns, which could indicate malicious or unauthorised activity. Some of the most well-known tools in this regard that employ AI for anomaly detection are listed, some of which have already been mentioned above:

- **Darktrace** (<https://www.darktrace.com/>)
- **Splunk User Behavior Analytics (UBA)** (https://www.splunk.com/en_us/software/user-behavior-analytics.html)
- **Vectra Cognito** (<https://www.vectra.ai/products>)
- **Gurukul Risk Analytics** (<https://gurukul.com/products/risk-analytics>)
- **Exabeam Advanced Analytics** (<https://www.exabeam.com/product/advanced-analytics/>)

As with all security tools, it is essential to keep them up to date and to use them as part of a broader security strategy.

2. **Bot identification:** through behavioural analysis, AI can identify and block bots based on their interaction patterns.

Bot identification is essential, especially in the context of web traffic, digital advertising and social media, where bots can artificially boost metrics, divert traffic or spread misinformation or disinformation. Several solutions use artificial intelligence and machine learning to identify and block bot traffic in real time. Some of the most popular tools are listed below:

- **Imperva Bot Management** (formerly Distil Networks) (<https://www.imperva.com/products/bot-management/>)
- **Akamai Bot Manager** (<https://www.akamai.com/us/en/products/security/bot-manager.jsp>)
- **Cloudflare Bot Management** (<https://www.cloudflare.com/bots/>)
- **DataDome** (<https://www.datadome.co/>)
- **Reblaze** (<https://www.reblaze.com/>)
- **Cofense Triage** (<https://cofense.com/product-services/triage/>)

These tools offer real-time protection against bot traffic, allowing organisations to protect their online assets and ensure their metrics and analytics are accurate. It is essential that organisations choose a solution that fits their specific needs and aligns with their infrastructure and objectives.

3. AI applications in Cybersecurity

- 3. Continuous learning:** As automated adversaries evolve, AI-based solutions can learn from attacks and adapt.

In the context of cyber security, "continuous learning" (also known as "online learning" or "real-time learning") refers to the ability of a system to continuously adapt to changing threats, in real time.

Some of the best-known tools and systems that use continuous learning and artificial intelligence techniques to protect against automated adversaries are listed below (some of which we have already mentioned):

- **Darktrace Antigena** (<https://www.darktrace.com/en/products/darktrace-antigena/>)
- **CylancePROTECT** (<https://www.blackberry.com/us/en/products/blackberry-protect>)
- **SentinelOne Singularity Platform** (<https://www.sentinelone.com/>)
- **Endgame** (<https://www.elastic.co/security>)
- **CrowdStrike Falcon** (<https://www.crowdstrike.com/products/falcon-platform/>)

The key advantage of such tools lies in their ability to adapt and learn from threats in real time, allowing them to stay one step ahead of adversaries even as their tactics change.

- 4. Rapid response:** Upon detection of an attack, AI can take immediate action to mitigate the attack, either by blocking traffic, shutting down processes or alerting security teams.

Rapid response against automated adversaries is essential, as these actors can escalate attacks or evolve rapidly. AI-based solutions can respond in real time to identified threats, and some can even take autonomous action to mitigate or neutralise the threat.

Listed below are some of the better-known tools that use AI to provide rapid response against automated adversaries (some of which have been discussed previously):

- **Darktrace Antigena** (<https://www.darktrace.com/en/products/darktrace-antigena/>)
- **Cisco Threat Response** (<https://www.cisco.com/c/en/us/products/security/threat-response.html>)
- **Palo Alto Networks Cortex XSOAR** (formerly Demisto) (<https://www.paloaltonetworks.com/cortex/soar>)
- **FireEye Helix** (<https://www.fireeye.com/helix.html>)
- **Symantec Endpoint Protection (SEP) Adaptive Threat Protection (ATP)** (<https://www.broadcom.com/products/cyber-security/endpoint/endpoint-protection>)
- **Netscout Arbor DDoS Protection** (<https://www.netscout.com/solutions/ddos-protection>)

These solutions use artificial intelligence techniques to analyse and respond to security events in real time. In addition, many of them offer the ability to integrate with other security solutions, allowing organisations to build a defence-in-depth approach and respond quickly to threats from multiple vectors.

3. AI applications in Cybersecurity

5. **Adversary simulation:** using AI to simulate attacks in controlled environments (network teaming) helps to identify weaknesses and better prepare defences.

Adversary simulation, also known as "network teaming", has adopted artificial intelligence to improve the simulation and to more effectively test defences in different scenarios. Some of the best-known tools are:

- **Endgame Red Team Tools** (now part of Elastic) (<https://www.elastic.co/what-is/endpoint-security>)
- **MITRE Caldera** (<https://github.com/mitre/caldera>)

These solutions help organisations understand their vulnerabilities and improve their defence postures by simulating realistic attacks. However, it is essential to remember that adversary simulations are only one part of a comprehensive cyber security strategy. Continuous training, system and software updates, and constant vigilance are crucial to effective defence.

Defending against automated adversaries is a constantly evolving race. With the ability of attackers to automate and adapt their attacks, traditional defences, based solely on signatures or static rules, can quickly become outdated. Integrating artificial intelligence into the defence provides the agility and adaptability needed to stay one step ahead of these advanced adversaries.

With the ability of attackers to automate and adapt their attacks, traditional defences, based solely on signatures or static rules, can quickly become outdated



3.7 Generative AI and Cybersecurity

Generative artificial intelligence (AI) has become a valuable tool in a variety of fields, from art creation to data synthesis. In the context of cybersecurity, generative AI can be both a solution and a potential threat - let's take a look at how::

BENEFITS OF GENERATIVE AI IN CYBERSECURITY:	
Synthetic data generation	Generative AI can be used to create synthetic datasets that simulate network traffic or user behaviour, without compromising real data. This data can be used to train intrusion detection systems, without violating user privacy.
Simulation of attacks	Through generative adversarial networks (GANs), it is possible to simulate how an attacker would act, allowing organisations to test the robustness of their systems and make improvements before real incidents occur.
Creating test scenarios	Generative AI can help create realistic penetration testing scenarios, improving on traditional practices that often rely on pre-defined and less dynamic scenarios.
Reinforcement of learning	Generative AI, especially GANs, can be useful in reinforcement learning, where an agent (the generative network) and an adversary (the discriminative network) work together. This technique can be used to teach cybersecurity systems how to improve their detection and response to threats in real time.

3. AI applications in Cybersecurity

THREATS AND CHALLENGES OF GENERATIVE AI IN CYBERSECURITY:	
Vulnerabilities during and after model training	Since generative AI models are trained on data that is collected from all sorts of sources - and not always transparently - it is unknown exactly what data is exposed to this additional attack surface. Combined with the fact that these generative AI tools sometimes store data for long periods of time and do not always have the best security rules and safeguards in place, it is very possible for threat actors to access and manipulate training data at any stage of the training process.
Violation of the privacy of personal data	There is no structure to regulate what kind of data users enter into generative models. This means that corporate users — and indeed anyone else — can use sensitive or personal data without complying with regulations or obtaining permission from the source. Again, with the way these models are trained and data are stored, personally identifiable information can easily get into the wrong hands and lead to undesirable situations.
Intellectual property exposure	There have been cases of companies unintentionally exposing proprietary data to generative models in a harmful way. This exposure occurs most frequently when employees upload proprietary items or works, API keys and other confidential information into the system.
Jailbreaks and cybersecurity solutions	Many online forums offer jailbreaks, or secret ways for users to teach generative models to work against their established rules. For example, ChatGPT recently managed to trick a human into solving a CAPTCHA on its behalf³⁵. The ability to use generative AI tools to generate content in such different and human-like ways has enabled sophisticated phishing and malware schemes that are harder to detect than traditional approaches.
Creation of malware and attacks	Generative techniques can be used by malicious actors to generate malware variants that can evade traditional detection systems.
Phishing and deception	Generative AI tools can be used to create fake websites, emails or communications that mimic legitimate ones, which undoubtedly increases the effectiveness of phishing attacks.

35 <https://cdn.openai.com/papers/gpt-4.pdf>

3. AI applications in Cybersecurity

Manipulation and falsification of data	GANs and other techniques can be used to create fake log records or manipulate data, which can make attacks undetectable or divert the attention of security teams.
Deepfakes in cybersecurity	The ability to create deepfakes (fake videos or audios that look real) can be exploited in targeted attacks to trick employees or executives into taking actions that compromise security.
Mitigation and adaptation	The key to tackling the threats associated with generative AI in cybersecurity lies in the constant adaptation and updating of defence tools and techniques, i.e: • Continuous monitoring of the latest research and trends in generative AI. • Regular training of cyber security teams on the capabilities and threats associated with generative AI. • Adoption of AI systems that can adapt and learn from generative techniques, staying one step ahead of threats.

Since the use of generative AI presents, as we have seen, significant challenges, it does not seem to be superfluous to select some Cybersecurity Tips and Best Practices for the use of generative AI³⁶, namely:

► **Carefully read the security policies of generative AI providers**

After initial protests about the lack of transparency of certain generative AI vendors in the training of their models, many of the major vendors have started to provide extensive documentation explaining how their tools work and what the user agreements are based on.

To find out what happens to the input data, you should consult the vendor's policies on data deletion and timelines, as well as what information they use to train their models. It is also good practice to look in their documentation for mentions of traceability, log history, anonymisation and other features you may need to meet specific regulatory requirements.

Especially important: look for any mention of opt-in and opt-out options and how to choose to have data used or stored.

36 Source: Hiter, S. Generative AI and Cybersecurity. eWeek (June, 2023)

3. AI applications in Cybersecurity

▶ **Do not enter sensitive data when using generative models**

The best way to protect the most sensitive data is to keep it out of generative models, especially those with which you are not familiar.

It is often difficult to know what data can or will be used to train future iterations of a generative model, let alone how much of the corporate data will be stored in the provider's data records and for how long.

Rather than blindly relying on the security protocols that these providers may or may not have, it is better to create synthetic copies of data or avoid using these tools altogether when working with sensitive data.

▶ **Keeping AI generative models up to date**

Generative models receive regular updates and sometimes these updates include bug fixes and other security optimisations. It is necessary to keep an eye out for opportunities to update the tools so that they remain effective.

▶ **Train employees on proper use**

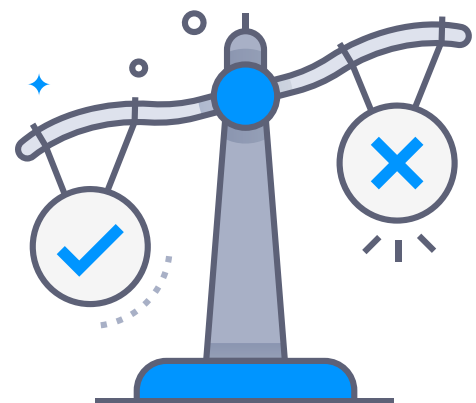
Generative AI tools are known to be easy to use and therefore easy to misuse. It is important for employees to know what kind of data they can use as inputs, what parts of their workflow can benefit from generative AI tools and what the compliance expectations are, in addition to meeting the organisation's general regulatory obligations regarding the use of electronic media.

▶ **Use data security and governance tools**

Data security and governance tools can protect your entire attack surface, including any third-party generative AI tools you may be using.

Consider investing in data loss prevention, threat intelligence, cloud-native application protection platform (CNAPP) and/or extended detection and response (XDR) tools.

The best way to protect the most sensitive data is to keep it out of generative models, especially those with which you are not familiar



3. AI applications in Cybersecurity

Some examples of **security tools and solutions using generative AI** are shown below.

