Artificial intelligence and the cyberthreat landscape: Understanding and navigating the challenges

March 2024

Authors

Samuel Tew
PRINCIPAL

Álvaro García
SENIOR MANAGER

Cem Çelebi
MANAGER

Mehmet Çavdar
ASSOCIATE

AXON
CONSULTING

Contents

1.
Introduction

Artificial intelligence (AI) is revolutionising business and industry as well as the everyday lives of individuals – and its take-up is growing. However, the more we rely on AI systems, the more these systems will be targeted by malicious actors.

How this happens is the subject of this white paper. Its focus is AI and cybersecurity in general and three types of attack in particular: data poisoning, white-box evasion, and membership inference attacks.

These cyberattacks and their effects on AI systems are examined in more detail in the first section of this white paper, but here are some very brief definitions.

The term **data poisoning attacks** usually refers to the injection of malicious data into the training data of a system. The result? Unreliable learning and compromise of the model's performance and behaviour, including its decision-making, or "intelligence". **White-box evasion** happens when misleading or harmful inputs are created that can enter a system without being detected. **Membership inference** is a way of picking out specific points of data in the training set of an AI model and then, potentially, reconstructing the model's training samples. This could result in privacy-related vulnerabilities or worse, especially if the training data consists of sensitive information.

Is anything being done to address these issues? This white paper offers an overview of some key public interventions that aim to enhance the cybersecurity of AI. Among the first countries or regions to address this issue have been the **USA**, the **UK**, the **EU**, **Germany**, and **China**, where specific measures to strengthen AI systems against cyberattacks have been discussed in several official publications, along with the responsible development of AI systems and general cybersecurity measures.

AI and cybersecurity are becoming increasingly linked. But this can be a complex area to navigate. This white paper aims to help readers understand more clearly both the cybersecurity threats that AI systems face and the measures that are being taken to address them.

2.
Examples of threats unique to artificial intelligence

So why should we be concerned about AI and cybersecurity? Partly because AI is not just subject to existing cybersecurity threats. AI systems also face many relatively recent security issues – such as adversarial machine learning (AML).

The term AML is used to describe the exploitation of fundamental vulnerabilities in machine learning components – such as hardware, software, workflows, and supply chains – which are fundamental to AI. Such attacks generally tend to manipulate input data to mislead an AI system's decision-making process.

The term AML is used to describe the exploitation of fundamental vulnerabilities in machine learning components – such as hardware, software, workflows, and supply chains – which are fundamental to AI

While the concept of AML has been around for a long time, the term has only recently come into more general use with the explosive growth of AI. It is now a significant concern. Like AI itself, adversarial tactics, techniques, and procedures have generated a lot of interest and are growing in number and significance[1]. A quick check on the number of papers referring to adversarial attacks confirms this. From 2014 to 2020 attacks increased from next to nothing to thousands. It is now close to eight thousand[2].

A striking example of the very real threat such attacks pose comes from the automotive sector, where some studies have shown that AML-based attacks could deceive driver assistance systems. Put very simply, that means misinterpretation of traffic signs[3] – and, quite possibly, some unpleasant consequences for passengers.

1 Source: The Challenge of Adversarial Machine Learning, Churilla et al. https://insights.sei.cmu.edu/blog/the-challenge-of-adversarial-machine-learning/

2 Source: The Challenge of Adversarial Machine Learning, Churilla et al. https://insights.sei.cmu.edu/blog/the-challenge-of-adversarial-machine-learning/

3 Source: https://www.forbes.com/sites/forbestechcouncil/2023/07/27/adversarial-attacks-on-ai-systems/?sh=1e16c65532be

But that is far from the only major worry. In early 2024, the American National Institute of Standards and Technology (NIST) published a document called *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*, which looked at three major paths for AML in the context of predictive AI applications: poisoning attacks, evasion attacks, and privacy attacks. It's a detailed document on the different variations for each of these three methods – but also highly technical. This white paper is intended to be more widely accessible and focuses on the best-known variations for these attacks: data poisoning, white-box evasion, and membership inference attacks.

