axonpartnersgroup.com

# AI vs AI

## How artificial intelligence can enhance – and threaten – cybersecurity

**April 2024**


**Authors**

Álvaro García
SENIOR MANAGER

Cem Çelebi
MANAGER

Mehmet Çavdar
ASSOCIATE


AXON CONSULTING

# Contents

# 1. Introduction

In our whitepaper **Artificial Intelligence and the cyberthreat landscape: Understanding and navigating the challenges** we discussed the cybersecurity threats that artificial intelligence (AI) systems face and the measures that are being taken to address them.

But could AI itself be a cybersecurity threat? AI technologies are certainly becoming more sophisticated – but they are also becoming more accessible to more people.

Which is why this whitepaper looks first of all at ways in which AI can actually boost cyberattacks and sidestep existing cybersecurity measures.

This is not the whole story, of course. AI also has a role in countering such attacks. Here we examine this role with reference to some specific threats and possible responses to those threats. We also discuss the state of security awareness.

But it's not just about balancing the good and bad things that AI can do. Yes, AI can help counter the attacks that AI itself may enable – but there are other issues that need addressing, which this paper does.

For example, are we likely to become so reliant on AI for our operations that its role in cybersecurity is simply accepted but not understood or explained? Could overreliance on AI lead to vulnerabilities? How do false positives and false negatives come about – and how do we avoid them? These challenges need to be addressed to ensure that the full potential of AI in cybersecurity can be achieved.

With that in mind, this paper ends with advice on reducing risk through a secure AI lifecycle: the creation of a secure ecosystem of AI technologies, encompassing design, development, deployment, operation and maintenance.

# 2. AI-based attacks and existing cybersecurity responses

AI has changed everything. Not long ago only a few organisations leveraged AI tools to enhance business processes. Now it's hard to find any that don't. And malicious actors aren't far behind.

In Hong Kong in late January an employee of a multinational company fell for an elaborate scam. He was part of a video call with high-ranking officials of his company demanding urgent and costly action. Except they weren't high-ranking officials; they were digital fabrications. This was a cyberattack utilising deepfake technology – and it cost the company USD 26 million.[1]

This example highlights the evolution of cybercrime tactics utilising AI technology. Today cybercriminals (usually motivated by financial gain), hacktivists (who use hacking techniques to promote political or social agendas) and others can leverage AI to enhance their attack capabilities and circumvent existing cybersecurity measures. Firewalls, endpoint security solutions, intrusion detection systems, security information and event management systems can all be targeted.

With this in mind researchers, academics, industry experts and government officials have got together to categorise AI-based attacks into several distinct groups.[2] Let's take a closer look at these groups.

1 "'Everyone looked real': multinational firm's Hong Kong office loses HK$200 million after scammers stage deepfake video meeting" Kong, 2024. https://www.scmp.com/news/hong-kong/law-and-crime/article/3250851/everyone-looked-real-multinational-firms-hong-kong-office-loses-hk200-million-after-scammers-stage