	Google Cloud Security AI Workbench	This new Google development is based on Google Cloud's Vertex AI and is powered by Sec-PaLM. Google Cloud Security AI Workbench is designed to support advanced threat and security intelligence, malware detection, behavioural analysis and vulnerability management. https://cloud.google.com/blog/products/identity-security/rsa-google-cloud-security-ai-workbench-generative-ai
	Microsoft Security Copilot	Microsoft Security Copilot is one of the most targeted security solutions in Microsoft's arsenal of generative AI products. It works to optimise incident response, threat detection and security reporting for users, and integrates insights and information from tools such as Microsoft Sentinel, Microsoft Defender and Microsoft Intune. https://www.microsoft.com/en-us/security/business/ai-machine-learning/microsoft-security-copilot
	CrowdStrike Charlotte AI	This CrowdStrike tool allows users to manage cybersecurity through natural language on the Falcon platform. Like many of these emerging cybersecurity AI tools, Charlotte AI is designed to complement existing security teams and reduce the impact of skills gaps. Charlotte AI is generally used to support threat detection and remediation efforts. https://www.crowdstrike.com/press-releases/crowdstrike-introduces-charlotte-ai-to-deliver-generative-ai-powered-cybersecurity/
	Cisco Security Cloud	Cisco is adding generative AI capabilities to the Security Cloud and its Collaboration and Security portfolios. The new features are designed to make policy management and threat response easier - even conversational. https://investor.cisco.com/news/news-details/2023/Cisco-Unveils-Next-Gen-Solutions-that-Empower-Security-and-Productivity-with-Generative-AI/default.aspx
	Airgap Networks ThreatGPT	Based on GPT-3 and graph databases, ThreatGPT is an Airgap Networks solution that helps enterprises more effectively and holistically analyse security threats in operational technology (OT) environments and legacy systems. https://airgap.io/embargo-until-tbd/

3. AI applications in Cybersecurity

	SentinelOne	The organisation recently upgraded (and restricted) its threat capture platform with generative AI capabilities. It is designed to scale security operations and threat detection, relying on integrated neural networks and extensive language modelling to provide better and closer to real-time information about potential threats and solutions. https://www.sentinelone.com/press/sentinelone-unveils-revolutionary-ai-platform-for-cybersecurity/
	Synthesis Humans	Synthesis Humans is one of the many generative tools offered by Synthesis AI. This solution is designed to train biometric access control systems in a more agile way. In combination with Synthesis Scenarios, this tool can be used to support facility security as well as cyber security. https://synthesis.ai/synthesis-humans/
	SecurityScorecard	SecurityScorecard has launched a security rating platform partly based on OpenAI's GPT-4. With this solution, security teams can ask open, plain-language questions about the security of their network and third-party vendors, and receive proactive answers and risk management guidance. https://securityscorecard.com/company/press/securityscorecard-launches-first-and-only-security-ratings-platform-with-openais-gpt-4-search-system-providing-customers-with-faster-security-insights/
	MOSTLY AI	MOSTLY AI is a synthetic data generation tool specifically designed to generate anonymised data that meets various security and compliance requirements. Due to its strong focus on security and compliance, it is frequently used in regulated sectors such as banking and insurance. https://mostly.ai/

4. Study scenarios

The study scenarios not only give us a tangible understanding of how artificial intelligence (AI) is being used in the real world to combat cyber threats, but also reveal the inherent strengths and weaknesses of these approaches.

Over the years, the integration of AI into cybersecurity has led to significant successes, as well as lessons learned from incidents where AI-based solutions failed to detect or prevent attacks. These case studies illustrate how organisations, public or private, large or small, are using AI to protect their digital assets.

In this section, we will explore some examples where AI solutions have successfully detected, prevented or mitigated cyber-attacks, highlighting how the technology has been able to help overcome traditional capabilities.

The study scenarios not only give us a tangible understanding of how artificial intelligence (AI) is being used in the real world to combat cyber threats, but also reveal the inherent strengths and weaknesses of these approaches

4.1 Modern threat detection and response systems

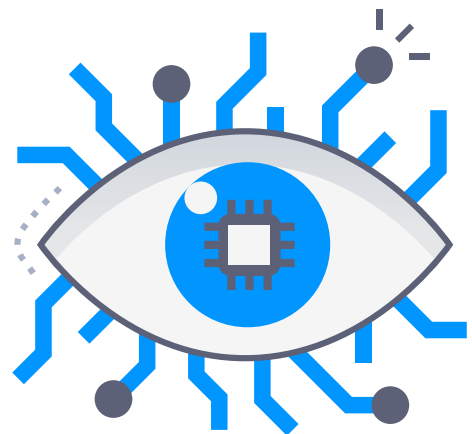
Threat detection and response has been a mainstay in the world of cyber security for years. However, today, with the massive adoption of artificial intelligence (AI)-based technologies, these systems have evolved considerably.

Indeed, these modern systems, commonly referred to as **managed threat detection and response (MDR) solutions** or **endpoint detection and response (EDR) systems**, often incorporate AI capabilities to improve the efficiency of their operations.

The key features of these current systems are **advanced automation**, as they use AI to recognise malicious patterns and behaviours in large datasets in real time, enabling faster response to threats; **continuous learning**, as they adapt and evolve over time, learning from new types of attacks and adapting to new threat patterns; **integration and orchestration**, given the ability to integrate with other tools and systems to provide a cohesive and orchestrated response to threats.

Some **successful application scenarios** are summarised below.

These modern systems, commonly referred to as managed threat detection and response (MDR) solutions or endpoint detection and response (EDR) systems, often incorporate AI capabilities to improve the efficiency of their operations



4. Study scenarios

Detection of zero-day attacks	Use of an AI-based MDR system to identify and prevent a zero-day attack before it becomes a widespread threat. ● Context: Zero-day attacks refer to unknown vulnerabilities in software, which malicious actors exploit before developers can create and distribute a patch. Given their nature, these attacks are difficult to prevent with traditional cybersecurity systems. ● Situation: Organisations can implement advanced AI-based MDR systems, which will lead to the detection of anomalous activity in, for example, software that is widely used in the organisation but has not yet been reported as vulnerable. ● Action: The AI system, using behavioural analysis, would have identified patterns of data access and modification that did not align with normal usage patterns. Instead of relying on known malware signatures, it would have focused on unusual behaviour. All of this would have made it easier for the organisation to be put on alert and isolate the affected software, preventing a potential large-scale security breach. ● Result: early detection could not only protect the organisation's information, but also alert the software developer and the security community, enabling a rapid response to protect other users.
Automated ransomware response	The case where an EDR system detects and mitigates a ransomware attempt in seconds, saving an organisation from significant disruption. ● Context: Ransomware, a type of malware that encrypts user data and demands a ransom to decrypt it, has evolved in complexity over the years. Ransomware attacks can cripple entire organisations, with significant costs in terms of downtime, data loss, financial and reputational losses. ● Situation: an organisation is attacked by an unknown ransomware variant. Within seconds, the malicious code would have started encrypting files on several systems. ● Action: The AI-based EDR system, which would have been installed in the organisation, would have detected the anomalous behaviour: rapid and massive file access followed by modifications consistent with encryption. The EDR could have automatically isolated the affected systems, being able to revert the changes made by the ransomware in a very short time. ● Result: the attack could have been contained quickly, and the organisation would have avoided significant losses and downtime. In addition, valuable data about the ransomware could have been collected, which can help strengthen the defences not only of the attacked organisation but also of other organisations by sharing indicators of compromise.

Notwithstanding the above, the use of these techniques, as we have seen in the preceding sections, also poses **challenges and lessons learned**:

4. Study scenarios

Challenges:

1. **False positive detection:** AI-based systems, especially when first trained or configured, can generate alerts about activity that, while unusual, is not necessarily malicious. This can trigger unnecessary responses and divert resources.
2. **Adaptability of adversaries:** Malicious actors are not static; they evolve and change their tactics, techniques and procedures (TTPs) to circumvent security systems. This means that what works today to detect an attack may not be effective tomorrow.
3. **Integration with existing infrastructure:** Not all organisations have the capacity to implement next-generation cyber security systems from scratch. Often, they must integrate new solutions with *legacy* systems, which can present compatibility and efficiency challenges.
4. **Difficulty in interpretation:** Decisions made by advanced AI models can often be "black boxes", i.e. difficult for humans to interpret or understand, which can lead to mistrust or confusion among security teams.

Lessons learned:

1. **Need for constant training:** Just as an anti-virus needs regular signature updates, AI systems require continuous training with fresh data to maintain their effectiveness.
2. **Importance of human feedback:** It is essential that cyber security analysts provide feedback to the system on the accuracy of alerts. This helps to adjust and improve the model over time.
3. **Defence in depth:** When it comes to cyber security, do not rely solely on an AI-based system. It is important to have multiple layers of defence and not to neglect basic security hygiene practices.
4. **Collaboration and intelligence sharing:** In today's interconnected world, sharing indicators of compromise, tactics and other forms of threat intelligence can help other organisations prepare for and defend against emerging threats.
5. **Gradual adoption:** It is prudent to implement and evaluate AI-based systems in controlled or *sandbox* environments before embarking on a full deployment. This allows potential problems to be identified and addressed in a more controlled environment.

These challenges and lessons underline the complexity of today's cyber-security AI landscape and the need for innovative, but also thoughtful and holistic approaches to address threats.

4. Study scenarios

Indeed, approaching the deployment of such tools requires careful planning that should consider the following elements and phases:

Adoption and adaptation

The adoption and adaptation of new technologies, particularly in the field of cybersecurity, requires a careful approach. We will focus on the adoption and adaptation of AI-based systems for threat detection and response:

Phase I: Pre-assessment

- **Current needs and gaps:** It is critical to first identify areas where the organisation faces cybersecurity challenges. These may include the detection of blind spots, slow response time or even a high volume of false positives.
- **Integration requirements:** how will the new system be integrated into the existing technology infrastructure? Technical aspects must be considered, but also process and team aspects.

Phase II: Selection of the solution

- **Customisation vs. off-the-shelf solutions:** some organisations may opt for customised systems, tailored to their specific needs, while others may find off-the-shelf solutions adequate.
- **Pilot testing:** before adopting a solution throughout the organisation, it is advisable to test in a limited environment to evaluate its effectiveness and ensure that it integrates well with existing systems.

Phase III: Implementation and fine-tuning

- **Staff training:** It is essential that cybersecurity staff understand how the new system works, how to interpret its results and how to act on them.
- **Initial feedback and fine-tuning:** The first few months of implementation are critical for gathering feedback. This feedback is used to fine-tune the system, reducing false positives and improving detection of legitimate threats.

Phase IV: Continuous evaluation and adaptation

- **Performance assessment:** As the threat landscape changes, it is essential to regularly assess how the system is performing and whether it is meeting expectations.
- **Adapting to new threats:** Artificial intelligence, especially in machine learning, may require new data or readjustments to address emerging threats. Cybersecurity solutions must be dynamic and adapt to the evolving threat landscape.

Phase V: Review and improvement

- **Incorporation of new features or capabilities:** as technology advances, new features or capabilities may emerge that it is desirable to incorporate into the existing system.
- **Iteration based on feedback:** lessons learned during system operation should be the basis for continuous improvements, thus ensuring that the solution remains relevant and effective in the face of emerging threats.

Corollary:

The adoption and adaptation of AI-based systems for cyber security is not a static process, but requires continuous engagement, evaluation and adjustment to ensure that the organisation remains protected against constantly evolving threats.

Evolving threats in response to modern systems:

The evolution of threats in response to modern defence systems is a complex and dynamic phenomenon. As new technological solutions are implemented, cybercriminals also adapt, developing more advanced tactics and techniques. This leads to a continuous cycle of adaptation and evolution between defenders and attackers.

1. AI-based detection evasion:

- **Polymorphic attacks:** these attacks automatically change their appearance/ code *signature* to avoid detection. This can be done through changes to the malicious code or obfuscation of its behaviour.
- **Adversarial machine learning techniques:** as discussed above, this refers to strategies specifically designed to confound AI models, such as introducing small perturbations in the data that can lead to the misclassification of malicious content as benign.

2. Harnessing automation:

- **Large-scale, fast-spreading attacks:** automated systems can launch attacks on a scale and speed that would be impossible for humans, such as the rapid spread of ransomware or worms.
- **Saturation attacks:** these attempt to overwhelm a system's detection and response capabilities by sending a flood of malicious traffic or requests, as in the case of DDoS attacks.

4. Study scenarios

3. More targeted attacks:

- **Spear phishing and targeted attacks:** instead of mass attacks, cybercriminals can use the information collected to specifically target individuals or organisations, often using highly personalised social engineering tactics.
- **ATAs, APTs (Advanced Targeted Attacks, Advanced Persistent Threats):** these attacks (many of them state-sponsored) are highly sophisticated and can involve a multiplicity of tactics and techniques to evade detection and reach their target.

4. Exploitation of emerging technologies

- **IoT (Internet of Things) and Edge Computing:** the proliferation of connected devices presents new opportunities for threat actors, especially as many of these devices lack adequate security measures.
- **Attacks in cloud environments:** As more organisations move their operations and data to the cloud, cybercriminals are looking to exploit vulnerabilities in cloud-based configurations and services.

5. Countermeasures and counterintelligence:

- **Defence discovery:** tools and tactics aimed at discovering a target's defences, identifying its weaknesses in order to adapt the subsequent attack accordingly.
- **Disinformation attacks:** which may involve the creation and dissemination of false information to divert attention from real defences or to discredit legitimate security alerts.