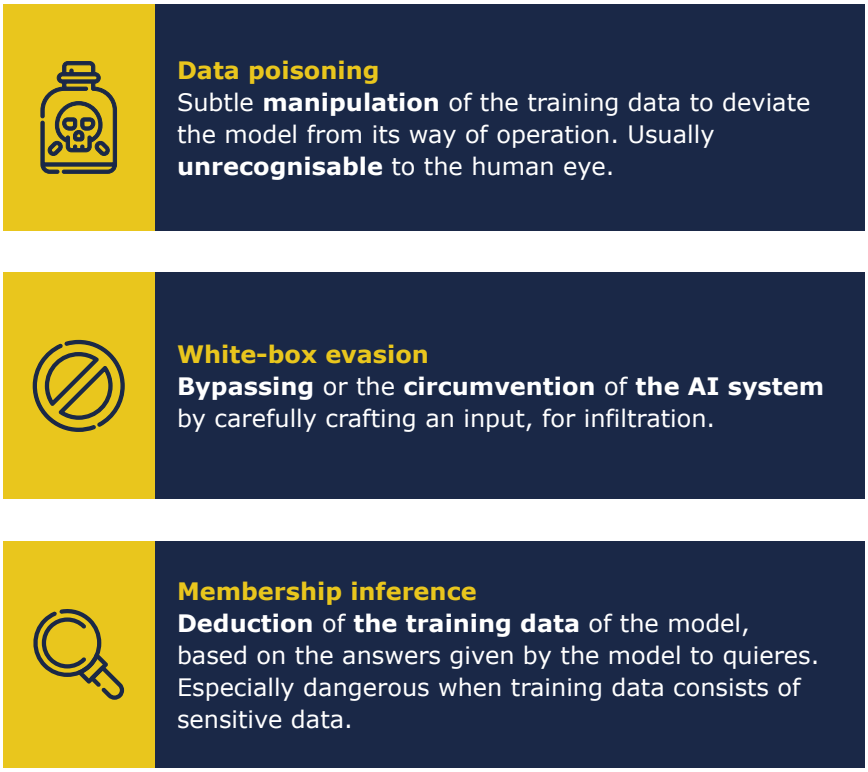


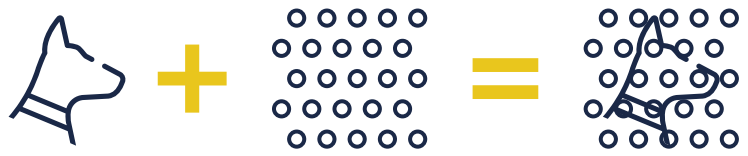
Figure 1: Overview of threats to AI technologies [Source: Axon]

2.1.
Data poisoning

Let's look first at one of the best-known examples of an adversarial attack: data poisoning. Data poisoning means manipulating the training data that is used in building an AI system. The malicious actor doing this manipulation may hope to change the intended outputs or compromise the performance of the model.

As we have indicated, data poisoning attacks take place in the training phase of a model. They are also very stealthy: changes to the training data are usually made only subtly by malicious actors. This means human detection of such attacks can be challenging.

As an example, let's assume an AI model is being trained to classify images of dogs. The model is fed a data set, consisting of correctly labelled images. However, a malicious actor then poisons the training images, so that the model learns from both actual and manipulated images.



Training data **Poison** **Poisoned image***

*The alteration caused in this version of the original image can not normally be be identified by a human.

Figure 2: Data poisoning attack example [Source: Axon]

As can be seen from the (slightly exaggerated) illustration above, the image of the dog is slightly altered; this means the model could confuse the characteristics of the manipulated image with those of a dog.

This could have serious consequences. If the model incorrectly classifies the data used in its training, there could be errors in the model's future predictions. This would compromise the model's integrity and reliability, and produce unreliable outcomes.