2 Mirsky et. Al https://arxiv.org/pdf/2106.15764.pdf

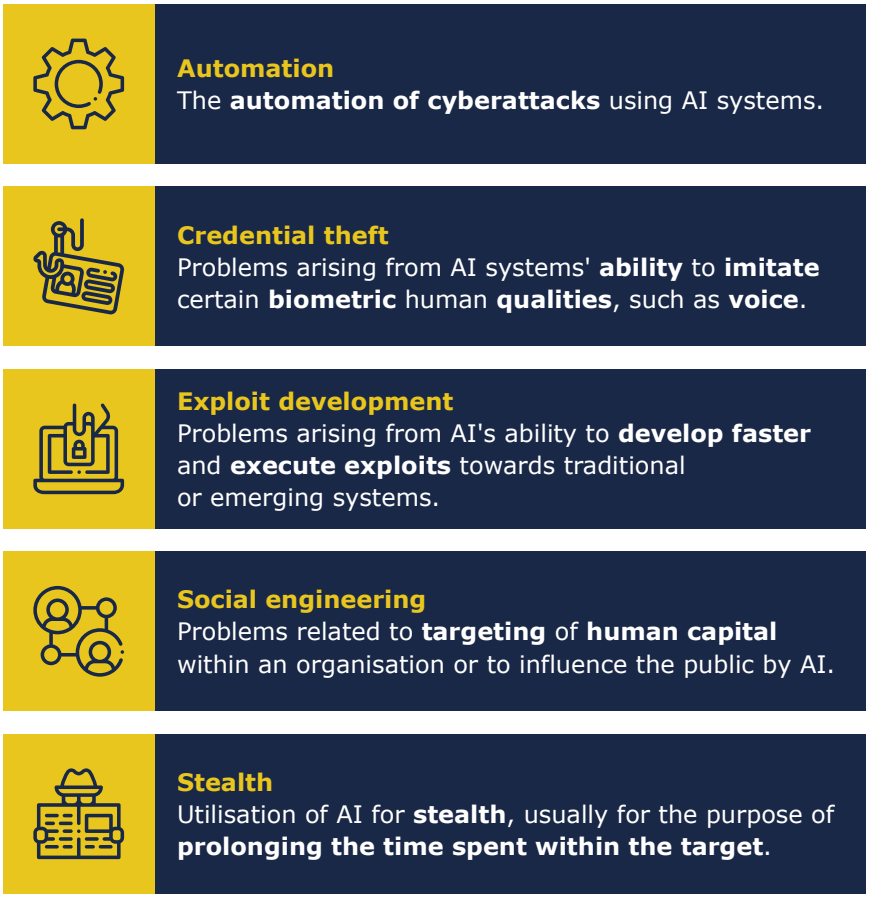


Figure 1: Overview of AI-based threats and cybersecurity measures [Source: Axon]

### Automation

Automation is fairly self-explanatory: cyberattacks with little or no manual intervention. In fact regular rule-based automation is already being used in the early stages of some cyberattacks. Automation allows a malicious actor to carry out reconnaissance procedures by scanning the network configuration of a host, gathering hardware and software information or working with open-source intelligence or publicly available information.

Some AI capabilities are especially helpful here. For example, image analysis and optical character recognition (OCR) libraries (such as OpenCV or Tesseract) can help with bypassing CAPTCHAs.[3] CAPTCHAs are widely used across the internet to prevent automated bots from abusing online services or websites. A well-known application is in user registration or the login pages of web applications.

The bottom line is that malicious actors can use autonomous methods to speed up or enhance attacks that would have been less effective carried out mainly by humans.

**The bottom line is that malicious actors can use autonomous methods to speed up or enhance attacks that would have been less effective carried out mainly by humans**

### Credential theft

AI can imitate certain types of human behaviour. That makes it very useful for credential theft. This is especially true for Generative AI (GenAI). GenAI models are designed to generate new content that is similar to, or based on, existing data. This is where ingenious new strategies come into play such as chat responses, designs, deepfakes, synthetic data – and even synthetic biometric data.

Synthetic biometric data closely resembles real data used by biometric security systems. Fake fingerprints, lifelike facial images, or voice mimicry could bypass authentication mechanisms – and give malicious actors access to a targeted organisation.

3 Short for Completely Automated Public Turing test to tell Computers and Humans Apart.

**Exploit development**

An exploit refers to a program or code segment crafted with the intention of identifying and exploiting security weaknesses or vulnerabilities within an application or computer system. The aim, almost always, is malicious; the installation of malware is the most obvious example.[4]

This isn't new. Malicious actors have long been developing exploits for all sorts of emerging fields – such as commercial software products, the Internet of Things (IoT) or blockchain – to identify vulnerabilities and exploit them.

The consequences? Unauthorised access, data theft, malware installation, privilege escalation… the list goes on. It even incudes reverse-engineering techniques such as de-compilation, which makes it possible to convert executable or ready-to-run program code into some form of higher-level programming language that humans can easily understand. This can then be used to understand how a type of commercial software works.