We emphasise that the continuing evolution of threats in response to advances in cyber security underscores the importance of constant adaptation and innovation in the field of cyber defence. Organisations must take a proactive approach, anticipating the emerging tactics of attackers and adjusting their defences accordingly.

In short, while modern AI-based detection and response systems are offering unprecedented capabilities in the fight against cyber threats, they also present new challenges and adaptation requirements for both the tools and the professionals who use them.



4.2 Successful implementations of AI in cybersecurity

Successful implementations of AI in cybersecurity provide valuable **case studies for** understanding how technology can strengthen an organisation's security posture. These examples also offer lessons on how to effectively integrate AI into existing infrastructures and how to overcome common challenges.

EXAMPLE SCENARIO		LESSON
Advanced threat detection	Organisations can detect suspicious activity within their network that other systems have missed. Using machine learning algorithms, the solution could analyse traffic patterns and detect anomalies that indicate a data compromise.	Machine learning can be particularly effective in detecting unknown or zero-day threats by observing deviations from normal behaviour.
Automated incident response	An organisation providing e-commerce or e-processing services can implement AI-based systems that, upon detecting a spike in web traffic that could be evidence of a DDoS attack, automatically redistribute and filter the traffic, minimising the impact on its operations.	A rapid and automated response can mitigate the damage of an attack in real time, especially when dealing with threats of the same type.
Biometric authentication	An entity can implement a facial recognition system for its mobile applications, providing an additional layer of security. AI could not only analyse facial features, but also behavioural patterns, such as the way a user holds their device.	AI can add layers of multi-factor authentication based on intrinsic characteristics and user behaviours.
Simulation of opponents	An organisation could use AI to simulate attacks on their own network, allowing them to identify vulnerabilities and strengthen their security posture before a real attack occurs.	AI-based simulations can help organisations prepare for real threats by identifying weaknesses in their infrastructure.
Forensic analysis	After an attack, an organisation could use AI tools to quickly analyse logs and records, identifying how the attackers were able to penetrate the system, what data they compromised and how they moved within the network.	AI can significantly speed up the forensic analysis process, enabling faster recovery and providing vital information to prevent future incidents.

4. Study scenarios

These scenarios show the variety of ways in which AI can be successfully integrated into the cybersecurity landscape. While each organisation will face unique challenges, these implementations offer tangible evidence of the benefits and advantages AI can provide in the fight against cyber threats.

Below are links to some concrete success stories related to the implementation of AI in cybersecurity. It should be noted, however, that due to the confidential nature of many cybersecurity incidents, some companies may not disclose specific details of the incidents or how they were resolved.

While each organisation will face unique challenges, these implementations offer tangible evidence of the benefits and advantages AI can provide in the fight against cyber threats

1. Advanced threat detection:

- **Company:** Darktrace
- **Details:** Darktrace uses machine learning and AI-based algorithms to detect, respond to and mitigate cyber threats in real time. One of its success stories involved an energy company where a compromise was detected on one of its workstations, which was being used to scan the internal network.
- **Outcome:** According to the entity, the activity was identified and stopped quickly, preventing a potential larger-scale compromise.
- **Referrer URL:** [Darktrace Success Stories](#)

2. Automated incident response:

- **Company:** Cloudflare
- **Details:** Cloudflare offers solutions to protect websites against all kinds of threats, including DDoS attacks. In one case, they protected one of their customers from a DDoS attack exceeding 400 Gbps.
- **Outcome:** The malicious traffic was successfully filtered out, and the client's website remained online without interruption.
- **Referrer URL:** [Cloudflare Blog](#)

4. Study scenarios

3. Biometric authentication:

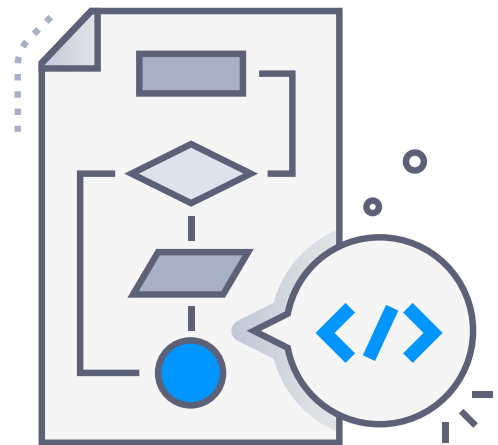
- **Company:** HSBC
- **Details:** Financial institution HSBC implemented voice recognition technology to verify the identity of its customers when they contact the institution. According to the bank, the solution's identification is based on more than 100 unique characteristics of a person's voice.
- **Outcome:** Reduced authentication time and improved customer experience, while adding an additional layer of security.
- **Referrer URL:** HSBC Voice ID

4. Simulation of opponents:

- **Company:** Cymulate
- **Details:** Cymulate is a platform that allows organisations to simulate attacks on their own networks. One client, an insurance company, used Cymulate to identify and mitigate vulnerabilities before they were exploited.
- **Outcome:** The company was able to strengthen its security posture and was better prepared to face real threats.
- **Referrer URL:** Cymulate

These success stories provide insight into how artificial intelligence and machine learning are being used in real-world situations to improve cybersecurity. It is crucial, however, to do in-depth research on each of these cases to get specific details and fully understand their impact and scope.

It is crucial, however, to do in-depth research on each of these cases to get specific details and fully understand their impact and scope



4.3 Failures and lessons learned

Analysing failures and extracting lessons learned is essential to understanding the bigger picture of any technology or application. In the context of artificial intelligence applied to cybersecurity, while there have been many successes, there have also been challenges and mistakes that have served as crucial learning points for the industry. Some examples are shown below.

TYPOLOGY AND DESCRIPTION		LESSONS LEARNT
Over-reliance on automated solutions	Organisations have sometimes relied too heavily on their AI systems for threat detection, assuming that AI would detect all possible threats. However, no system is infallible.	It is essential to have a balance between AI solutions and human oversight. Human experience and judgement are essential in the world of cybersecurity.
Adversarial attacks against AI models	Attacks have emerged that seek to trick or confuse machine learning models. For example, malware samples can be slightly modified to make them undetectable by AI-based systems.	It is crucial to constantly update and train AI models with recent and relevant data. In addition, specific defence techniques against adversarial attacks must be applied.
False positives	In some deployments, AI systems have generated a significant number of false positives, which can lead to overloading security teams and the possibility of missing real threats amidst the noise generated.	It is essential to constantly adjust and optimise AI models and algorithms to reduce the number of false positives and improve accuracy.
Dependence on quality data	The effectiveness of AI depends on the quality of the data it is trained on. If an AI system is trained on inadequate or biased data, its predictions or detections may be incorrect or ineffective.	It is essential to ensure the quality, diversity and representativeness of the data used to train AI systems.
Cost of implementation	The adoption and implementation of AI solutions can be costly, not only in economic terms, but also in terms of time and resources. Some organisations may have underestimated these costs and experienced difficulties in the implementation phase.	A detailed cost-benefit analysis is essential before implementing AI solutions in cyber security.

These failures and lessons learned highlight the importance of taking a balanced and careful approach when implementing AI in cybersecurity. While AI offers powerful tools and capabilities, it remains essential to be mindful of its limitations and challenges.

5. Challenges and limitations of AI in Cybersecurity

As mentioned above, artificial intelligence is directly impacting the field of cybersecurity, offering innovative solutions for threat detection and prevention, behavioural analysis and automated incident response. However, like any emerging technology, AI is not without its challenges and limitations. Despite its transformative potential, expectations towards AI must be balanced with a clear understanding of its constraints.

These challenges encompass not only technical aspects, such as the quality of data training or the interpretation of results, but also ethical dilemmas and privacy concerns. Moreover, as cybercriminals adapt and evolve, new obstacles to AI-based systems emerge, from adversarial attacks to model manipulation.

In this section, we will explore in detail the challenges inherent in the use of AI in cybersecurity, the current limitations of this technology, and the areas where, despite advances, human intervention and judgement remain irreplaceable. In doing so, we seek to provide a balanced and realistic perspective that allows organisations to maximise the benefits of AI while remaining alert to its potential limitations.

Despite its transformative potential, expectations towards AI must be balanced with a clear understanding of its constraints

5.1 Adversarial attacks against AI models

Adversarial attacks against AI models have emerged as a critical concern in the field of cybersecurity. As we have pointed out in this paper, these attacks are designed to deceive or confuse machine learning models, which could lead to erroneous or malicious decisions by such systems.

Effectively, an adversarial attack involves the introduction of small perturbations in the input data, designed to be almost imperceptible to humans, but which can lead the model to make incorrect predictions. These perturbations are carefully calculated to maximise the model's prediction error.

Adversarial attacks can be of two **types**:

WHITE BOX ATTACKS	In this scenario, the attacker has complete knowledge of the model, including its architecture and parameters. This allows him to design perturbations that are especially effective against the specific model.
BLACK BOX ATTACKS	In this case, the attacker does not have direct access to the model and its parameters, but may have access to its predictions. Although this scenario is more challenging for the attacker, it is still possible to generate effective adversarial perturbations.

These adversarial attacks have several **negative implications for cyber security**.

5. Challenges and limitations of AI in Cybersecurity

For example, in the case of **malware detection**, if an AI system is used to detect malware, an attacker could design malware that, once altered, goes undetected by the model. In the case of **authentication systems**, if an AI-based system handles authentication, e.g. through facial recognition, an adversarial attack could allow unauthorised access to an intruder. Finally, in the case of **network traffic analysis**, attackers can manipulate specific characteristics of network traffic to evade detection by an AI-based system.

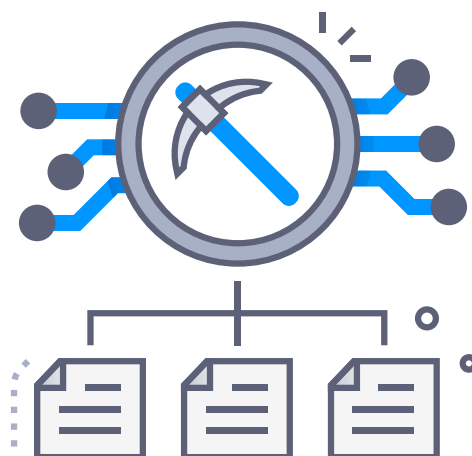
In response to this, several **countermeasures** can be developed, among them:

- ▶ **Adversarial training:** this technique involves training the model with adversarial examples, which can increase its robustness against such attacks.
- ▶ **Disturbance detection:** some methods aim to detect adversarial disturbances directly rather than attempting to make accurate predictions in their presence.
- ▶ **Regularisation and defence techniques:** These are techniques designed to make models inherently more resistant to adversarial attacks by adjusting their behaviour during training.

Adversarial attacks against AI models are a manifestation of a fundamental truth in cybersecurity: any system, no matter how advanced, has vulnerabilities. The goal would be to stay one step ahead of attackers, constantly adapting and evolving in response to new threats.

Both attackers and defenders make use of **advanced tools**, and many of these tools integrate AI capabilities. Some of the most popular ones, both for attack and defence, are shown below.

Adversarial attacks against AI models are a manifestation of a fundamental truth in cybersecurity: any system, no matter how advanced, has vulnerabilities



5. Challenges and limitations of AI in Cybersecurity

OFFENSIVE tools, which can be used by attackers	DeepExploit: is an automated pentesting tool that uses deep learning. It is able to learn from the results of previous penetration tests and adapt its techniques accordingly. https://github.com/13o-bbr-bbq/machine_learning_security/tree/master/DeepExploit
	Snallygaster: a tool that searches for exposed files on web servers, using AI techniques to identify potential attack vectors. https://github.com/hannob/snallygaster
	GPT-2: although not originally designed as an attack tool, this natural language technology developed by OpenAI can be used to generate fake content, such as phishing emails. https://github.com/openai/gpt-2
DEFENSIVE tools	TensorFlow Privacy: a library that helps developers train machine learning models with differential privacy, which can help protect training data. https://github.com/tensorflow/privacy
	IBM's Adversarial Robustness Toolbox (ART): is a Python library that provides tools to improve the robustness of machine learning models and deepen them against adversarial attacks. (https://github.com/Trusted-AI/adversarial-robustness-toolbox)
	DeepArmor: is a cybersecurity solution that uses deep learning techniques to detect and prevent malware in real time. https://www.sparkcognition.com/deeparmor-endpoint-security/
	CylancePROTECT: this is endpoint software that uses AI models to predict and prevent malware and advanced scripting. https://www.cylance.com/cylanceprotect

These tools are only a small subset of the options available on the market.