What could this mean in real-world situations? Let's consider a life-or-death medical scenario. A model is trained to recognise cancerous cells, aiding diagnosis. If a poisoned model mistakes blurs for cancerous cells, this could affect treatment and recovery – and expenditure on both. But what if the poisoned model mistakes a tumour for a healthy cell? This could lead to non-diagnosis or even loss of life.

If a poisoned model mistakes blurs for cancerous cells, this could affect treatment and recovery – and expenditure on both

And it's not just about classifying images. Predictive AI systems are also used in the cybersecurity industry. If data poisoning attacks result in compromises in the cybersecurity operation chain of an organisation, what could that mean for the AI system for which a machine learning (ML) model, for example, is used?

Here's an example. An intrusion detection system (IDS) uses an AI system to help classify network traffic data as malicious or genuine. A malicious actor targets the IDS's AI module with a data poisoning attack. The poisoned data could cause the AI intrusion detection system to misclassify normal network activities as malicious. It could also authorise malicious behaviours. In both cases, malicious actors could benefit.

2.2. White-box evasion

As we have seen, data poisoning attacks are used during the training phase of AI. By contrast, evasion attacks attempt to cause false classification within a trained model.

How? Well, let's consider an AI system trained to classify emails as legitimate or spam. An attacker might aim to evade the spam detection system by carefully crafting an email to include specific keywords, phrases, or obfuscation techniques[4] that are designed to bypass the system's filters. The manipulated email looks legitimate to the human eye and is also designed to deceive the trained AI model. Thus, the spam email is misclassified as legitimate and ends up in the recipient's inbox.

An attacker might aim to evade the spam detection system by carefully crafting an email to include specific keywords, phrases, or obfuscation techniques that are designed to bypass the system's filters

This is called a white-box attack because the malicious actor has prior knowledge of the AI system in question – its algorithms, or parameters, for example. In a black box attack, the malicious actor has no prior knowledge of the targeted AI system.

But what if an organisation employs an AI-based phishing detection mechanism? There is still a way around this for a malicious actor, who can employ evasion techniques, such as polymorphy[5] to slip past the AI system's detection capabilities and successfully deliver a phishing email to an employee. You can guess the worst possible result: a harmful URL in the email, the end user clicks on it, ransomware is downloaded, and all files connected to the host computer are encrypted.

And that's not all. What could a white-box evasion attack mean for facial recognition systems? This technology matches a human face from a live camera feed. It's often used to safeguard the security of facilities housing sensitive information, such as airports or office buildings, as well as facilitate phone unlocking through a front-facing camera. However, malicious actors can potentially trick the system by manipulating their faces on the camera feed. Some researchers have been able to compromise a facial recognition system by producing a rectangular paper sticker that includes a meticulously crafted image generated through a novel algorithm[6].

4 Definition: Using techniques to hide or encode malicious content, making it challenging for AI systems to recognize the underlying phishing elements.

5 Definition: Constantly changing the structure and content of phishing emails to create variations that may go undetected by static AI models relying on fixed patterns.

6 Source: AdvHat: Real-World Adversarial Attack on ArcFace Face ID System, Komkov and Petiushko, 2021. https://www.computer.org/csdl/proceedings-article/icpr/2021/09412236/1tmhLeqyrHq

2.3. Membership inference

Our third attack variation is membership inference attacks. This means reverse-engineering an AI system by supplying inputs aiming to infer if a particular data sample has been used – or not used – during training. The result? Attackers get information that helps them to reconstruct the model's training samples.