Neural machine translation (NMT) is an example of this. NMT is a proposition to machine translation that uses an artificial neural network to predict the probability of a sequence of words, typically modelling whole sentences in a single integrated model.[5] It is used to de-compile and reverse-engineer code.[6] Using NMT, malicious actors can identify frequently used and publicly available code to find patterns in binary code. They can then deduce the source code of the targeted software.

And this takes us back to automation. Once any vulnerabilities are identified, AI can be used to automate the process of generating exploits by creating code or attack payloads that specifically target the discovered vulnerabilities. AI algorithms can also optimise the exploit code. That makes it not just more efficient but harder to detect.

4 https://www.cisco.com/c/en/us/products/security/advanced-malware-protection/what-is-exploit.html

5 https://www.theaidream.com/post/how-ai-is-changing-personal-data-tracking

6 https://arxiv.org/pdf/2112.15491.pdf

### Social engineering

The weakest links in an organisation's security are its people. Social engineering exploits this weakness.

Simply put, social engineering refers to all techniques aimed at talking a target into revealing specific information or performing a specific action for illegitimate reasons.[7] And AI can be part of this process. AI can support attack phases involving human victims, through reconnaissance and then addressing initial access, privilege escalation, and credential access.[8]

**AI can support attack phases involving human victims, through reconnaissance and then addressing initial access, privilege escalation, and credential access**

And, once again, there's GenAI, which can replicate human writing styles convincingly – in emails, for example. It's a real boon for impersonation, persona building on social media, spear phishing and business email compromises, to name only a few.

### Stealth

Stealth means making malware inconspicuous – concealing any alterations made by the malware to the compromised system.[9] AI algorithms can help with this process: they can be employed to obfuscate malicious code and conceal its true intent. For example, by dynamically generating polymorphic code (code that mutates while keeping the original algorithm intact) or leveraging machine learning algorithms to modify code patterns, attackers can evade detection by traditional measures. That is, they can evade signature-based antivirus systems and intrusion detection or prevention systems.

The better attackers do this, the longer they can avoid detection, keep an attack going, and maximise its impact. AI's stealth capability boosts tactics like reconnaissance, initial access, persistence, lateral movement, collection, and exfiltration. AI can also help attackers evade intrusion detection systems by mimicking normal network behaviour.

7 https://www.enisa.europa.eu/topics/incident-response/glossary/what-is-social-engineering

8 https://www.traficom.fi/sites/default/files/media/publication/TRAFICOM_The_security_threat_of_AI-enabled_cyberattacks%202022-12-12_en_web.pdf

9 https://www.kaspersky.com/resource-center/definitions/stealth-virus

# 3.
# AI in cybersecurity: the state of play

AI can attack. But it can also defend. It's ironic perhaps, but not surprising, that AI's vast promise makes it both a threat and a boon to cybersecurity. To illustrate the point, this chapter focuses on a few specific threats, responses to those threats, and the state of security awareness, in five specific areas.


Figure 2: Various use cases of AI in cybersecurity [Source: Axon]

### Threat detection and prevention

As the term suggests, threat detection and prevention solutions involve tools and processes that identify and respond to security threats before they can damage systems or data.[10]

How does AI benefit this process? To begin with, AI algorithms can analyse vast amounts of data in real time. That means AI-driven cybersecurity solutions can identify both familiar and unfamiliar threats, such as malware, ransomware, phishing attacks, and unknown (or zero-day) vulnerabilities – and can do so more quickly and accurately than traditional rule-based approaches.

AI algorithms can also learn from relevant data. That means AI can predict and find potential anomalies – a major boost to threat detection and prevention.

**AI algorithms can also learn from relevant data. That means AI can predict and find potential anomalies – a major boost to threat detection and prevention**

### Network security

Networks have become more complex and distributed.[11] How can AI and machine learning (ML) technologies help to keep them secure?