5.2 Overreliance on automated solutions

The **over-reliance on automated solutions** in cybersecurity, and in particular those based on Artificial Intelligence (AI) and Machine Learning (ML), has significant implications and associated risks, namely:

1. Lack of interpretability:

AI, especially deep learning, can function as a "black box". This means that even if a model can predict or classify with high accuracy, it is often difficult to understand how it arrived at a particular decision. This raises concerns in cybersecurity, where traceability and understanding of the decisions made are critical to assess the effectiveness and reliability of the system, and could constitute a legal non-conformity, if the system in question is subject to specific regulation, as prescribed by the European Artificial Intelligence Regulation, for risky systems.

2. False sense of security:

The deployment of AI solutions can lead organisations to believe that they are fully protected against threats. However, no system is foolproof. If organisations rely solely on automated solutions, they may overlook critical areas of vulnerability or be unprepared to respond when these solutions fail or are circumvented.

3. Threat evolution:

Attackers are constantly adapting and evolving their methods to evade defence systems. If AI solutions are not continuously updated and adapted to the changing threat landscape, they can quickly become obsolete.

The over-reliance on automated solutions in cybersecurity, and in particular those based on Artificial Intelligence (AI) and Machine Learning (ML), has significant implications and associated risks

5. Challenges and limitations of AI in Cybersecurity

4. Targeted attacks against AI:

Attackers are increasingly aware of how AI-based systems work and are developing specific techniques, such as adversarial attacks, to deceive or circumvent these systems. Over-reliance on AI solutions without due diligence can expose organisations to these specialised attacks.

5. Failures in automation:

AI systems are only as good as the data they are trained on. In the context of cybersecurity, this means that if a system has been trained on unrepresentative or biased data, it may make incorrect predictions or fail to detect certain threats.

6. Displacement of human judgement:

Despite advances in AI, human judgement and expertise remain crucial in cybersecurity. The cybersecurity team has an intuitive and contextual understanding of the systems and networks they manage, which is extraordinarily important in identifying and responding to threats that might go unnoticed by an automated system.

7. Cost of maintenance and updating:

While automation may seem cost-effective in the short term, maintaining and upgrading AI systems to ensure they remain effective in the face of emerging threats can require significant investments in time and resources.

So, in **conclusion**, while AI and automation can offer revolutionary capabilities in the field of cyber security, it is essential to approach these systems with a balanced approach, and they should be seen as one tool in a broader defence arsenal, complementing, not replacing, other traditional methods and techniques.

Combining human expertise with AI capabilities and compliance with applicable legal standards is the best defence against evolving cyber threats.

Attackers are increasingly aware of how AI-based systems work and are developing specific techniques, such as adversarial attacks, to deceive or circumvent these systems

5.3 False positives and false negatives

False positives and **false negatives** are a crucial challenge in any detection or classification system, and their prevalence in Artificial Intelligence (AI) or Machine Learning (ML)-based systems can have serious consequences in the field of cybersecurity.

A False Positive (FP) is said to occur when the system misidentifies benign activity as malicious. In security terms, this could be legitimate software mistakenly identified as malware.

A False Negative (FN) is said to occur when the system fails to detect malicious activity, misclassifying it as benign. For example, a real malware that was not detected by the security system.

Both types of false positives and false negatives have important implications for cyber security:

IMPLICATIONS OF FALSE POSITIVES	IMPLICATIONS OF FALSE NEGATIVES
Unnecessary disruptions: False positives can lead to blocking or stopping legitimate applications and processes, causing disruptions to normal business operations.	Undetected security breaches: A false negative allows real threats to bypass defences, which can lead to data breaches, system compromise or other cyber damage.
Wear and tear on security equipment: A high number of false positives can consume significant resources, as security personnel have to review and verify each alert.	Ill-founded confidence: Believing that a system is secure when in reality there are active threats can lead to complacency and a lack of preparedness for possible incidents.
Desensitisation: If security alerts are commonly perceived as false alarms, staff may begin to ignore them, which could lead to the omission of truly critical alerts.	

5. Challenges and limitations of AI in Cybersecurity

At this point, it is worth recalling two of the most significant challenges posed by the use of AI and ML in cybersecurity. Firstly, **data quality**, knowing, as we have said, that the accuracy of ML models is directly linked to the quality of the data with which they are trained, so that unrepresentative or unbalanced data can lead to higher rates of false positives and negatives. Secondly, regarding **complex models**, some advanced ML techniques, especially in deep learning, can act as "black boxes", making it difficult to understand why certain decisions are made and, therefore, complicating the task of adjusting the model to reduce these errors.

In the face of these realities, it is therefore necessary to develop **AI risk mitigation strategies**, in particular:

- ▶ **Constant training:** ML models must be regularly retrained and adjusted with updated data to improve their accuracy.
- ▶ **Incorporating feedback:** By integrating human feedback, systems can learn from errors and adjust their detection criteria.
- ▶ **Combination of techniques:** Using a hybrid approach combining different detection techniques can help reduce both false positives and false negatives.

In **conclusion**, as we have seen, false positives and negatives present significant challenges in cyber security, especially when using AI-based systems. While it is difficult to eliminate them completely, proper understanding and effective management of these errors can minimise their impact and ensure a more robust cyber defence.

False positives and negatives present significant challenges in cyber security, especially when using AI-based systems.

5.4 Privacy and ethics in the application of AI

Privacy and ethics in the application of Artificial Intelligence is an increasingly important issue, also in the context of cybersecurity, where data and personal information may be at stake. AI-powered security solutions have the potential to be extremely effective, but also present concerns about how data is collected, stored and used.

Among the aforementioned **problems**, we can quote the following aspects::

IN RELATION TO...	PROBLEMS CAN ARISE FROM...
...data collection:	Oversizing: To train and operate, AI systems require large data sets. In the process, there is the potential for more data to be collected than necessary, which may invade users' privacy. Consent: Data is often collected without the user's knowledge or consent, raising ethical and legal concerns.
... storage and use of data:	Data security: By storing large data sets, organisations become attractive targets for cybercriminals. A security breach could expose personal and/or confidential information. Profiling: With sufficient data, AI can be used to profile individuals based on their online behaviour, which can lead to biased decisions or discrimination.
...transparency and decision-making:	Black box' decisions: Many AI models, especially those based on deep learning, do not provide clear visibility of how they make decisions. This can lead to lack of trust and difficulties in verifying the fairness or appropriateness of these decisions. Bias and fairness: if the data used to train AI models is biased, the decisions made by the model will be biased too. This can reinforce stereotypes or lead to discrimination.

5. Challenges and limitations of AI in Cybersecurity

... monitoring and supervision:	Potential abuse: AI-based cybersecurity solutions that monitor networks and systems for threats can also be used to monitor user behaviour for malicious or invasive purposes.
... accountability and responsibility:	Lack of accountability: Determining responsibility for failures or errors in an AI-based system can be complicated, especially if it is unclear how the system made a particular decision.
... regulations and ethical guidelines:	Need for regulatory frameworks: To ensure that ethical concerns are addressed, clear guidelines and regulations are essential to guide the development and application of AI solutions in cybersecurity.

The ethics and regulation around AI and cybersecurity are evolving rapidly as technologies advance and the potential problems and consequences become more apparent.

As far as the European Union is concerned, there are two most significant regulatory frameworks:

► General Data Protection Regulation (GDPR) of the European Union³⁷:

Although not specifically designed for AI, the GDPR has set important standards for data privacy and the rights of individuals, such as the right to be forgotten and transparency in data processing. These principles also apply when using AI in cybersecurity.

The GDPR is a fundamental regulation for the privacy and data protection of all individuals within the European Union (EU). It entered into force on 25 May 2018. These are its essential aspects::

- **Territorial scope:** The GDPR applies not only to organisations located within the EU, but also to organisations located outside the EU if they offer goods or services to individuals in the EU or monitor the behaviour of individuals in the EU.
- **Consent:** Organisations can no longer use long and difficult-to-understand terms and conditions. The request for consent must be given in an easily accessible and understandable form. Moreover, it must be as easy to withdraw consent as it is to give it.

³⁷ REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of individuals regarding the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation).

5. Challenges and limitations of AI in Cybersecurity

► Rights of the data subject:

- **Right of access:** Individuals have the right to know whether their personal data is being processed and, if so, to access that data.
- **Right to rectification:** Individuals have the right to correct inaccurate data.
- **Right to be forgotten:** Also known as the right to erasure, this allows individuals to request the deletion of their data.
- **Right to portability:** Individuals can obtain and re-use their personal data across different services.
- **Right to restriction of processing:** Individuals may request that their data not be processed except for specified, defined, adequate and lawful purposes.
- **Right to object:** Individuals have the right to object to the processing of their data in certain circumstances.

- **Data breach notification:** In case of a data breach, organisations must notify the relevant data protection authorities within 72 hours, unless the breach does not pose a risk to the rights and freedoms of individuals. If the breach poses a high risk to the rights and freedoms of individuals, they must also be notified.

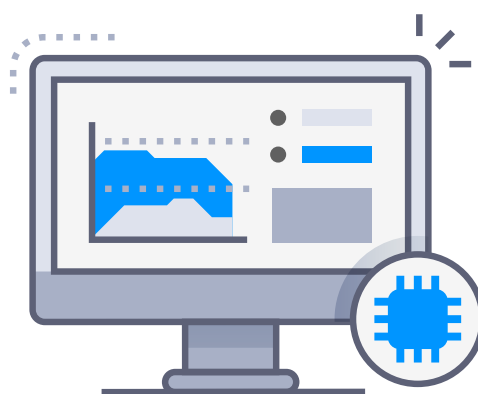
- **Responsibility of the Controller (RT) and Processor (ET):** Sets out the responsibility of RTs and ETs to ensure compliance with the GDPR, including the need to keep detailed records of data processing activities.

- **Data protection by design and by default:** Organisations should consider data protection in the design of new systems, processes or products and also ensure that, by default, only the data necessary for each specific use are processed.

- **Data Protection Officers (DPOs):** Organisations must appoint a DPO if they belong to certain groups of entities or carry out certain types of data processing, such as large-scale processing of sensitive data.

- **International transfers:** Stricter conditions are set for the transfer of personal data outside the EU.

Individuals have the right to know whether their personal data is being processed and, if so, to access that data



5. Challenges and limitations of AI in Cybersecurity

- **Penalties:** Organisations can be fined up to 4% of their global annual turnover or €20 million (whichever is higher) for serious breaches. There is a graduated system of fines for less serious breaches.

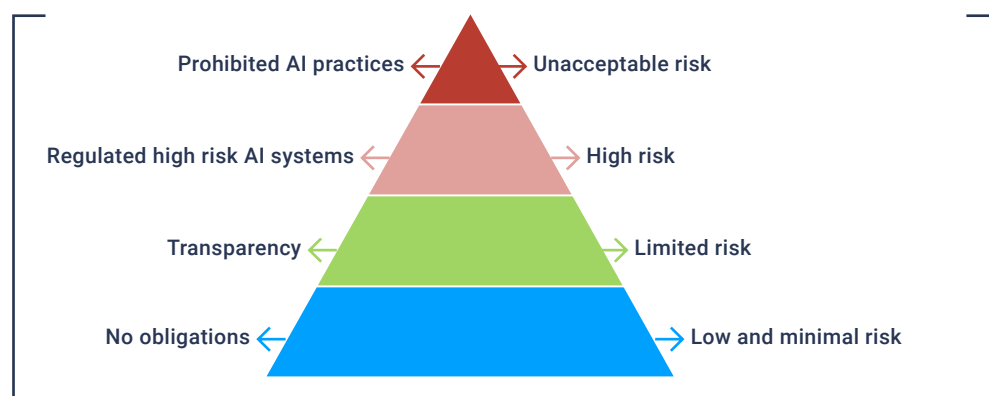
► **European Commission's proposal for AI regulation (2021)**³⁸:

The European Commission presented in April 2021 a proposal for an EU regulatory framework on artificial intelligence (AI)³⁹. The draft law on AI is the first attempt to enact a horizontal regulation on AI. The proposed legal framework focuses on the specific use of AI systems and the associated risks.

In the text, the Commission proposes to establish a technology-neutral definition of AI systems in EU legislation and to establish a classification for AI systems with different requirements and obligations adapted to a "risk-based approach".

Thus, some AI systems presenting "unacceptable" risks would be banned; a wide range of "high risk" AI systems would be authorised, but they would be subject to several requirements and obligations for access to the EU market. AI systems presenting only "limited risk" would be subject to very light transparency obligations.

