11

Companies developing compliance with requirements

Guide 11. Cybersecurity

European
Artificial Intelligence Act

Financiado por la Unión Europea
NextGenerationEU

GOBIERNO DE ESPAÑA
MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

Plan de Recuperación, Transformación y Resiliencia

España | digital 2026

This guide has been developed within the framework of the development of the Spanish pilot for the regulatory AI Sandbox, through collaboration among participants, technical assistance providers, potential competent national authorities, and the sandbox's expert advisory group.

The aim of the guide is to serve as an introductory support to the European Regulation on Artificial Intelligence and its applicable obligations. Although **it is not legally binding and does not replace or develop the applicable legislation, it provides practical recommendations** aligned with regulatory requirements, pending the approval of the harmonised implementing standards for all Member States.

This document **is subject to an ongoing process of evaluation and review, with periodic updates** in line with the development of standards and the various guidelines published by the European Commission, and it will be updated once the Digital Omnibus amending the Artificial Intelligence Act is approved.

Among the currently applicable relevant technical references are **ISO/IEC 27001:2022 "Information security, cybersecurity and privacy protection – Information security management systems – Requirements" and ISO/IEC 27017:2015 "Information technology – Security techniques – Code of practice for information security controls based on ISO/IEC 27002 for cloud services,"** which will serve as a basis to ensure information security and the protection of cloud services in the development and deployment of artificial intelligence systems within the context of compliance with the European Regulation on Artificial Intelligence.

Revision date: 10 December 2025

General content

1. Preamble 6
2. Introduction 9
3. European Regulation on Artificial Intelligence 13
4. How to approach the requirements? 16
5. Technical documentation 46
6. Self-assessment questionnaire 48
7. Annexes 49
8. References, Standards & Norms 63

Detailed Index

1. Preamble 6
1.1 Purpose of the document 6
1.2 How to read this guide? 6
1.3 Who is it for? 7
1.4 Use cases used in the guide 7
2. Introduction 9
2.1 What is cybersecurity for AI? 9
2.2 What are the main threats? 10
3. European Regulation on Artificial Intelligence 13
3.1 Preliminary analysis and relationship of the articles 13
3.2 AI Act Requirements 14
3.3 Correspondence of the article with the sections of the guide 15
4. How to approach the requirements? 16
4.1 Level of cybersecurity and consistency 16
4.1.1 Applicable measures 17
4.2 Systems resistance to alterations of use 20
4.2.1 Applicable measures 21
4.3 Data security vulnerabilities and controls 23
4.3.1 Applicable measures 24
4.4 Protection against adversarial attacks 27
4.4.1 Applicable measures 28
4.5 Attacks on AI system flaws 36
4.5.1 Applicable measures 36
5. Technical documentation 46
6. Self-assessment questionnaire 48
7. Annexes 49
7.1 Annex I: Access policies. An approach to high-risk artificial intelligence systems 49
7.1.1 For the provider 49
7.1.2 For the deployer 51
7.2 Annex II: Training in cybersecurity and high-risk artificial intelligence systems 52
7.2.1 For the provider 52

7.2.2 For the deployer 54
7.3 Glossary 55
8. References, Standards & Norms 63
8.1 Standards 63
8.2 ENISA 63
8.2.1 Artificial Intelligence cybersecurity challenges 63
8.2.2 Securing Machine Learning 63
8.2.3 Cybersecurity of AI and standardization 64
8.3 Other references 64

1. Preamble

1.1 Purpose of the document

This guide is intended to develop compliance with the necessary cybersecurity measures for AI systems, **specifically in aspects related** to **artificial intelligence**, so that it is integrated into a broader cybersecurity scheme. The main objective of this guide is to provide companies with the knowledge and steps necessary to implement cybersecurity requirements in AI systems, as established in Article 15 of the European Regulation on Artificial Intelligence. Its application is the responsibility of both the provider and the system deployer.

1.2 How to read this guide?

In the introductory section, what we mean by cybersecurity for high-risk AI systems, in relation to the European Regulation on Artificial Intelligence, is addressed. This section lays the conceptual foundations of the guide.