The answer is analysis. Through the analysis of packet headers, payloads, and flow data, AI-driven systems can differentiate various forms of network traffic – find out, in other words, whether they are legitimate, suspicious, malicious, or unauthorised. Organisations can, in turn, implement appropriate security controls and access policies based on these findings.

### Incident response and forensics

Automation of incident response processes means faster and more effective reactions to security incidents. It also minimises the likelihood of human error.[12] This is again where AI really comes into its own. AI algorithms can examine vast amounts of data to find concealed threats, advanced persistent threats (APTs), and zero-day exploits, uncovering the patterns, anomalies, and behavioural markers that indicate cyberattacks or insider threats.

Automated response actions are just one example of this – albeit an important one. This is when machine learning algorithms identify anomalies and spot unusual behaviours. Automated response actions can help isolate infected systems from uninfected ones, block malicious IP addresses, or cut specific connections that may enable infiltration.

10 https://www.cisco.com/c/en/us/products/security/what-is-threat-detection.html

11 https://www.cisco.com/c/en/us/solutions/artificial-intelligence/artificial-intelligence-machine-learning-in-networking.html

12 https://www.paloaltonetworks.com/cyberpedia/what-is-security-automation

### Phishing detection

AI can automate phishing – but AI technology can also automate ways of combatting phishing. AI analysis of suspicious emails, messages, and URLs involves extracting relevant features and training machine learning models on that data to find out whether phishing is indeed taking place.

For example, natural language processing (NLP) techniques allow AI models to detect hidden patterns, grammatical errors and inconsistencies that may indicate phishing mails. Real-time monitoring capabilities offered by AI systems can also allow the detection and automatic flagging of suspicious mails or URLs.

Again this is about AI's ability to process large volumes of data, identify patterns and indicators of phishing attacks, and make accurate predictions – in real time.

### Malware analysis and detection

AI technologies can analyse malware samples, malicious code, and forensic artifacts to identify the tactics used by malicious actors. More specifically, AI algorithms can be used to 'decompose' files (such as executables or any electronically stored information) and analyse decompiled code to identify relevant patterns and behaviours to detect and identify new and evolving forms of malware.

This is particularly important in combating two threats in particular. One is zero-day threats – usually an unknown vulnerability in your computer or mobile device's software or hardware. The other is polymorphic threats. Here malicious actors constantly change the structure and content of malware to create variations that may go undetected by static measures relying on fixed patterns.

# 4.
# AI in cybersecurity: too good to be true?

A recently published survey shows that 64% of respondents (made up of more than 800 cybersecurity decision-makers in North America and the United Kingdom) are highly likely to add an AI-centric technology or solution to improve their cybersecurity readiness within the next year. In fact 61% of respondents believe AI will outperform humans when it comes to identifying threats.[13]

But is that a good thing? Not necessarily. In this section we discuss five possible downsides of overreliance on AI in cybersecurity.