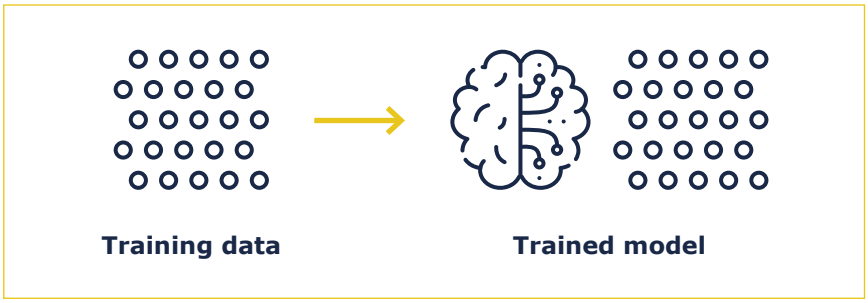


Figure 3: Training of an AI model

Here's a simple guide to how it's done. The figure above shows the training of a model. The data represented by the blue and yellow circles are used in the training process of the deep neural network.

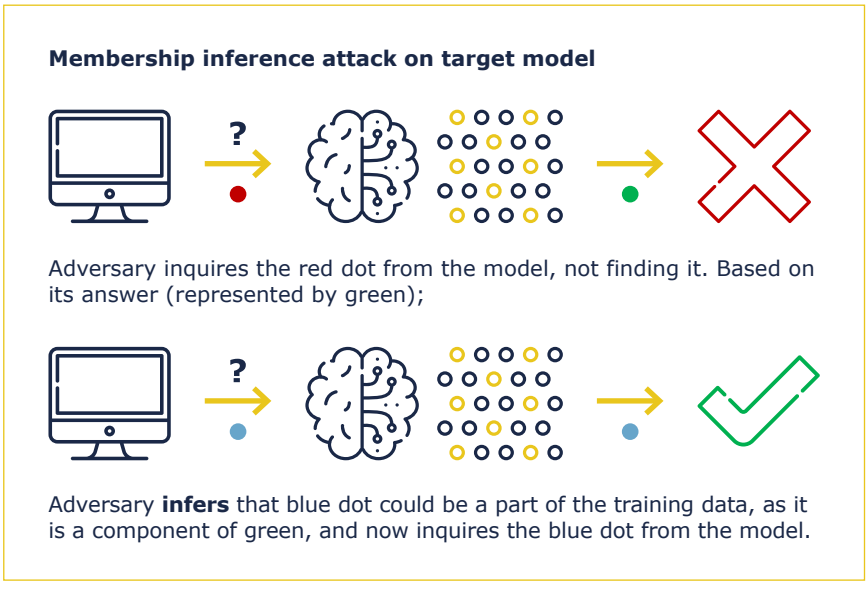


Figure 4: Membership inference attack on target model

In a membership inference attack, the attacker feeds a query to the system, including a specific point of data (represented by the red dot in the figure above). Based on the model's answers to the query, the attacker tries to **infer** whether that specific point of data was part of the initial training data. Based on the answer of the model in the first trial (represented by the green dot), the attacker correctly assumes that the blue dot could be a part of the training data. **Repeating the trial can therefore result in an attacker obtaining a comprehensive overview of the initial training data of the model**.

The figure above shows how this cycle works. In this example, the attacker continues its queries until it reaches the yellow dot, deducing that it could also be part of the training data. It may then continue its queries – trying, for example, to find out the location of the dots, along with the number of blue and/or yellow dots until it feels it has all the training data information it needs.

Another example involves an organisation that has been using an AI system designed to classify files as either benign or malicious based on their content. This model is trained on a dataset containing a variety of files, including both safe and malicious samples. An attacker would seek to determine whether a specific file was part of the AI model's training dataset. Again, the goal is to exploit potential vulnerabilities in the model that may inadvertently reveal information about its training data.

All of which may seem fairly abstract, until we consider the real-world consequences of this approach. One of the most worrying implications is that successfully inferring membership can have serious privacy implications. In certain circumstances uncovering information about the characteristics of a training dataset could mean that personal information – such as a name, address, ID or health information – can be determined by malicious actors. Where the organisation attacked is subject to privacy regulations (such as GDPR), there could also be severe legal and reputational consequences.