In the second section, we establish how we understand the requirements established in Article 15: accuracy, robustness and cybersecurity, of the AI Act. It is important to clarify that this guide focuses exclusively on cybersecurity aspects, and not on accuracy or robustness requirements, which are specifically addressed in other guides. Throughout this guide, the reader (provider or, depending on its scope, the deployer) should approach the section with the following approach:

- Identify the assets and actors of your AI system, in relation to the life cycle.
- Associate assets and actors in order to establish their relationships.
- Identify vulnerabilities to which AI system assets are exposed.
- Define and implement security controls to protect the system.
- Periodically review the effectiveness of these controls.

For each of these vulnerabilities and in relation to the objective of achieving the requirements.

Section 3 reviews the relevant articles of the European Regulation on Artificial Intelligence, as well as the requirements it requires.

Section 4 discusses how to address the requirements listed in paragraph 3. Section 5 sets out the basis for what must be documented in relation to cybersecurity, in line with the process described in section 4.

In the following section, section 6, a series of generic self-assessment questions are established for compliance with the requirements demanded by the AI Act.

Section 7 sets out the recommendations on training and access policies, which are mentioned throughout the other sections of the guide. In addition, this section includes

subsection 7.3 which is the list of all the terms and technical aspects indicated throughout the guide.

The guide closes with section 8, which brings together the sources consulted for the preparation of the guide and recommended to delve into those aspects that are more specific and specific to the AI system and its intended purpose. Specifically, in section 8.1, a list of standards is provided, which can help the organization in considering a broader cybersecurity scheme, and which is outside the scope and objective of this guide, focused on cybersecurity for AI, within the framework of the European Regulation on Artificial Intelligence.

1.3 Who is it for?

It is the responsibility of the provider of the high-risk AI system to take appropriate measures (both organizational and technical) to ensure that the protection requirements in the cybersecurity aspect for AI are met. Likewise, within its scope of application, the deployer of the system also has responsibilities that will materialize in specific measures (again organizational and technical).

This guide is aimed at high-risk AI systems in advanced stages of development (from technological maturity level TRL 6) and systems that are already in operation. Therefore, it applies both to systems that are close to putting into service and to those that require adjustment or reinforcement measures during operation.

Throughout the guide, the steps and recommendations necessary to implement the specific cybersecurity requirements set out in the EU AI Act are detailed, both by the provider and the deployer of the AI system. It is important to note that this guide does not address general cybersecurity measures, but rather those specific to high-risk AI systems.

For complementary topics, such as risks management, transparency, technical documentation or quality management, it is recommended to consult the relevant guides.

1.4 Use cases used in the guide

Throughout the guide, two **use cases** will be used as **examples** of how **to develop** the technical documentation. The examples will focus on the AI system provider, who is responsible for generating and curating the documentation. However, it is important to note that the obligation to comply with cybersecurity requirements does not only fall on the companies that develop AI, but also on those that commercialize, implement or deploy it.

A detailed description of the use cases used can be found in **the Guide of practices and examples to understand the AI Act.**

Whenever an example is given, it will be done **in an illustrative way**. Both **providers and deployers** must apply the measures and controls indicated in this guide.

The use cases have been selected based on two reasons:

- The ability to explain the information and procedures detailed in the guide, to establish the necessary criteria, in the understanding of the application of the requirements established in Article 15 of: Accuracy, robustness and cybersecurity.

- The relevance of both use cases in terms of the impact of suffering an attack that may generate serious incidents on the rights of natural persons, including people at risk of social exclusion or minorities.

The selected cases have been, with the following considerations:

- **Aid granting automatic system.**
- **Attendance at work.**

2. Introduction

2.1 What is cybersecurity for AI?

Cybersecurity in high-risk Artificial Intelligence systems is not an option, and all systems are exposed to specific threats that require rigorous protection measures adapted to their context. These threats include manipulation of training data, which can compromise the integrity of models; attacks designed to force errors into results, such as adversarial examples; or those aimed at obtaining private information from data used during system training. In addition, defects in the model or errors in its integration can be exploited by third parties to alter its operation or performance. As will be explained later, you should always be vigilant against the appearance of new threats that may appear from the categories or other new ones, and in no case should this guide be considered to contain a closed catalogue of threats and vulnerabilities.

We live in a technological context in which cyberthreats to information systems are varied, and the risks they contemplate are wide-ranging. Organizations of all kinds have internalized the concept of cyber threat, and what is more relevant, cybersecurity. We have all seen some news that refers to a *cyberattack* received by a company, or institution that, in the best of cases, stops the service provided from a few hours to a few days. In the worst case, the loss of data, the risks to individuals and their rights are irreparable.