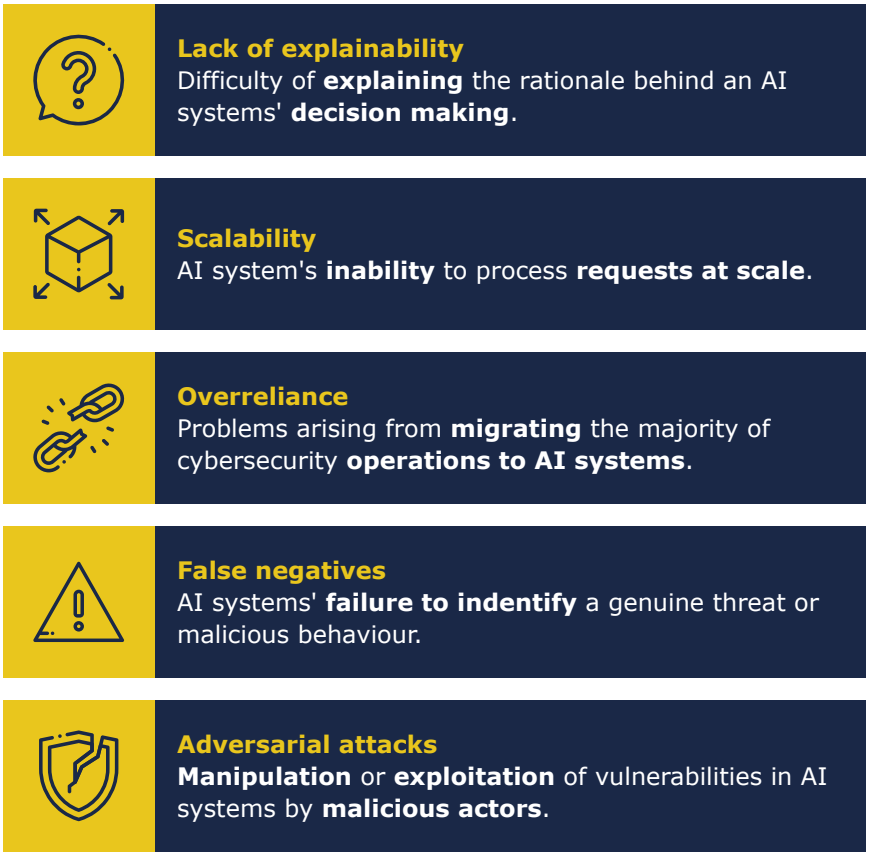


Figure 3: Overview of threats associated to using AI in cybersecurity [Source: Axon]

13 https://arcticwolf.com/the-future-of-artificial-intelligence-in-cybersecurity/

### Lack of explainability

Trust is a key factor in cybersecurity – and an important part of trust is explainability. Can you understand the rationale behind an AI system's decision or prediction? Can you say why the AI system flagged a particular activity as malicious or made a specific decision? Should you follow the decisions made by AI without understanding the implications? In other words, is the whole process explainable?

Here's an example. A production process is halted based on a recommendation made by AI. Can the person implementing that recommendation say why it happened? Both revenue and the supply chain are hit. Why was this necessary?

A decision-maker deferring to AI may be undermining or negating their own expertise – and failing to explain the decision to everyone's satisfaction.

More importantly, who takes responsibility? When decisions are made that aren't clearly understood stakeholders inside and outside an organisation might question the accountability of an organisation or employee – and the reputation of both.

### Scalability

With the rise of cloud computing, and cloud-based cybersecurity solutions from large cloud providers such as Amazon Web Services or Microsoft, scalability of any IT system is generally seen as less of an issue than it once was.

However, scalability of cybersecurity systems still requires careful consideration and planning, especially as applications and systems grow in complexity and are more widely used. Experts need to monitor tools and metrics to help track resource utilisation, response times, and throughput in order to optimise resource utilisation. They also need to implement auto-scaling policies to control costs while ensuring scalability.

**Experts need to monitor tools and metrics to help track resource utilisation, response times, and throughput in order to optimise resource utilisation**

**Overreliance**

Overreliance on technology may lead to a false sense of security. Yes, AI technologies enhance the efficiency and effectiveness of cybersecurity measures, but even here there may be limitations and vulnerabilities.

For example, AI systems are usually trained with historic data. Cyberthreats are constantly evolving. Is the AI data out of date? Could this lead to false positives, false negatives, or missed threats?

This, again, takes us back to explainability. Should we rely blindly on the outputs produced by AI systems without understanding how such outputs are produced? In the event of a sophisticated attack, human technical expertise, intuition and creativity can still play a part in defending an organisation.

**In the event of a sophisticated attack, human technical expertise, intuition and creativity can still play a part in defending an organisation**

How do AI systems make their decisions? How are they relevant to the cyberthreats being addressed? How are they relevant to the overall cybersecurity management strategy? Context and understanding matter. The human factor is still important.

### False positives and false negatives

Which brings us to false positives and false negatives.