One of the most worrying implications is that successfully inferring membership can have serious privacy implications

3.
Key cases of public intervention in AI

We know that AI has many benefits. We've also shown that it can be targeted by malicious actors in some innovative ways. Not surprisingly, these threats and benefits have led many countries to consider regulating some aspects of AI. However, much of the attention so far has been focused on the ethics of AI, there are relatively few references to cybersecurity. That said, there are interventions – planned or happening – where the cybersecurity aspect of artificial intelligence is also recognised, the most relevant of which are discussed in this section.

EU	■ **EU AI Act, first draft 2021.** Not yet issued, underlines the importance of cybersecurity in critical systems. ■ **"ENISA AI Cybersecurity ("CS") Challenges, December 2020.** CS breaches of AI systems, classifies threats. ■ **ENISA AI and CS Research, June 2023.** Aims to identify the areas where research is needed on AI for cybersecurity and on securing AI.
Germany	■ **German Standardisation Roadmap "on Artificial Intelligence, December 2022.** Aims to standardise various aspects of AI, efforts mainly under basic topics of security of AI systems, testing and certification, and socio-technical aspects. ■ **AI Security Concerns in a Nutshell, April 2023.** Identifies common AI security threats.
USA	■ **AI Bill of Rights, December 2022.** ■ **NIST's AI Risk Managment Framework, January 2023.** Aims to help organisations and entities manage many risks of AI systems. ■ **Executive Order on Safe, Secure and Trustworthy AI, October 2023.** A comprehensive order, elaborates on key actions to be taken for security of AI, among other aspects. ■ **NIST's Taxonomy, January 2024.**
UK	■ **AI Safety Summit, November 2023.** Fruited the guidelines. ■ **Guidelines for Secure AI System Development, November 2023.** Aims to address unique AI cybersecurity risks, is a joint effort by multiple countries. ■ **UK Artificial Intelligence Regulation Impact Assessment, March 2023.** Elaborates on three options, with each one getting more comprehensive.
China	■ **Interim Administrative Measures for Generative Artificial Intelligence Service, July 2023.** ■ **Internet Information Service Algorithmic Recommendation Management Provisions, March 2022.** ■ **Draft Internet Information Service Deep Synthesis Management Provisions, December 2022.**

Figure 5: Overview of key public interventions towards crossroads of AI and cybersecurity

3.1. European Union

Noteworthy EU intervention includes an AI act (not yet passed) and two reports by The European Union Agency for Cybersecurity (**ENISA**): one from December 2020, the other from June 2023.

The **EU** proposed its **AI Act** in April of 2021. It was described as "hopefully under the last round of discussion"[7] as of 8 December, 2023.

The risk-centred European AI Act[8] underlines the importance of cybersecurity within its text, noting that high-risk systems should be designed to be resilient against cyberattacks to prevent the leveraging of sensitive data.

Although no detailed cybersecurity provisions are supplied within the act itself, ENISA has published reports that underline AI cybersecurity vulnerabilities and how they can be addressed.

ENISA has published reports that underline AI cybersecurity vulnerabilities and how they can be addressed

Of the two reports published by ENISA, the first (*AI Cybersecurity Challenges*) identifies the assets within the AI ecosystem that ought to be protected and the scenarios where security breaches might occur. Additionally, the report classifies threats posed to AI in various stages of their lifecycles alongside their potential effects. As well as confirming the EU's wish to secure AI systems, the report is, as its title suggests, a useful guide to identifying the cybersecurity challenges that AI faces.

ENISA's other relevant report (*AI and Cybersecurity Research*) is a sort of follow-up to the first: it focuses on the potential areas of study where AI and cybersecurity overlap. The report mostly looks at how existing AI capabilities could extend the capabilities of existing cyberthreats. It also looks at use cases involving AI systems in specific cybersecurity contexts, including secure-by-design practices, as well as ways to deflect common cybersecurity attacks on AI systems.