The field of high-risk artificial intelligence systems does not escape general cybersecurity threats, *which are not the objective of this guide* and must be taken into account, in the appropriate way. These artificial intelligence systems also add and exacerbate, by their nature, a series of cyberthreats of their own that expose their vulnerabilities and are vectors of specific attacks of **artificial intelligence systems,** especially in the context of high-risk systems. Not taking these threats into consideration, and therefore being exposed to them, poses **a serious risk to any system** and, without a doubt, an **unacceptable action in the context of high-risk systems**.

The potential damage that high-risk systems can cause to the safety and health of people, material damage, privacy, limitations of rights, discrimination, access to employment and a long etc., are of great magnitude, given the context of high-risk consideration that they have.

The cybersecurity measures for AI systems presented in this guide aim to **mitigate** the **risks** that threaten the **rights and freedoms of individuals** and society at large. These can be seriously threatened by the existence of backdoors in the models, the possibility of exfiltration attacks or vulnerability to adversarial attacks.

It is therefore important that this guide is approached with the consideration of **identified risks** that threaten the **rights and freedoms of individuals** and society in general, identified in the risks plan (see risks management guide) for the AI system, and that there is a process that begins in the identified risks, is associated with the vulnerabilities present in this guide and the application of the applicable security controls can be demonstrated. All this is always within the life cycle of the AI system: conception, design, implementation, validation/verification and commissioning.

Cybersecurity plays a crucial role in ensuring that AI systems are resilient to attempts to alter their use, behaviour, and performance or compromise their security properties by malicious third parties who exploit system vulnerabilities. Cyberattacks against AI systems can leverage AI-specific assets, such as training datasets (e.g., data poisoning) or trained models (e.g., adversary attacks), or exploit vulnerabilities in the AI system's digital assets or the underlying ICT infrastructure. Therefore, in order to ensure a level of cybersecurity appropriate to the risks, providers of high-risk AI systems should take appropriate measures, also taking into account, as appropriate, the underlying ICT infrastructure.

Likewise, the AI Act in its article 15, indicates that cybersecurity (oriented to Artificial Intelligence) must be adequate and consistent throughout the life cycle of the system, which requires organizations (especially the provider, but also the deployer) to adequately size the efforts in costs and personnel. Similarly, considering the need to properly **monitor** new forms of attack on artificial intelligence systems is a strongly changing field.

2.2 What are the main threats?

Generally grouped under the concept of adversarial attacks, to which an artificial intelligence system is exposed are:

- Poisoning attacks: They seek to manipulate the AI system during its training or upgrade phase to compromise its behaviour. These attacks can be classified into two main types:
    - Data poisoning: This involves the introduction of malicious or manipulated data into the training data set. The goal is to alter the model's learning so that it generates controlled results or insert backdoors that can be exploited later. Although its effects usually manifest during the inference phase, the attack occurs mainly in the development and training phase of the system.
    - Model poisoning: Unlike data poisoning, in this case the attack directly manipulates the AI model itself, usually during its update or in federated learning environments. In these scenarios, the model is trained in a distributed manner across multiple devices or entities, increasing the risk of tampering. The attacker can modify the model's parameters or insert malicious behaviours that compromise its operation without the need to act on the training data.
- Evasion attacks: In this type of attack, the adversary tries to get the AI system to make incorrect predictions, either globally or in specific cases. These attacks occur in the **inference phase**, when the system is already in **production**, so it is essential that the system has security mechanisms in place to detect them. Countermeasures should be designed during the design and training stages to maximize the resilience of the system to these attacks.
- Investment attacks: Privacy is the main risk in these types of attacks, but combined with other types of attacks (e.g., extraction) they have devastating effects. The attacker's goal is to gain knowledge from the model's training data. Like other attacks, it occurs in the **inference phase** with the **AI system** in **production**, but the system must be designed and trained with the aim of resisting them to the greatest extent possible.