False positives happen when an AI system incorrectly identifies benign activities or normal behaviour as security threats. Operational disruption is a possible consequence: an organisation allocating resources to an incident that isn't actually a threat.

By contrast, false negatives – missing a threat – can prolong an unseen danger, ultimately leading to reputational damage or financial losses.

In both cases there's the danger of increased exposure. False negatives mean some security breaches aren't detected, giving attackers a free hand. False positives, on the other hand, divert resources from actual threats – good news, yet again, for would-be attackers.

The end result? Diminished trust and confidence – an organisation that doesn't trust AI-based security solutions, and stakeholders and clients that don't trust an organisation to protect critical assets or sensitive information.

So what's the answer? Firstly, regular updates of the AI model – that is feeding with a diverse dataset. After that, regular supervision and monitoring of the system's performance and its outputs. Common sense? Yes, but both approaches can still be very effective in preventing false negatives and false positives.

### Adversarial attacks

Malicious actors can manipulate or exploit the vulnerabilities of AI systems, crafting inputs specifically designed to bypass detection mechanisms. How and why this affects cybersecurity is discussed in a previous whitepaper, accessible on the Axon website.

# 5. The secure AI lifecycle: reducing risk

Secure AI lifecycle practices aim to reduce the overall risk level of any AI system. But traditional software systems may follow existing security practices within their development lifecycles. Does AI make them irrelevant? Not always. AI systems can still draw on the established secure lifecycle developments of traditional systems – for example, the principles outlined in the ISO/IEC 27000-series of standards relating to IT security, cybersecurity and privacy protection.

However, such standards may not address certain novel challenges of AI systems – such as the unique relationship of an AI system with data and how that should be managed. An approach specifically tailored for the AI lifecycle is needed.

Hence the secure AI lifecycle, which draws from UK guidelines[14] and offers a solid methodological approach to defining the secure AI lifecycle.

14 https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf

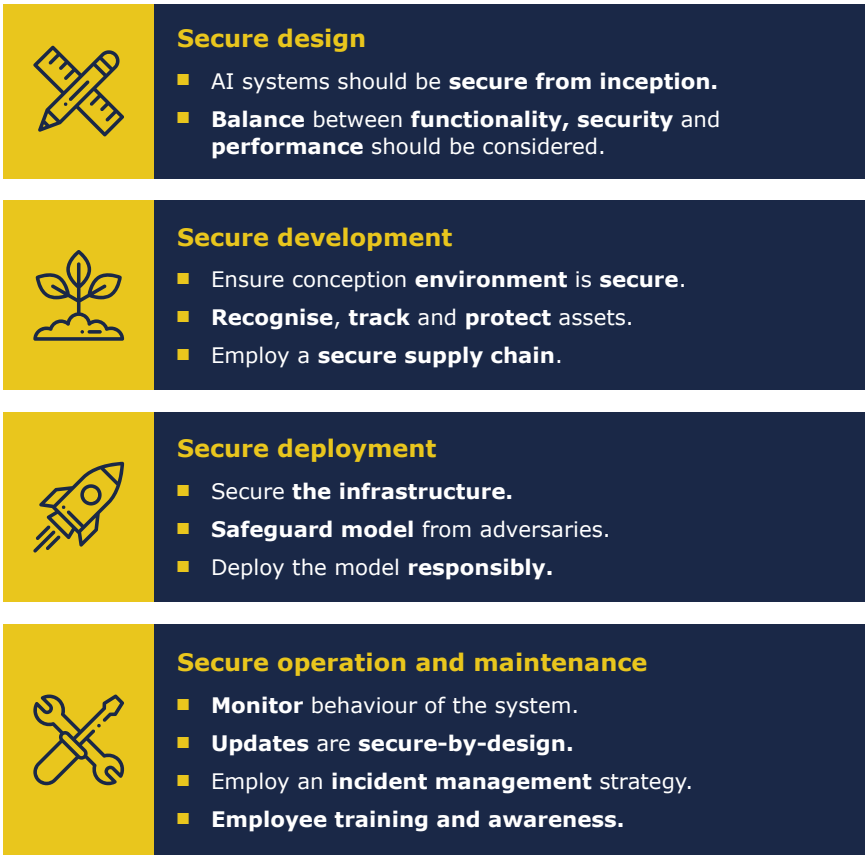


Figure 4: Overview of secure AI lifecycle practices [Source: Axon]