7 Source: "What Is the EU AI Act and When Will Regulation Come into Effect?", Reuters, 2023 https://www.reuters.com/technology/what-are-eus-landmark-ai-rules-2023-12-06/

8 Source: The AI Act was proposed in April 2021.

Key Insights from the EU

We may soon have a risk-centred AI act from the EU – and we already have an EU expression of interest in cybersecurity measures for high-risk systems. Could these factors imply that more rigid EU initiatives for regulating cybersecurity aspects of such systems are on the way? Indeed, while ENISA reports provide an overview of what those measures could be, it seems to indicate that mandatory controls in the field may not be far off.

With this in mind, some research areas already identified (such as decision trees, K-means clustering, and artificial neural networks) in the ENISA reports could be potential areas for future regulatory intervention by the EU. Additionally, secure design practices identified in the *AI and Cybersecurity Research* report could indicate measures the EU might require manufacturers and/or deployers of AI systems to implement in the future – such as necessary intrusion/malware detection, vulnerability assessment, or feature extraction measures.

3.2. Germany

Germany has also addressed the overlap of cybersecurity and AI, though without recommending concrete regulatory action.

In fact, the German **Standardisation Roadmap on AI** states the need for standardising cybersecurity in the field of AI. It also suggests that generating standardised cybersecurity testing tools particularly tailored to the needs of AI systems is a priority. Its headings include **basic topics**, **security of AI systems**, **testing and certification**, and **socio-technical aspects**, while contexts for standardisation efforts include **industrial automation**, **mobility**, **health**, **medicine**, and **financial services**.

There's another German document that identifies AI security concerns. Called **AI Cybersecurity Concerns in a Nutshell**, it provides an overview of **common threats against AI systems**. Although short, it offers information regarding **evasion**, **information extraction**, and **poisoning and backdoor attacks**, and discusses **general measures for securing AI systems** and the **limitations** of addressing such attacks.

Key Insights from Germany

As of early January 2024, Germany had not passed any specific law regarding the cybersecurity aspects of AI. However, the specifications within the roadmap and the identification of security concerns within AI systems indicate that Germany is aware of the security of AI systems. As in the EU, concrete regulatory action may follow.

3.3. United States

The late 2022 AI Bill of Rights (AIBoR) and NIST's *AI Risk Management Framework* (AIRMF) in early 2023 suggest that the US is working towards comprehensive regulation of various aspects of AI, including the cybersecurity aspect.

The AIBoR could be read as a foundational document for the country's regulatory plans in this area. It does not, however, include a definitive call for action; it could perhaps be seen as a problem-identifying statement of sorts.

Indeed, the **AIBoR** laid the groundwork for the NIST standards that followed and President Biden's executive order of 30 October 2023. AIBoR aims to help guide the **design**, **use**, and **deployment** of what is described as "Automated Systems" **under five principles**, all of which underline the **connection between cybersecurity and AI**.

Not long after the AIBoR, NIST released its *AI Risk Management Framework* (AIRMF). The framework, published in late 2022, aims to help organisations and individuals (referred to as AI actors) **manage** what the framework calls **"many risks of AI systems"**. It does deal with individual risks associated with the use of AI systems; however, it does not explicitly deal with cyber risks in AI systems. The *Risk Management Framework* could be seen as the precursor to, and even part of, a more comprehensive intervention, as it is followed by the **Executive Order on Safe, Secure and Trustworthy AI.** This is the most important regulatory development from the USA to date.

The executive order outlines action to be taken by government agencies such as[9] NIST in areas like standards for team testing. It tasks the Department of Homeland Security to apply such standards to critical infrastructures and establishes the **AI Safety and Security Board**. The executive order also discusses the establishment of "an advanced cybersecurity program that develops AI tools to find and fix vulnerabilities in critical software."

9 Note: List is not exhaustive.

The most recent development to be discussed here (it was published on 4 January 2024), is the NIST taxonomy. Its full title is *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*. This document classifies attacks under two categories, depending on the type of system targeted: **predictive** and **generative**. These attacks are further classified according to a "conceptual hierarchy", based on the attack, the goals of attackers, and their objectives.