PIRÁMIDE DE RIESGOS



³⁸ Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL ESTABLISHING HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE LAW) AND AMENDING CERTAIN LEGISLATIVE ACTS OF THE UNION (Brussels, 21.4.2021).

³⁹ Source: European Parliament. Artificial intelligence act. Briefing. EU Legislation in Progress. (2023).

5. Challenges and limitations of AI in Cybersecurity

The Council adopted the general position of the EU Member States in December 2021. The Parliament voted its position in June 2023.

At the time of writing, EU legislators are now starting negotiations to finalise the new legislation, with substantial amendments to the Commission's proposal, including revising the definition of AI systems, expanding the list of prohibited AI systems, and imposing obligations on general purpose AI and generative AI models such as ChatGPT.

The Proposal for a Regulation on Artificial Intelligence (AI) marks a significant step towards the regulation of AI applications in the European Union. These are its essential elements:

- ▶ **Objective:** The proposal aims to ensure that AI is used in a way that is safe and respects the fundamental rights of EU citizens.
- ▶ **Risk classification:** AI applications are classified according to the level of risk they present:
 - **Unacceptable risk:** Some practices would be completely prohibited due to a clear potential to harm people's rights. This includes, for example, AI systems that distort human behaviour.
 - **High risk:** Applications in critical areas, such as biometric identification systems and critical infrastructure systems. These systems will be subject to strict regulations and will require assessment prior to implementation, in some cases being prohibited altogether.
 - **Limited risk:** Applications must follow specific transparency requirements. For example, chatbots should be declared as such so that users know they are interacting with a machine.
- ▶ **Transparency:** The proposal emphasises transparency in the use of AI systems, especially in areas such as deepfakes or interactions with chatbots.
- ▶ **Establishment of a European AI Committee:** It is proposed that a committee be set up to help implement and update the Regulation.

EU legislators are now starting negotiations to finalise the new legislation, with substantial amendments to the Commission's proposal, including revising the definition of AI systems, expanding the list of prohibited AI systems, and imposing obligations on general purpose AI and generative AI models such as ChatGPT

5. Challenges and limitations of AI in Cybersecurity

- ▶ **Penalties:** The proposal provides for substantial penalties for non-compliant companies, including fines of up to 6% of their annual global turnover.
- ▶ **Applicability:** The Regulation will apply not only to AI system providers established in the EU, but also to providers offering their systems on the EU market.
- ▶ **Innovation and support:** While the proposal has a regulatory focus, it also highlights the importance of fostering innovation in the field of AI and supporting the development of AI capabilities in the EU.

Ethical frameworks

We can mention the following:

1. **Asilomar Principles on Artificial Intelligence**⁴⁰:

These principles, established at the 2017 AI conference, cover areas such as research to make AI safe, ethics in its application and the need for it to benefit all.

2. **OpenAI AI Principles**⁴¹:

This AI research organisation has established principles that seek to ensure that AI benefits all of humanity, prioritising security and long-term cooperation.

3. **Google's AI Principles**⁴²:

Although they come from a specific company, these principles have been influential. They address security, fairness, transparency and accountability, among other aspects.

4. **IEEE AI Ethics**⁴³:

The IEEE, one of the world's largest professional organisations for the advancement of technology, has established ethical standards for AI and robotics that focus on incorporating human values into design and operation.

In **conclusion, therefore**, while regulation and ethical frameworks are essential to guide the application of AI in cybersecurity, it is crucial that these guidelines are kept up to date and flexible to adapt to the rapidly evolving technology.

40 <https://futureoflife.org/open-letter/ai-principles/>

41 <https://openai.com/charter>

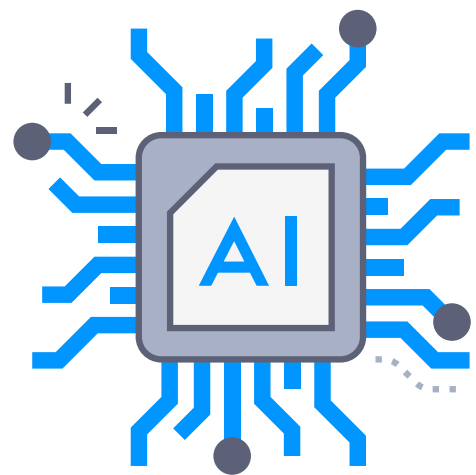
42 <https://ai.google/responsibility/principles/>

43 <https://standards.ieee.org/news/get-program-ai-ethics/>

6. Future of AI in Cybersecurity

As we venture into the coming decades, the integration of AI into cyber security will become even deeper and more complex, promising significant transformations in how we detect, respond to and prevent cyber threats.

In this section, we will explore the projections and trends that will define the future of AI in cybersecurity. From autonomous defence systems and continuous learning to ethical challenges and the need for robust regulatory frameworks, we will address the expectations and concerns surrounding this technological horizon. In addition, we will highlight how current innovations can outline the contours of future solutions and how the global community can prepare and adapt to these impending changes.



6.1 Emerging trends

The following is a summary of some of the issues that are currently being worked on and that could shape the future of AI applied to cybersecurity.

Autonomous cyber self-defence

Autonomous cyber self-defence refers to the use of advanced technologies, particularly artificial intelligence and machine learning, to enable computer systems and networks to automatically detect, respond to and mitigate threats in real time, without human intervention.

Its characteristics are:

- 1. Proactive detection:** Traditionally, many security systems operate in a reactive mode, responding to threats after they occur. Autonomous self-defence, by contrast, focuses on detecting patterns and anomalies in real time, enabling an almost immediate response.
- 2. Immediate response and containment:** Once a threat is detected, autonomous systems can take action to contain it, which may include isolating a compromised device, blocking a suspicious IP address or limiting access to certain parts of the network.
- 3. Adaptability:** Given the ever-changing nature of cyber threats, autonomous cyber self-defence is designed to constantly learn and adapt. This means that with each detected threat, the system becomes smarter and more effective in its response.
- 4. Reduced human burden:** With automated response, the need for constant human intervention is reduced, allowing security teams to focus on more complex threats or long-term cyber security strategy.



6. Future of AI in Cybersecurity

5. **Challenges:** Despite its advantages, autonomous cyber self-defence is not without its challenges, which include the possibility of excessive or erroneous responses, complexity in implementation and maintenance, and dependence on AI, which can be susceptible to specific attacks, such as adversarial attacks.
6. **Real-life applications:** There are solutions on the market today that offer autonomous response capabilities, especially in the field of endpoint detection and response (EDR). These solutions can identify malicious behaviour on network devices and take immediate action to neutralise the threat.

EXAMPLE

Darktrace

(<https://www.darktrace.com/>)

This company has its 'Enterprise Immune System', a solution that uses machine learning algorithms to detect, respond to and mitigate cyber threats in real time. Its system learns and understands the normal 'life pattern' of each user and device on the network, enabling it to detect significant deviations that indicate potential threats. In addition, its Darktrace Antigena product acts as a 'digital antibody', making autonomous decisions on how to respond to specific threats without human intervention.
<https://es.darktrace.com/resources/autonomous-response-darktrace-antigena>

Federated learning

Federated learning is an approach to training machine learning models in which multiple devices or servers retain their data locally and share only model updates with a central server, rather than sharing the data itself.

Its essential characteristics are:

- Each device trains a model locally, using its own data.
- Once a device has processed its local batch of data and updated the model, it sends only the updates or the model summary to the central server.
- The central server aggregates updates from all devices to form an up-to-date global model.
- This global model is sent back to all devices for the next round of training.
- This process is repeated until the model converges or satisfies certain stopping criteria.

As we say, one of the advantages of this model is that since the raw data never leaves the local device, there is less risk of exposure, which is especially useful for sensitive or personal data. In addition, the bandwidth required is reduced, as only model updates are shared, which are typically much smaller in size than the entire dataset, which is also very suitable for scenarios where data is distributed, such as mobile devices or geographically dispersed locations.

6. Future of AI in Cybersecurity

Regarding its **applications in cybersecurity**, we can highlight the following two:

- ▶ **Threat detection in distributed networks:** By allowing each node or device in the network to learn locally about threats and share its updates with a central server, a global detection model can be built without compromising data privacy at each node.
- ▶ **Real-time model updates:** Devices in a network can quickly adapt to new threats by learning locally, and then updating a global model.

Explainable AI systems (XAI)

As AI plays more critical roles in cybersecurity, it is essential that all parties involved (especially cybersecurity teams) can understand and trust the decisions made by these systems. The acronym XAI refers to methods and techniques in AI research that make the results of algorithms understandable to humans.

With the popularity of deep learning models, such as neural networks, AI has achieved significant levels of accuracy in many tasks. However, as we have been repeating, these models often act as "black boxes", where even experts have difficulty understanding why a specific decision was made. This lack of transparency can be problematic, especially in fields such as medicine, law and banking, where incorrect decisions can have serious consequences and where justification is required, even as a legal imperative.

XAI models can be developed on the basis of different approaches:

1. **Local vs. global interpretability:** Interpretability can focus on understanding individual decisions (local) or on understanding how the model generally works (global).
2. **Intrinsically Interpretable Models:** These models, such as decision trees or linear regression, are naturally explainable. However, they may not be as accurate as complex models.

EXAMPLES

TensorFlow Federated (TFF):

is an open-source platform developed by Google that allows developers to use APIs to implement federated learning. It is built on top of TensorFlow and provides tools to simulate federated learning on distributed data.
(<https://www.tensorflow.org/federated?hl=es-419>)

PySyft:

is a flexible extension of PyTorch for federated learning and other privacy techniques in machine learning. It focuses on decentralisation and offers tools for secure multiparty computation, among others.
(<https://github.com/OpenMined/PySyft>)

Federated AI Technology Enabler (FATE):

is an open-source platform that provides a secure and convenient framework for collaborative training and federated learning. It was an initiative of WeBank and has attracted contributions from many other companies and organisations.
(<https://fate.fedai.org/>)

While these tools offer frameworks and libraries for federated learning, it is important to note that many large technology companies, such as **Google** and **Apple**, are already implementing federated learning in some of their products to enhance user privacy. A classic example is text prediction on smartphone keyboards, where the model is trained locally on the user's device, based on their inputs, without sending the actual data to central servers.

6. Future of AI in Cybersecurity

3. **Post-hoc methods:** These methods are applied after the model has been trained. They can be visualisations, such as heat maps, or techniques that decompose the model's decisions, such as LIME (Local Interpretable Model-agnostic Explanations) or SHAP (SHapley Additive exPlanations).
4. **Attribute Decomposition Techniques:** These techniques attempt to explain the contribution of each characteristic to a specific decision, giving an idea of which characteristics are the most influential.

Be that as it may, the benefits of the XAI approach are clear: they underpin **trust** (when users, especially those who are not AI experts, understand how a system works, they are more likely to trust it); they help maintain **accountability** (XAI can help ensure that AI systems act responsibly and fairly, reducing biases and errors); facilitates model **improvement and refinement** (by understanding how a model makes decisions, it is easier to identify and correct errors or inaccuracies); and facilitates regulatory compliance (since in some jurisdictions, such as the EU, automated decision-making systems are being required to be transparent and able to justify their decisions).

Naturally, the application of XAI models is not without its challenges, such as the **accuracy-interpretability trade-off** (knowing that there is often a trade-off between the accuracy of the model and its interpretability, simpler models may be easier to understand but less accurate); **subjectivity** (since "explainability" can be subjective, i.e. what is clear and understandable to one technical expert may not be clear and understandable to another or to a non-specialist); or **generalisability** (since explanations generated for a specific instance may not generalise well to other instances).

EXAMPLES

LIME (Local Interpretable Model-agnostic Explanations):

explain the predictions of any classifier or regressor in a way that is understandable to humans. It works by creating an interpretable model that is locally faithful to the predictions of the original model.

(<https://github.com/marcotcr/lime>)

SHAP (SHapley Additive exPlanations):

is based on game theory to explain the output of any machine model. It is a unified measure of the importance of features.

(<https://github.com/slundberg/shap>)

DeepLIFT (Deep Learning Important Features):

presents an approach for decomposing the outputs of neural networks and calculating the importance of each input for the output. It is especially useful for deep neural networks.

(<https://github.com/kundajelab/deeplift>)

AI Explainability 360:

is a toolkit that includes algorithms, libraries and tutorials to help developers understand, explain and visualise decisions made by AI models.