We go into more detail below, but, briefly, the main steps come under four categories. The lifecycle begins with the secure design stage – that is, incorporating security principles into AI systems right from the start. Then there's the secure development stage. Here we ask questions like: How secure is the environment in which the AI system is designed? How well protected are assets? Is the supply chain secure? We then move on to the deployment processes of AI systems. Finally, we address secure operation and maintenance.

These lifecycle considerations may appear to be relevant only to suppliers of AI systems. However, knowing the basic security principles AI systems should adhere to can help with safe use of AI systems and encourage a healthy AI environment.

**Secure design**

Do you aim to incorporate security measures and considerations into the core design and architecture of a system right from its inception? Then you're using secure design – and that is essential to the security of AI systems.

This approach involves threat modelling: identifying assets, threat vectors and types. Here the STRIDE model can be used. STRIDE is a mnemonic to identify various threat types – spoofing, tampering, repudiation, information disclosure, denial of service and escalation of privilege.

Other approaches include generating threat scenarios[15], conducting risk assessments, putting in place robust testing practices, and using secure coding practices and software development frameworks.[16]

There is of course a balance to be struck between security and functionality and performance. This will also be part of the secure design process. Functionality, user experience, and whether an AI system satisfies ethical and/or legal requirements are relevant here.

**There is of course a balance to be struck between security and functionality and performance**

In practice this also means performing due diligence assessments of the model: after all some of its components come from third parties. Potential security threats posed by imported data must also be addressed.

15 https://www.csa.gov.sg/docs/default-source/csa/documents/legislation_supplementary_references/guide-to-cyber-threat-modelling.pdf?sfvrsn=a367caff_0

16 https://www.enisa.europa.eu/publications/artificial-intelligence-and-cybersecurity-research

**Secure development**

Is the environment in which the system developed secure? That is, have you secured the various tools, libraries, modules, middleware, frameworks, network systems, and external APIs from which the AI system is designed?

- **Libraries** – reusable collection of codes that provide various functionalities in helping develop AI systems.
- **Modules** – code segments with specific functions that enhance the modularity of AI systems.
- **Middleware** – software which facilitates the communication between different components of a system.
- **External APIs** – external services which enable software communication for data exchange.

They all need to be secure.

Recognising, tracking, and protecting the assets related to AI systems is also an important consideration in secure development. The assets, their level of importance to the AI system, and any associated risks should all be identified.

As for data imported from the third parties, developers of AI systems should secure their supply chains in this stage. Suppliers of imported components (such as data, software libraries, or modules) should follow suit when it comes to other software assets an organisation implements.

**Secure deployment**

The software is ready to be shipped. But if the system is not deployed properly, the time spent on secure designing and development could be wasted. Incorrect deployment of AI systems could affect performance, expose user data to attackers, or risk theft of the model.

What is the appropriate response? Firstly, secure the infrastructure. This is an essential step in the design and development of an AI system – and one that can be strengthened by segregation of the environments holding sensitive code or data. Logging plays a crucial role here: it provides an audit trail and detailed record of system activities, enabling effective incident response, and generally enhancing cybersecurity.

Ensuring continuous protection of the model is important in both the development and the deployment stages. Continuous protection can deter anyone aiming to steal the functionality of the model or the data it was trained on. Continuous protection also helps guard against indirect threats to the model (via privacy attacks such as white or black box membership inference attacks). This can be carried out with well-established and standard cybersecurity best practices.