- Extraction attacks: These attacks attempt to gain insight into the model, with the goal of replicating or training a similar model. The most common way is through interactions with the system, although they can also involve lateral mechanisms, such as stealing the model or extracting information from its responses. Extraction attacks can facilitate **transfer** attacks, where, once the model is known, it is possible to apply other types of attacks (evasion, reversal) to the high-risk AI system attacked. Like other attacks, these occur in the **inference phase** with the **AI system** in **production**, but the system must be designed and trained to withstand them to the greatest extent possible.
- Side-channel attacks: This type of attack seeks to obtain information from the AI system through physical characteristics observable during its operation, such as energy consumption patterns, response times, or electromagnetic emissions. Unlike direct attacks on the model or data, these attacks exploit vulnerabilities in the physical deployment or hardware where the system runs. Side-channel attacks can occur both in the **inference phase** with the system in **production** and during training and can facilitate other types of attacks by revealing information about the model's architecture or parameters. It is critical to implement specific countermeasures at the hardware and software level to minimize information leakage through these channels.
- Supply chain attacks: These attacks focus on compromising the integrity of the AI system through vulnerabilities in its supply chain, from development to deployment. The attacker can introduce malicious or tampered with components at any point in the system lifecycle, including:
    - Software components: Introducing compromised dependencies, malicious libraries, or vulnerable code into the development process can create exploitable backdoors or vulnerabilities once the system is in production. This is especially critical in systems that rely on multiple open-source or third-party components.
    - Infrastructure and hardware: Tampering with physical components or computing infrastructure can enable persistent attacks that are difficult to detect and mitigate.

  These attacks can affect both the training and inference phases and require rigorous verification and validation of all elements of the supply chain.
- Denial of Service (DoS) attacks: This type of attack seeks to disrupt or degrade the normal functioning of the AI system, overloading its computational resources or saturating its processing capabilities. Unlike traditional DoS attacks that focus on network infrastructure, AI-specific DoS attacks can exploit unique features of these systems:
    - Computational overhead: This consists of generating queries specifically designed to maximize the consumption of model resources, for example, through inputs that require especially intensive calculations or that trigger the worst cases of system performance. These attacks are particularly effective on AI systems that operate in real-time or have strict latency requirements.

- Attacks of exhaustion of other resources: They focus on exploiting limitations in the resources used by the intelligent system to be able to carry out the task entrusted to it. For example, the system can generate a massive number of files that compromise the operation of the operating system. Like other attacks, they occur primarily in the **inference phase** with the system in **production** but require special considerations during design and development to implement effective protection mechanisms.

Unauthorized third-party access to assets presents in the value chain of a high-risk intelligence system, without being a directly AI-related vulnerability, is a broad attack vector for any of the aforementioned attacks, with the risk of escalating **black box attacks** where the attacker has no knowledge of the system to **white or grey** box where there is partial or even total knowledge. In this guide, the relationship between access policies and high-risk artificial intelligence systems is presented in an Annex.

Another fundamental AI cybersecurity concept is feedback loops, which are particularly important in continuous training systems that continue to learn after their deployment. These loops can result in the amplification of biases or unwanted system behaviour if the output results influence future input data operations.

It is often said that *"a chain is only as strong as its weakest link"*, and this is especially relevant in the context of cybersecurity. Training is a very powerful tool within an organization to establish solid links. The guide addresses in a summarized way, in another Annex dedicated to it, aspects related to how organizations (provider and deployer) should train their teams in the field of cybersecurity oriented to artificial intelligence. This allows for a distribution throughout the organization of the concepts and risks inherent in cybersecurity for high-risk artificial intelligence systems.