(<https://aix360.mybluemix.net/>)

InterpretML:

is an open-source library from Microsoft for interpreting machine models. It provides a variety of techniques and tools for model interpretation.

(<https://interpret.ml/>)

Adoption of Blockchain for security

Although blockchain is best known for its use in cryptocurrencies, it has applications in cybersecurity, especially in identity management and ensuring data integrity.

The ways in which blockchain is playing a role in cyber security are summarised below:

- 1. Data integrity and authentication:** Blockchain provides an immutable and transparent record of data. Once a block is added to the chain, it cannot be altered without altering all subsequent blocks, which is considered extraordinarily difficult due to the decentralised nature of the blockchain network. This ensures data integrity and prevents tampering.
- 2. Decentralisation:** Traditional cybersecurity relies on centralised servers, which are vulnerable points of attack. Blockchain is inherently decentralised, meaning that there is no single point of potential compromise, which, depending on the security adopted for each node, can be an advantage or a source of potential risk.
- 3. Secure identity:** Blockchain-based systems can provide digital identity solutions where users' identities are verified and stored on the blockchain.
- 4. Secure communications:** Blockchain solutions can ensure secure, authenticated communications between devices in the Internet of Things (IoT). These devices are often vulnerable to attacks, but with blockchain-based identity management, they could be validated and communicate more securely.
- 5. Auditing and traceability:** Blockchain provides a clear and verifiable trace of all transactions. This is invaluable for auditing operations and facilitates transparency and accountability.
- 6. Censorship resistance and availability:** Due to their decentralised nature, blockchain networks are resistant to censorship and disruption. It is difficult to shut down or censor a blockchain network without the consensus of most of its participants.
- 7. Smart Contracts for secure automation:** Smart contracts are autonomous programmes that run on blockchain when certain pre-set conditions are met and can be used to automate and validate agreements and transactions without intermediaries, thus reducing the possibility of fraud or malicious intervention.

Although blockchain is best known for its use in cryptocurrencies, it has applications in cybersecurity, especially in identity management and ensuring data integrity

6. Future of AI in Cybersecurity

Despite the benefits of blockchain technology for cybersecurity, there are also challenges. For example, although the blockchain is immutable and transactions cannot be altered once validated, if an attacker manages to gain control of most of the network (a 51% attack), they could potentially validate fraudulent transactions. Furthermore, like any emerging technology, the practical implementation of blockchain in cybersecurity is still developing, and organisations should be cautious and diligent in adopting it.

AI models based on user behaviour

Instead of relying solely on passwords or biometric data, AI could analyse continuous behavioural patterns (such as the way someone types or moves their mouse) to authenticate users and detect anomalous behaviour.

AI models based on user behaviour are an emerging trend in cybersecurity and represent one of the most advanced ways to detect anomalies and suspicious activity in a system. As we say, these models are trained to learn normal patterns of user behaviour in order to identify any deviation from those patterns as potentially suspicious activity.

Some of the most common **applications** of this type of model would be:

- 1. Fraud detection:** In the financial sector, for example, if a user transacts large sums of money suddenly or in patterns that do not match their historical behaviour, AI can generate an alert.
- 2. Access control and authentication:** If a user's login behaviour suddenly changes (e.g., they log in at unusual times or from unknown geographic locations), it could be a sign that someone else is using their credentials.
- 3. Protection against insider threats:** Disgruntled or malicious employees can pose threats to organisations. If they start accessing files or systems they do not normally use, this can be detected by behaviour-based systems.

EXAMPLES

Guardtime:

Uses blockchain technology to ensure data integrity and authenticity.
(<https://www.guardtime.com/>)

Civic:

Is a secure identity solution based on blockchain. It provides businesses and individuals with tools to control and protect identities. Through its decentralised platform, Civic enables authentication without the need for traditional passwords.
(<https://www.civic.com/>)

REMME:

Is a solution that aims to eliminate phishing attacks, passwords and certificates. It uses blockchain to authenticate users and devices instead of a password.
(<https://remme.io/>)

Chain of Things (CoT):

Researches and develops applications that combine blockchain with the Internet of Things (IoT). This has applications in areas such as energy, transport and logistics, where the integrity and security of data collected by IoT devices is critical.
(<https://www.chainofthings.com/>)

6. Future of AI in Cybersecurity

There are undoubted **advantages** to using such models, including: **personalisation** (since they are based on individual user behaviour, they are highly customised and adaptive); **proactive detection** (they can detect threats in real time, allowing for a faster response); and **reduction of false positives** (since they are more attuned to actual user behaviour, they tend to generate fewer alerts for legitimate activity that deviates slightly from the norm).

However, the use of these models also entails certain **challenges**, such as: **requiring time to learn** (since, like any machine learning system, these models need time to learn user behaviour patterns); **changes in user behaviour** (so that if a user changes roles or responsibilities, their behaviour may change, which can lead to false positives until the system adapts) or **privacy** (since these systems collect and analyse a lot of information about user behaviour, which raises concerns about privacy and how this data is handled and stored).

Quantum AI

As quantum computing becomes a more everyday reality, we are likely to see specific developments in quantum AI. These systems may be able to process information at exponentially faster speeds and handle security problems that are currently intractable for conventional systems. Quantum AI" is an emerging field that combines techniques and concepts from artificial intelligence with quantum mechanics and quantum computing. Although still in its early stages, this field promises to revolutionise the capabilities and performance of AI systems.

As is well known, quantum computing is based on quantum mechanics, a theory of physics that describes how subatomic particles work. Unlike classical computing, which uses bits (0s and 1s) to represent and process information, quantum computing uses "qubits". Qubits have the amazing ability to represent multiple states at once (superposition) and to be "entangled", meaning that the state of one qubit can depend on the state of another, regardless of the distance separating them.

EXAMPLES

Darktrace:

Previously reviewed, uses what it calls an "Enterprise Immune System" to learn from the normal 'life pattern' of each user and device on a network and then identify anomalous behaviour. (<https://www.darktrace.com>)

Cylance:

As noted above, it uses AI to deliver behaviour-based endpoint threat prevention. Its platform focuses on stopping malware, ransomware and other threats based on the identification of suspicious behaviour rather than known virus signatures. (<https://www.cylance.com>)

Exabeam:

Is a security information and event management (SIEM) platform that uses machine learning to analyse user behaviour and detect threats. (<https://www.exabeam.com>)

UserInsight de Rapid7:

Focuses specifically on detecting anomalous behaviour of users and attackers within a network. It can identify when a user's credentials have been compromised and are being used by an attacker. (<https://www.rapid7.com>)

These tools combine traditional security techniques with advanced machine learning to detect threats in real time based on user behaviour.

6. Future of AI in Cybersecurity

The use of quantum AI models would have the following **advantages**:

- ▶ **Speed and Scalability:** Quantum algorithms can perform certain operations much faster than their classical counterparts. In theory, quantum AI could tackle problems that are currently intractable for classical computers due to their complexity.
- ▶ **Optimisation:** Problems such as combinatorial optimisation, which are crucial in fields such as logistics, economics and many AI applications, could greatly benefit from the ability of quantum computers to explore multiple solutions simultaneously.
- ▶ **Deep learning and model training:** Quantum computers have the potential to significantly speed up the training of complex AI models, which could revolutionise areas such as deep learning.

However, as always, the use of these models poses certain **challenges**:

- ▶ **Hardware:** Although significant progress has been made in building quantum computers, we still face challenges in terms of stability, coherence and scalability.
- ▶ **Quantum algorithms:** The adaptation of classical AI algorithms to quantum computing remains an active area of research. Not every problem will benefit from a quantum solution.
- ▶ **Quantum-Classical Interaction:** Integrating quantum computing systems with classical infrastructures and algorithms is a considerable challenge.

IBM Q Experience:

IBM has been a leader in the field of quantum computing, and through IBM Q, offers access to real quantum computers for researchers and developers to experiment and develop quantum algorithms. While not exclusively for AI, it is a platform that could be used to experiment with quantum AI algorithms.
(<https://quantum-computing.ibm.com/>)

D-Wave Systems:

D-Wave is known for its advanced quantum systems, different from gate-based computer models. They have worked on optimisation and machine learning using their quantum systems.
(<https://www.dwavesys.com/>).

Google AI Quantum:

Google has been actively researching the field of quantum computing and has made significant advances. Although they focus on many aspects

EXAMPLES

of quantum computing, artificial intelligence is one of the applications they are exploring
(<https://quantumai.google/>)

Rigetti Computing:

A start-up company that focuses on building quantum computers and also provides a cloud platform for developers and scientists to experiment with quantum computing. While the platform is not exclusively dedicated to quantum AI, it does offer the potential to explore applications in that domain.
(<https://www.rigetti.com/>)

It is important to note that many of these tools and systems are designed to be platforms for quantum computing in general, not specifically for quantum AI. However, given that quantum computing has potential applications in AI, these platforms can play a crucial role in the future development of quantum AI.

Human-machine collaboration

Despite advances in AI, humans will remain essential in cybersecurity. The emerging trend will be for systems where humans and machines work together, complementing each other's strengths.

Indeed, human-machine collaboration is an approach that leverages the unique capabilities of both humans and machines, especially in the context of artificial intelligence, to improve decision-making, efficiency and overall outcomes. The main idea behind this collaboration is that while machines are excellent at calculating, analysing and processing large volumes of data, humans possess intuition, contextual understanding, empathy and creativity.

This so-called man-machine symbiosis has some **key aspects** that need to be taken into account:

- 1. Complementarity:** AI and humans have complementary points. For example, AI can handle large amounts of data, perform repetitive operations and complex calculations at speeds that humans cannot achieve. However, humans bring creativity, judgement and experience based on intuition.
- 2. Natural interaction:** The development of intuitive and natural interfaces, such as voice conversation or gestural interaction, allows for more seamless collaboration between machines and humans. These systems rely on natural language processing, speech recognition, and gesture recognition to understand and respond appropriately to human interactions.
- 3. Two-way learning:** While the machine learns from human behaviour to improve its predictions and actions, the human also learns to trust and understand how the machine works, establishing a constant feedback and improvement cycle.
- 4. Transparency and explainability:** For humans to trust decisions made or suggested by machines, it is essential that machines can provide understandable explanations for their decisions.

The emerging trend will be for systems where humans and machines work together, complementing each other's strengths

6. Future of AI in Cybersecurity

- 5. Human intervention:** In many collaborative systems, a mechanism is built into the system that allows for human intervention in certain scenarios. For example, in an autonomous driving system, there may be situations where the system asks the human driver to take control.

EXAMPLES

IBM Watson:

One of the most notable uses of Watson is in the medical field. Watson Health helps healthcare professionals make informed decisions about patient treatment by analysing large amounts of medical data and scientific literature. Although Watson provides recommendations, it is always the doctor who makes the final decision. (<https://www.merative.com/>)

Google's DeepMind AlphaGo:

Although best known for beating human champions at the game of Go, the real achievement here is how humans and machine learned from each other, enabling Go players to study AlphaGo's moves to improve their own strategies. (<https://www.deepmind.com/research/highlighted-research/alphago>)

KUKA LBR iiwa:

A collaborative robot designed to work alongside humans in an industrial environment. These "cobots" are touch-sensitive and can stop or slow down if they detect an object or person in their path, allowing humans and

robots to work side-by-side on the same task.

(<https://www.kuka.com/en-us/products/robotics-systems/industrial-robots/lbr-iiwa>)

OpenAI Codex:

An AI-based platform that helps programmers write code. It can generate code snippets based on user-supplied descriptions, acting as a programming assistant. (<https://openai.com/blog/openai-codex>)

Adobe Sensei:

Integrated into Adobe's tools, Sensei uses AI and machine learning to assist with creative tasks, from photo and video editing to design and illustration. Although it automates some functions, the creative is in ultimate control and uses the tool to improve and speed up the design process. (<https://www.adobe.com/sensei.html>)

These examples represent a variety of applications in different industries, showing how AI-based tools can work in collaboration with humans to improve efficiency, accuracy and creativity.

6. Future of AI in Cybersecurity

AI at the edge (Edge AI)

Instead of relying on centralised data centres, AI could be processed on the device (such as mobile phones, IoT, etc.). This has significant implications for cybersecurity, enabling faster responses and reducing the risks associated with data transmission.