**Continuous protection can deter anyone aiming to steal the functionality of the model or the data it was trained on**

Responsible deployment of the model is next. This includes continuously monitoring the model to protect it from would-be attackers. It also includes assessing security concerns. Benchmarking and red teaming are key terms here. Benchmarking compares the performance of any given system to other AI systems on the market. Red teaming simulates attacks on AI systems to identify vulnerabilities. Both help uncover weaknesses and ensure that the model performs to the desired standards.

**Secure operation / maintenance**

The secure operations and maintenance of AI systems includes monitoring the system's behaviour, checking the inputs provided to the system, employing a solid incident management strategy, releasing software updates to strengthen the system, and sharing the lessons learned during these processes. All of this can be reinforced through the promotion of employee education and awareness.

Monitoring a system's behaviour can help spot abnormal behaviour and aid early detection of anomalies – and potential security breaches – in AI systems. Monitoring the inputs to a system can also help detect attacks. It can also help identify areas where updates are necessary. Updates should be deployed with the secure-by-design principle in mind – that is, integrating security measures into systems from their inception, rather than as an afterthought.

Correct deployment of software updates should also involve an incident management strategy. Incident management provides a structured and coordinated approach to identifying, protecting and mitigating security incidents that may impact AI systems. It could also be argued that regular software updates aid ongoing incident management and in turn sustained cybersecurity improvement.

Employees (such as developers, cybersecurity, and IT professionals) are often the first line of defence against cyberthreats.[17] Employee education about, and awareness of, the most-likely-to-occur threats associated with AI is crucial.

A simple example is phishing mail. If the employee is aware that content could be harmful, they can prevent any contamination before it reaches the organisation.

Another is data poisoning. This is the deliberate and malicious contamination of data to compromise the performance of AI and ML systems – such as the wrong classifications being offered in an image-based model. If an employee suspects a data poisoning attack and knows how to help detect and deal with it, by fixing the model or deploying appropriate solutions, time and cost could be saved and risk avoided.

Education and awareness are key to better security. They should not be overlooked.

**If an employee suspects a data poisoning attack and knows how to help detect and deal with it by fixing the model or deploying appropriate solutions, time and cost could be saved and risk avoided**

17 https://www.isaca.org/resources/isaca-journal/issues/2018/volume-4/roles-of-three-lines-of-defense-for-information-security-and-governance

# 6. Conclusion

The threat landscape continues to evolve. AI and cybersecurity overlap – and not always in positive ways. AI systems can be used for malicious purposes. They can present unique challenges of their own. But they can also mitigate risks.

The use of AI within cybersecurity will also throw up unexpected challenges. As well as malicious use of AI, thoughtless overreliance on AI can undermine cybersecurity. That said, the secure AI lifecycle can limit such threats and make secure development of the technology more likely.

But innovation continues to shape how AI interacts with cybersecurity – and it's a double-edged sword. Tools will get more accessible and smarter – but that will appeal to malicious actors who will deploy them in cyberattacks.[18] However, the same tools can help to defend us – if we use them correctly.

And that's the point. It's not really about good AI vs bad AI. It's about us – how we navigate a time of increasingly complex digital environments. We can foster an AI environment that allows us to keep innovating. We can use AI to achieve the promise of a bright digital future. But we can't let our defences down. Ever.

18 https://www.cnet.com/tech/services-and-software/the-biggest-ai-trends-in-cybersecurity/

AXON CONSULTING

# About **Axon Consulting**

Axon is an international firm founded in 2006 that provides investment and advisory services to a broad client base in the ICT and digital space in more than 70 countries across the world.

Axon's cybersecurity practice is central to its business and an increasingly important service area for clients. Its work in this field includes strategy, policy and regulation, governance and research at business and governmental level. Axon works closely with government representatives and other clients to help them understand their cybersecurity needs and challenges in their territories, and to define actionable recommendations aimed at improving their cybersecurity ecosystems.

**Tel:** +34 913 102 894
**Email:** marketing@axonpartnersgroup.com