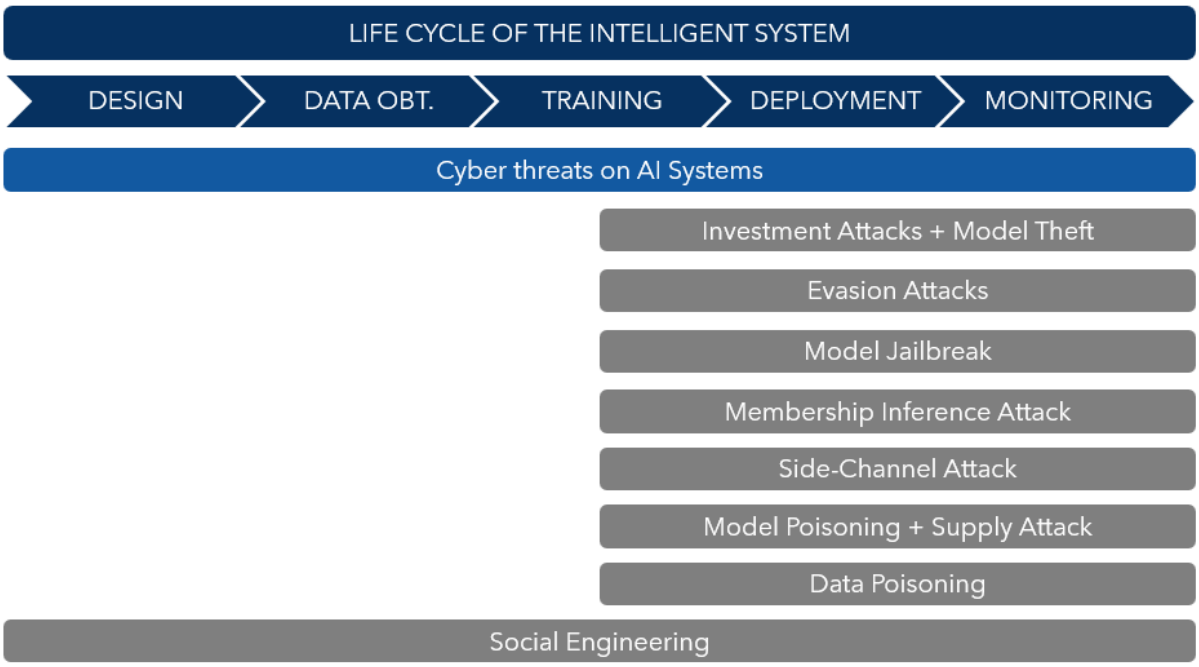

Image 1. Diagram of the main threats within cybersecurity for intelligent systems, relating their moment of appearance to the life cycle of the intelligent system.

3. European Regulation on Artificial Intelligence

The putting into service or use of high-risk AI systems should be subject to compliance with certain mandatory requirements, including cybersecurity requirements. Those requirements aim to ensure that high-risk AI systems available in the Union or whose output outputs are used in the Union do not pose unacceptable risks to important public interests recognised and protected by Union law.

Regulation 2024/1689 of the European Parliament and of the Council of 13 June 2024 (European Regulation on Artificial Intelligence), already in its recital (76) clearly establishes its objectives, indicating that there is specific cybersecurity for artificial intelligence, closely related to the existing concept, but with its own particularities. This section includes the articles referring to the generation of cybersecurity of the AI Act and details in which sections of this guide the different elements of these articles are addressed.

3.1 Preliminary analysis and relationship of the articles

Article 15 establishes the requirements that must be met in terms of three fundamental aspects *"Accuracy, robustness and cybersecurity"*. Accuracy and robustness are specifically addressed in their guides.

Article 15 also states that these systems must be protected against **data poisoning attacks, adversary attacks** in all their variants and must be designed and developed in such a way that they can be resistant to the exploitation of failures **by third parties**. Third-party attacks may be carried out by users, legitimate or not, who could use the AI system's capacity for illicit or illegal purposes that do not correspond in any way to the intended purpose of the system. For example, legitimate users could use the system to extract queries made by previous users.

In this guide, emphasis will be placed on the parts of this article that are specifically oriented to cybersecurity in AI, which are from Article 15, points 1 and 5.

Article 15: Accuracy, robustness and cybersecurity, point 1: Establishes the need for an adequate level of cybersecurity in a uniform manner throughout its life cycle.

Article 15: Accuracy, robustness and cybersecurity, point 5: It focuses on the AI-specific cybersecurity measures and components on which the content is developed throughout the guide.

3.2 AI Act Requirements

AI Act

Art.15 – Accuracy, robustness and cybersecurity

1. High-risk AI systems shall be designed and developed in such a way that they achieve an **appropriate level** of accuracy, robustness, and **cybersecurity**, and that they perform **consistently** in those respects **throughout their lifecycle**.

2. To address the technical aspects of how to measure the appropriate levels of accuracy and robustness set out in paragraph 1 and any other relevant performance metrics, the Commission shall, in cooperation with relevant stakeholders and organisations such as metrology and benchmarking authorities, encourage, as appropriate, the development of benchmarks and measurement methodologies.

3. The levels of accuracy and the relevant accuracy metrics of high-risk AI systems shall be declared in the accompanying instructions of use.