Under **predictive AI**, there are three types of attacks identified: **evasion**, **poisoning**, and **privacy**. This approach is similar to our own in this white paper, though with much more detail on the types of attacks and their mitigation.

Under predictive AI, there are three types of attacks identified: evasion, poisoning, and privacy

Generative AI is not addressed in this white paper. However, the NIST document includes a detailed taxonomy and attack classification under that section as well. Following the **attack classification**, the document addresses mitigation in the context of attacks on the **AI supply chain**, as well as **direct prompt injection**, and **indirect prompt injection**.

Key Insights from the USA

It seems then that the USA is underlining a willingness to regulate the cybersecurity aspect of AI by issuing gradually more detailed documents, starting with the AIBoR, then NIST's AIRMF and the executive order, and finally NIST's taxonomy document.

The executive order stood as an actionable agenda item: it solely laid out areas where cybersecurity relating to AI would be the primary focus within the regulatory efforts. It didn't enforce any action, but it can be regarded as an initial step towards identifying cybersecurity threats specific to AI systems, and ways of protecting such systems against attacks. Its length and depth appear to prove that the US aims to drive forward work on creating an environment where AI systems can be secure.

3.4. **United Kingdom**

The UK published its Regulation Impact Assessment (RIA) in March 2023. This document focuses on four options: three that suggest specific actions and one "do nothing" option. In addition, the UK jointly issued (with many other countries) a set of rules under the heading *Guidelines for Secure AI System Development* at the end of November 2023.

The RIA suggests that regulatory efforts in AI are inherently related to cybersecurity. That said, reference to cybersecurity in each of the options is minimal; the options mainly focus on general AI regulations.

The options are scaled: their enforceability ranges from voluntary to mandatory, and their applicability ranges from suppliers only to everyone in the AI value chain. Only one option (a centralised AI regulatory function) includes fines or penalties for non-compliance. The costs of the options rise in each case: the implementation cost associated with option three is almost 20 times the sum of the implementation costs of all preceding options.

The RIA favours the second option – the development of a central regulatory function – which could mean imposing new, mandatory obligations on businesses. However, it should also be mentioned that a pro-innovation approach has been adopted by the UK[10]. This involves managing AI through existing regulators and regulations. Cross-sectoral principles including safety, security, robustness and transparency are outlined in this approach for the consideration of existing regulators.

The RIA favours the second option – the development of a central regulatory function – which could mean imposing new, mandatory obligations on businesses

The UK is also involved in a jointly issued document, *Guidelines for secure AI system development*, which appeared on 27 November 2023. The document was led by the UK and the US but included input from Australia, Canada, Chile, Israel, Japan, New Zealand and others, including much of the EU.[11] It stated that AI systems house unique vulnerabilities[12], differing from other systems, in addition to conventional cyber threats. The guidelines aim to address such risks under four security areas: the processes of **design**, **development**, **deployment**, and **operation and maintenance**.

Key Insights from the UK

As well as the RIA, the UK has hosted the first conference of its kind[13] on AI security. Although the RIA has yet to translate into regulatory intervention, the jointly issued guidelines indicate that the cybersecurity aspect of AI matters to UK regulators. Secure design, deployment, and operation and maintenance practices are laid out in the document, indicating a will to securely deploy AI systems, end-to-end, within the UK.

10 Source: "Policy implications of artificial intelligence", Bhatnagar & Gajjar, 2024 https://researchbriefings.files.parliament.uk/documents/POST-PN-0708/POST-PN-0708.pdf

11 Source: "US and UK Lead 18 Nations in Adoption of AI Security", CPO Magazine, 2023 https://www.cpomagazine.com/cyber-security/us-uk-lead-18-nations-in-adoption-of-ai-security-guidelines/#:~:text=The%20%E2%80%9CGuidelines%20for%20Secure%20AI,and%20much%20of%20the%20EU.