The use of this type of model has the following **advantages**:

1. **Reduced latency:** By processing data directly on a device, the need to send that data to a centralised server for processing is eliminated, which in turn reduces latency. This is especially crucial in real-time applications such as autonomous vehicles.
2. **Privacy and security:** Keeping data processing on the device can minimise the security risks associated with data transmission and ensure that sensitive data does not leave the device.
3. **Offline operation:** Devices with Edge AI capabilities can operate and make decisions without the need for an active connection to the cloud or central server.
4. **Bandwidth efficiency:** By reducing the need to transmit large amounts of data to the cloud, bandwidth is saved.
5. **Power consumption:** While Edge AI devices may require more power than devices without AI capabilities, they often consume less power than is needed to constantly communicate with a central server.

Applications of this model would be useful in several contexts: **autonomous vehicles** (since vehicles need to rapidly process huge amounts of data from their sensors to navigate safely, latency in decision-making can have serious consequences); in **smart home appliances** (refrigerators, hoovers, ovens and other devices that use AI for optimisation and decision-making); in **wearable health devices** (heart rate monitors, glucose devices and other medical devices that need to process real-time data); in security cameras (that detect anomalous activity or recognise faces and make decisions based on such data) or in drones (used for navigation, object detection and real-time decisions).

Instead of relying on centralised data centres, AI could be processed on the device (such as mobile phones, IoT, etc.)



6. Future of AI in Cybersecurity

On the other side of the scale, the **challenges** involved in this type of model are:

- ▶ **Hardware limitations:** Although Edge devices are evolving rapidly, there are still limitations in terms of processing power, memory and storage compared to centralised data centres.
- ▶ **Management and updating:** Maintaining and updating AI models on numerous dispersed devices can be a challenge.
- ▶ **Model development:** Models often need to be optimised and adapted to be small and efficient enough to run on Edge devices without sacrificing excessive accuracy or capacity.

All these emerging trends in AI for cyber security represent only a small part of what is undoubtedly to come.

EXAMPLES

TensorFlow Lite:

A Google solution designed to bring machine learning models to mobile and Edge devices. It provides tools to convert and optimise standard TensorFlow models to be efficient on these devices.

(<https://www.tensorflow.org/lite>)

ONNX Runtime:

Comprises a machine learning inference library for ONNX (Open Neural Network Exchange) models. ONNX Runtime is lightweight and can be used on different platforms, including Edge devices.

(<https://onnxruntime.ai/>)

Intel OpenVINO

(Open Visual Inferencing & Neural Network Optimization) Toolkit:

Intel's toolkit for accelerating and optimising AI models, specifically designed to run efficiently on Intel hardware, including chips designed for Edge devices.

(<https://software.intel.com/content/www/us/en/develop/tools/openvino-toolkit.html>)

NVIDIA Jetson Platform:

A series of NVIDIA-designed embedded systems that integrate a GPU and are optimised for AI Edge inference tasks.

(<https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/>)

Azure IoT Edge:

A Microsoft solution that enables cloud-based workloads to be deployed directly to Internet of Things (IoT) and other edge devices, including machine learning models.

(<https://azure.microsoft.com/en-us/services/iot-edge/>)

These tools and platforms represent only a fraction of the growing AI Edge industry.

6.2 Ongoing investigations

The field of cybersecurity is fertile ground for research, especially with the integration of emerging technologies such as Artificial Intelligence. This section explores the most recent areas of research.

Deep learning against advanced network attacks:	Network attacks are evolving to become more sophisticated and harder to detect. Current research is focused on how to use deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to identify hidden patterns in network traffic that may indicate an attack.
Adversarial techniques to strengthen AI systems:	The idea is to use adversarial techniques (attacks specifically designed to fool AI models) in a controlled environment to improve the robustness of AI models in cybersecurity. This involves training models with perturbed data to make them more resistant to adversarial attacks in real-world scenarios.
AI supply chain security:	As AI becomes an essential part of many systems, ensuring the integrity of the entire AI supply chain -from training to deployment- has become crucial. Research focuses on how these systems can be infiltrated and compromised and how to defend against such vulnerabilities.
Detection of deepfakes and, in general, synthetic content:	With the increasing ability of AI tools to create realistic synthetic content (such as audio or video deepfakes, for example), research into the automatic detection of such content has intensified. This has direct applications in cybersecurity, particularly in areas such as authentication and protection against misinformation.

6. Future of AI in Cybersecurity

Incident response automation:	Rather than simply detecting threats, AI systems are being developed that can automatically respond to security incidents, making real-time decisions on how to mitigate or neutralise a threat.
Understanding the "black box" of AI:	A major area of research is explainable AI (XAI). In the context of cybersecurity, it is essential to understand why an AI system makes a particular decision, especially if it is related to threat detection or incident response.
IA for insider threat protection:	Insider threats, whether malicious or inadvertent, remain a major challenge in cybersecurity. AI can play a role in monitoring user behaviour to identify suspicious activity or deviations from the norm.
Safe training methodologies:	Investigate ways to train AI models without exposing sensitive data, such as federated learning or differential privacy learning.

These research areas show the continuous evolution and adaptation of AI in the cybersecurity domain. As threats change and become more sophisticated, it is essential that research in this field keeps pace to provide effective and proactive solutions.

EXAMPLES

Deep learning against advanced network attacks:

"DeepDefense", a platform that uses deep learning techniques to detect attacks in real-time.
(<https://ieeexplore.ieee.org/document/7946998>)

Adversarial techniques for hardening AI systems:

Google's "CleverHans" project.
([GitHub.com/tensorflow/cleverhans](https://github.com/tensorflow/cleverhans))

AI supply chain security

MITRE Corporation has been doing research in related areas.
(mitre.org).

Detection of deepfakes and synthetic content:

"Deepware Scanner" by the company Cyabra.
(cyabra.com)

Incident response automation:

Palo Alto Networks' "Cortex XSOAR".
(paloaltonetworks.com/cortex/xsoar)

Understanding the AI "black box":

LIME (Local Interpretable Model-Agnostic Explanations).
(github.com/marcotcr/lime)

AI for Insider Threat Protection:

Varonis Insider Threat Solution.
(varonis.com/solutions/insider-threat-detection)

Safe training methodologies:

"TensorFlow Federated" for federated learning.
(tensorflow.org/federated)

6.3 Potential impact on industry and society

The advancement of AI in cybersecurity has the potential to redefine many aspects of the industry and exert a noticeable influence on society. As autonomous systems become more sophisticated, the impact will be felt on a variety of levels:

1. Optimising enterprise security:

- ▶ Businesses can expect greater protection against threats with systems that can learn and adapt in real time.
- ▶ A reduction in incident response times and an improved ability to prevent breaches before they happen is anticipated.

2. Transforming the work of cybersecurity professionals:

- ▶ AI can handle routine tasks, allowing cybersecurity professionals to focus on more strategic or complex tasks.
- ▶ This could result in a restructuring of roles and responsibilities, as well as the need for new skills and training.

3. A digitally safer society:

- ▶ As AI-based solutions become integrated into more platforms and services, the average citizen can benefit from more robust digital security in their daily lives.
- ▶ Online transactions, storage of personal data and other digital activities may experience less associated risks.

4. Changes in threats:

- ▶ Attackers will also evolve and adapt their methods in response to more advanced defence systems.
- ▶ We could see an increase in highly sophisticated and targeted attacks that use AI to find and exploit vulnerabilities.

The advancement of AI in cybersecurity has the potential to redefine many aspects of the industry and exert a noticeable influence on society

6. Future of AI in Cybersecurity

5. Ethical challenges and privacy:

- ▶ As AI becomes a common tool in cybersecurity, concerns will arise about the use and misuse of personal data.
- ▶ Companies and organisations should be transparent about how they use AI and ensure that privacy rights are respected.

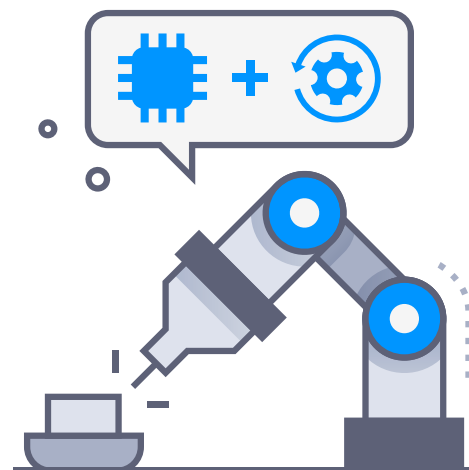
6. Economy and labour market:

- ▶ The mass adoption of AI solutions could influence the demand for cybersecurity professionals, possibly increasing the demand for specialised AI experts and decreasing the need for more traditional roles.
- ▶ Startups and companies developing AI solutions for cybersecurity could experience significant growth, influencing the economy and creating new market opportunities.

7. Rules and regulations:

- ▶ Governments and regulators around the world –as it is currently happening in the European Union– could introduce new regulations to ensure that AI is used responsibly, including in cybersecurity.
- ▶ These regulations could influence how companies develop, implement and use AI solutions.

The impact of AI on cybersecurity is vast and spans many facets of industry and society. It is essential that stakeholders, from practitioners to regulators, are informed and prepared to address these changes in a proactive and ethical manner.



7. Recommendations and best practices

As Artificial Intelligence (AI) becomes increasingly relevant in the world of cybersecurity, both for defence and attack, it becomes imperative that organisations adopt informed strategies for its implementation. However, AI is not a magic bullet that can be applied unceremoniously; its effective use requires a nuanced understanding and a strategic approach.

In this section, we will explore recommendations and best practices that organisations should consider when integrating AI solutions into their cyber security systems. From model selection and training to implementation and real-time monitoring, it is essential that organisations are equipped with best practices to ensure that AI is an asset and not a weakness.

We will address how to ensure the robustness and reliability of AI systems, how to manage and protect the data that feeds these systems, and how to ensure that ethics and transparency are central to any AI implementation. We will also highlight the importance of continuous training and adaptability, given the changing nature of cyber threats.

As Artificial Intelligence (AI) becomes increasingly relevant in the world of cybersecurity, both for defence and attack, it becomes imperative that organisations adopt informed strategies for its implementation

7.1 Integration of cybersecurity teams and AI teams

The effective convergence of artificial intelligence and cyber security requires not only the combination of technologies, but also collaboration between experts in both areas. Effective integration of these teams can boost an organisation's cyber defence capabilities and ensure that AI solutions are robust, reliable and adequate to deal with real cyber threats.

Essential elements for success:

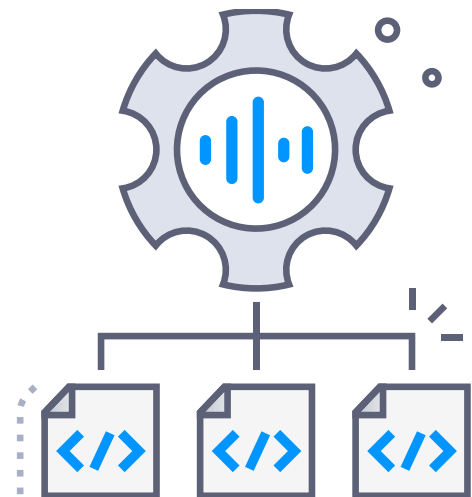
- 1. Communication and collaboration:** A constant and efficient flow of communication between cybersecurity and AI teams is essential. The challenges, objectives and solutions of both fields must be shared and mutually understood.
- 2. Cross-training:** Providing training on cyber security fundamentals to the AI team and, reciprocally, on AI basics to the cyber security team can build mutual understanding and improve collaboration.
- 3. Joint development of solutions:** Rather than working in silos, AI and cybersecurity teams should collaborate on the design, development and deployment of solutions. This ensures that AI solutions are relevant and aligned with cybersecurity objectives.
- 4. Regular reviews and feedback:** AI solutions in cyber security should be subject to regular reviews by both teams. These reviews can identify deficiencies, areas for improvement and necessary adaptations to address emerging threats.
- 5. Joint testing:** As in development, testing of solutions should also be a joint activity. This can help identify and rectify bugs before they are exploited by adversaries.

The effective convergence of artificial intelligence and cyber security requires not only the combination of technologies, but also collaboration between experts in both areas

7. Recommendations and best practices

- 6. Integration of tools and platforms:** Using tools and platforms that enable integration and collaboration between the two teams can be key, including the use of collaborative development platforms, project management systems and communication tools.
- 7. Mutual respect and appreciation:** For effective collaboration, it is essential that both teams recognise and value each other's competencies and contributions. Artificial intelligence and cybersecurity are complex and specialised disciplines, and it is essential that they respect each other to ensure effective collaboration.
- 8. Crisis scenario planning:** In the event of a security incident, it is critical to have plans in place for how teams will work together. This includes determining roles, responsibilities and communication flows.
- 9. Continuous updates and training:** As both the AI and cybersecurity fields are constantly evolving, it is essential that both teams keep up to date on the latest trends, techniques and threats in their respective areas.