12 Note: Including cybersecurity risks

13 Note: The conference in question resulted in the *Guidelines for secure AI system development*.

3.5. China

China was one of the first countries to implement several regulations aimed at generative AI. They include the **Interim Administrative Measures for Generative Artificial Intelligence Service (IAMGAIS)**, **Internet Information Service Algorithmic Recommendation Management Provisions**, and the draft **Internet Information Service Deep Synthesis Management Provisions.** Here we focus on the first of these: the **IAMGAIS**.

The IAMGAIS covers areas like data governance, quality of training data, bias mitigation, and transparency – mainly as they relate to public service providers. IAMGAIS introduces strict obligations on providers of generative AI systems. The measures even include monitoring and controlling the content generated by generative AI systems, including the removal of illegal content and labelling content created by such systems.

IAMGAIS also says that generative AI systems must be aligned with cybersecurity law. Security assessments are also required before the deployment of generative AI systems.

Key Insights from China

So far China is the only country we have discussed that has introduced AI regulation laws[14]. That said, while cybersecurity measures have become law in China, the emphasis of AI regulations seems to be on ethical and/or political concerns, as can be seen by the regulation of the content generated by AI systems.

14 Note: The EU's AI act has not yet been enforced.

3.6.
Key takeaways on public interventions

The EU, Germany, the US, and the UK have been driving recognition of the need to consider AI and cybersecurity together. This is sure to have an impact on the safe usage and deployment of the technology.

There is, however, one interesting takeaway from this activity. While all these countries or organisations have offered detailed guidelines, so far none of them have brought in mandatory measures to ensure the cybersecurity of AI systems[15]. Why is this?

One possible answer is that these countries wish to encourage AI innovation, and not to slow down the development of AI systems within their jurisdictions. The AI market is expected to grow at an extraordinary rate: the projection for Europe in 2024 is USD 83.67 bn. For 2030 it is USD 202.46 bn. That's a compound annual growth rate of 15.87%[16]. Many countries do not want to issue mandatory controls that could stall AI R&D.

It seems that balancing the need for innovation and the need for security is driving both the pace and focus of regulatory action in these areas.

The AI market is expected to grow at an extraordinary rate: the projection for Europe in 2024 is USD 83.67 bn. For 2030 it is USD 202.46 bn

15 Note: As of early January 2024.

16 Source: Artificial Intelligence – Europe Statista Market Forecast https://www.statista.com/outlook/tmo/artificial-intelligence/europe#:~:text=The%20market%20size%20in%20the,US%24202.50bn%20by%202030.

4.
Conclusion

We have highlighted some of the complex and novel tactics used to exploit vulnerabilities within AI systems. Addressing these attacks requires equally novel approaches and protection mechanisms. Encouraging a culture of cybersecurity awareness is also a must; it aids stakeholder identification and mitigation of such threats.

Our look at public interventions highlights the importance of regulatory frameworks in increasing the resilience of AI systems. Most of the countries we mentioned have yet to employ legally enforceable controls involving the cybersecurity of AI systems. However, the attacks included in this white paper – and others – have been noted as we can see in publications from the UK, the USA, Germany, and the EU, among others.

So how do we move forward? We suggest that a combination of technical approaches with policy interventions can help foster trust in AI applications and confident deployment of AI systems.

The bottom line is that AI systems already face cyberattacks, often carried out in new and innovative ways – and measures are needed to mitigate the risks of such attacks. The novel risks this white paper discusses sometimes require solutions that are both technical and process-oriented. Such efforts should be supported with cooperation across sectors and regulatory frameworks.

In short, responding to the cybersecurity threat to AI is about fostering a more secure environment for AI deployment. Achieve this and all the stakeholders in the AI value chain will benefit.

The bottom line is that AI systems already face cyberattacks, often carried out in new and innovative ways – and measures are needed to mitigate the risks of such attacks

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

Axon Consulting
Axon Partners Group