Ultimately, effective integration of cybersecurity and AI teams is not simply an option, but a necessity in today's world. Cyber threats are evolving rapidly, and combining cybersecurity expertise with advanced AI capabilities can provide a robust defence against increasingly sophisticated adversaries. However, for this collaboration to be fruitful, it is essential that good practices are adopted to foster communication, understanding and collaboration between these specialised teams.



7.2 Further training

In a world where technology and cyber threats evolve at a dizzying pace, continuous training and education are essential to stay current and ensure effective defence against adversaries. This imperative applies not only to cybersecurity professionals, but also to those working at the intersection of AI and cybersecurity.

In order to achieve all this, it is advisable to take into account:

- 1. Updated training programmes:** Educational institutions and training bodies should regularly review and update their curricula to reflect the latest developments in cybersecurity and AI. This will ensure that new professionals possess the most up-to-date knowledge.
- 2. Workshops and seminars:** Organising or attending workshops and seminars on emerging topics can provide in-depth insight into specific techniques, emerging threats or new approaches in cybersecurity and AI.
- 3. Professional certifications:** Certifications such as CISSP, CISM, and others related to IA and/or cybersecurity, can help professionals validate and enhance their skills. These certifications often require continuing education, ensuring that professionals stay current.
- 4. On-the-job training:** Organisations should foster a culture of continuous learning, providing opportunities for employees to be trained in new tools, techniques and best practices.
- 5. Participation in communities and forums:** Active membership in online communities and specialised forums can provide a platform to learn from colleagues, share knowledge and keep abreast of the latest developments and challenges.

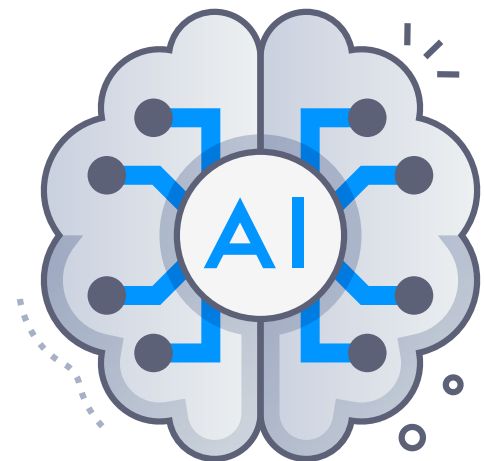
In a world where technology and cyber threats evolve at a dizzying pace, continuous training and education are essential to stay current and ensure effective defence against adversaries

7. Recommendations and best practices

- 6. Simulations and practical exercises:** Conducting simulations and practical exercises, such as "capture the flag" or wargames, can help practitioners apply their knowledge in real-world scenarios, improve their skills and learn from their mistakes.
- 7. Online resources and MOOCs:** With the proliferation of online education platforms, there is an abundance of courses (often free) covering a wide variety of topics in AI and cybersecurity. These can be an excellent way to learn at your own pace.
- 8. Conference attendance:** Conferences such as Black-Hat, DEF-CON, STIC CCN-CERT and other AI-specific conferences offer an opportunity to learn about recent research, discoveries and developments in the field.

In short, training and continuing education in cybersecurity and AI is not a luxury, but a necessity. To mount an effective defence against current and future threats, professionals must constantly be updating and expanding their knowledge and skills. Organisations and professionals that invest in continuing education will be better positioned to address and mitigate risks in the evolving cybersecurity landscape.

Below is a list of training references covering both Artificial Intelligence (AI) and cybersecurity. This is a small number of references including courses, certifications and resources.



7. Recommendations and best practices

1. Courses and Specialisations:

- ▶ **Coursera:**
 - Specialisation in Cybersecurity
 - Introduction to Artificial Intelligence
 - Deep Learning Specialization
- ▶ **edX:**
 - Fundamentals of Cybersecurity
 - Principles of Artificial Intelligence
- ▶ **Udacity:**
 - Nanodegree in Cybersecurity
 - Nanodegree in Artificial Intelligence

2. Certifications:

- ▶ **Certified Information Systems Security Professional (CISSP):**

A globally recognised certification that demonstrates capability and knowledge in cybersecurity. More information on the official ISC² website.
- ▶ **Certified Information Security Manager (CISM):**

Offered by ISACA, this certification is essential for information security management. More details on the official ISACA website.
- ▶ **TensorFlow Developer Certificate:**

A certification focused on AI and deep learning. Find more information on the official TensorFlow website.

3. Additional Resources:

- ▶ **MIT OpenCourseWare:**
 - Introduction to Deep Learning
 - Computer and Network Security
- ▶ **Cybrary:**

A platform offering free courses in cybersecurity and related fields. Visit the Cybrary website.
- ▶ **ArXiv:**

An invaluable resource for researchers, ArXiv is a repository of preprint articles in various fields, including AI and cybersecurity. Check out ArXiv to stay up to date with current research.
- ▶ **ANGELES Platform, CCN-CERT**

<https://angeles.ccn-cert.cni.es/es/>

7.3 Designing robust and resilient systems

Robustness and resilience in the design of systems, especially those embedded with Artificial Intelligence, are essential to ensure that they continue to perform optimally under adverse conditions and can recover quickly from failures or attacks.

By **robustness** we mean the ability of a system to withstand disturbances and continue to function properly without degradation. A robust system can handle unforeseen conditions and variations in the operating environment.

Resilience is the ability of a system to recover quickly from failures, adapting to new conditions and restoring its normal functioning.

In order to achieve robust and resilient systems, the **following elements** need to be considered:

Design Principles	Attack Area Minimisation:	Reduce the attack surface by limiting entry points into the system and eliminating unnecessary components.
	Redundancy:	Implementing duplicate systems and components that can take over workload in case of failure of a primary component.
	Segmentation/segregation:	Dividing the system into smaller, independent segments, so that a failure or attack in one segment does not compromise the entire system.
	Continuous monitoring/surveillance:	Use monitoring and surveillance tools and solutions to quickly detect any irregularities or anomalies.
	Regular Updates:	Keep software and hardware up to date to correct known vulnerabilities.

7. Recommendations and best practices

Considerations for AI	Robust Training Data:	Ensure AI models are trained with varied and up-to-date datasets to handle diverse situations
	Rigorous Validation:	Evaluate and validate models under various scenarios and conditions.
	Defence against Adversarial Attacks:	Implement techniques such as regularisation and data augmentation to protect AI models from attacks that seek to exploit their weaknesses.
	Transparency and Explainability:	Use AI models that can be interpreted and audited to understand their functioning and decision-making.
Tests and Simulations	Continuous Monitoring	Regular testing in controlled environments to identify and correct vulnerabilities.
	Fault Simulations:	Simulate failures in different parts of the system to evaluate response and recovery time.
	Incident Response Exercises:	Practice, in accordance with applicable regulations, responding to potential security incidents to improve efficiency and effectiveness in real situations.
Culture of Continuous Improvement	Foster a mindset of constantly seeking to improve the robustness and resilience of the system, learning from incidents and adapting to new threats and challenges.	

In short, designing robust and resilient systems is essential to ensure that systems, especially those embedded with AI, can handle and recover quickly from adverse conditions. This is an ongoing task that requires a combination of design techniques, rigorous testing and a culture of constant improvement.

8. Conclusion

8.1 Final reflections on the current state and future of AI in cybersecurity

The evolution of cybersecurity and artificial intelligence has proven to be a fascinating but also challenging pairing. Both disciplines, separately, have complex trajectories, and their intersection has provoked both revolutions and dilemmas. Reflecting on their current state and the future landscape, several key considerations can be drawn:

1. Increasing Interdependence:

Cybersecurity can no longer be considered a discipline independent of AI. The vastness and complexity of cyberspace, combined with the overwhelming amount of data generated, makes AI-based solutions essential for effective defence.

2. Changing Challenges:

As AI becomes more advanced, so do the threats. Malicious actors quickly adopt new technologies to improve their tactics. It is a constant game of cat and mouse, where defence and attack evolve in parallel.

8. Conclusion

3. Relevance of the Human Factor:

Despite automation and the advanced capabilities that AI brings, the human factor remains irreplaceable. Ethical decisions, data interpretation and understanding context remain a human responsibility. Human-machine collaboration will be critical to the success of cybersecurity in the future.

4. Ethical and Regulatory Challenges:

The adoption of AI in cybersecurity brings with it ethical and regulatory challenges, such as the case of the European Regulation on AI that we have referred to in this paper on several occasions. Privacy, consent and transparency are areas that need to be approached with caution and responsibility, especially when balanced against the need for security.

5. Untapped Potential:

While we have seen impressive advances in AI applied to cybersecurity, there is still vast untapped potential. Emerging technologies, such as quantum AI and federated learning, may further reshape the cybersecurity landscape in the next decade.

6. Preparing for the Future:

Organisations and cybersecurity professionals must be prepared to adapt quickly. Continuous education, research and interdisciplinary collaboration will be essential to keep up with the latest trends and threats.

7. Holistic Vision:

Cybersecurity, at its core, is a holistic discipline. It is no longer just about technology, but also about processes, people and policies. The adoption of AI should be seen as part of a broader, more strategic approach to securing cyberspace.

In conclusion, the intertwining of AI and cybersecurity is redefining the future of digital security. While it presents unprecedented opportunities for more effective defence and faster detection, it also introduces complex challenges that must be addressed with prudence, innovation and collaboration. The future trajectory of this intersection will undoubtedly be exciting and defining for the digital future of humanity.

Human-machine collaboration will be critical to the success of cybersecurity in the future

8.2 Subsequent actions and recommendations for future research

The evolving landscape of cybersecurity and artificial intelligence requires not only a reflection on what we have learned so far, but also a clear vision of the next steps. As we move towards a more digitised and interconnected future, it is essential that the global community - from researchers and practitioners to policymakers and ordinary citizens - unite in the mission to secure our cyberspace.

Some recommendations and subsequent actions are set out below:

1. Establishment of Collaborative Research Centres:

It is essential to establish more centres and platforms that enable interdisciplinary collaboration in cybersecurity and AI. These centres can act as focal points for innovative research, bringing together experts in AI, cybersecurity, law and other related fields.

2. Promote specialised education and training:

There is an urgent need to develop education and training programmes that focus on the intersection of AI and cybersecurity. This is not only intended to address the skills shortage in this field, but also to ensure that future professionals possess the necessary knowledge.

3. Global Norms and Standards:

The international community should work together to develop standards and regulations around the application of AI in cybersecurity. These regulations will not only provide a frame of reference, but also ensure that the technology is used in an ethical and responsible manner. To this end, it is considered essential to formalise certification frameworks for trusted AI technologies, products and services⁴⁴.

As we move towards a more digitised and interconnected future, it is essential that the global community – from researchers and practitioners to policymakers and ordinary citizens – unite in the mission to secure our cyberspace

⁴⁴ On this subject, see the excellent work of the Rand Corporation. *Labelling Initiatives, codes of conduct and other self-regulatory mechanisms for artificial intelligence applications* (2022). https://www.rand.org/pubs/research_reports/RRA1773-1.html

8. Conclusion

4. Emerging Threats Research:

With the rapid evolution of AI, threats are also changing and adapting. It is crucial to fund and prioritise research on emerging threats, especially those that originate from the latest technological advances.

5. Development of Explainable AI Tools:

The world needs more research into tools and methods that make AI more transparent and understandable. Decisions made by AI algorithms in the field of cybersecurity can have significant repercussions, so it is essential that they can be explained and understood.

6. Promoting Privacy and Ethics:

Future research should not only focus on the effectiveness and efficiency of AI solutions in cybersecurity, but also on their ethical and privacy impact. Privacy should not be a sacrifice to achieve security.

7. Rigorous Testing and Validation:

Before implementing AI-based solutions in real-world environments, it is crucial to conduct thorough testing and validation. This will ensure that solutions are robust and reliable in the face of real-world threats.

8. Incentives for Innovation:

Governments and private organisations should provide incentives for innovation in cybersecurity and AI. This can take the form of grants, competitions or awards.

In short, we are at a pivotal point at the intersection of cybersecurity and AI. As these disciplines continue to evolve and intertwine, it is essential that we take a proactive, collaborative, regulatory and ethical approach to meet the challenges of the future. The call to action is clear: we must unite in the mission to secure our digital future, protecting our data and our infrastructures - ultimately our societies.





centro criptológico nacional



www.ccn.cni.es

www.ccn-cert.cni.es

<https://oc.ccn.cni.es/>