4. High-risk AI systems shall be as resilient as possible regarding errors, faults or inconsistencies that may occur within the system or the environment in which the system operates, in particular due to their interaction with natural persons or other systems. Technical and organisational measures shall be taken in this regard.

The robustness of high-risk AI systems may be achieved through technical redundancy solutions, which may include backup or fail-safe plans.

High-risk AI systems that continue to learn after being placed on the market or put into service shall be developed in such a way as to eliminate or reduce as far as possible the risk of possibly biased outputs influencing input for future operations (feedback loops), and as to ensure that any such feedback loops are duly addressed with appropriate mitigation measures.

5. High-risk AI systems **shall be resilient against attempts by unauthorised third parties** to alter their use, outputs or performance by exploiting system vulnerabilities.

The technical solutions aiming to **ensure the cybersecurity** of high-risk AI systems shall be **appropriate to the relevant circumstances and the risks.**

The technical solutions to address AI specific vulnerabilities shall include, where appropriate, **measures to prevent, detect, respond to, resolve and control** for attacks **trying to manipulate the training data set** (data poisoning), **or pre-trained components used in training** (model poisoning), **inputs designed** to cause the AI model to make a mistake (adversarial examples or model evasion), **confidentiality attacks or model flaws.**

3.3 Correspondence of the article with the sections of the guide

The following table details the sections of this guide that address the different elements of this article:

Article	AI Act requirements	Guide Section
15.1	Level of accuracy, robustness and cybersecurity	Section 4.1
15.5	Systems resistance to alterations of use	Section 4.2
15.5	Prevent and control dataset manipulation	Section 4.3
15.5	Protection against adversarial attacks	Section 4.4
15.5	Protection against system defects	Section 4.5

4. How to approach the requirements?

4.1 Level of cybersecurity and consistency

In the European Regulation on Artificial Intelligence, Article 15 on accuracy, robustness and cybersecurity states the following:

AI Act

Art.15.1 – Accuracy, robustness and cybersecurity

High-risk AI systems shall be designed and developed in such a way that they achieve an **appropriate level** of accuracy, robustness, and **cybersecurity**, and that they perform **consistently** in those respects **throughout their lifecycle**.

Throughout this section we will indicate the high-level measures that are considered to be able to achieve the objective described in the article, taking into account that they will be further developed in specific aspects in the following sections. The focus of this section is to enable the provider or deployer to understand the scope of AI cybersecurity controls required to cover the vulnerabilities detailed in the guide that apply to them. It is important to be aware that information is presented in the guide that in no case is intended to be exhaustive regarding existing vulnerabilities, but that these are shown as a list of those possible, and that, given the areas indicated, (data poisoning, adversarial attacks or manipulation of defects) these should be expanded according to the needs of the AI system for the intended purpose. Especially with the aim of covering the risks in terms of rights and freedoms.

In each element of the life cycle of a high-risk AI system, a level of cybersecurity is established in accordance with the intended purpose of the system, establishing appropriate measures. Depending on the phase of the life cycle, the measurements must be carried out by provider (design, development) or operation (deployer).

In this context, the "life cycle of the AI system" is understood as the duration for the AI system, from its conception, design, implementation and commissioning to its retirement. Different phases can apply to different AI systems depending on their intended purpose and way of commercialization, so in general terms we consider these phases:

- Conception.
- Design.
- Implementation.
- Verification and testing.

- Commissioning.
- Post-market monitoring.
- Withdrawal.

This life cycle described here applies to the AI system, exclusively, and in no case can it be considered representative of other elements, such as the life cycle of the data (which is addressed in the data guide), or the life cycle of other information systems or infrastructure used to support the AI system. Similarly, the relationship of cybersecurity as the establishment of security controls is directly related to the risks management system (see risks management guide), especially those related to the rights and freedoms of natural persons, damage to health, serious damage to property and/or the environment. The protection of the system against the type of threats described in this guide is directly related to the robustness of the high-risk AI system (see Robustness guide).

4.1.1 Applicable measures

The technical measures will be translated into **security controls**, applied to the inventory of assets to defend against threats and prevent risks, so that they are so according to the AI Act in its Art. 15:

AI Act

Art.15.5 – Accuracy, robustness, and cybersecurity

[...] **appropriate to the relevant circumstances and the risks.**

Once the measures have been defined and implemented, they must **be guaranteed and maintained throughout the life cycle**, so that the level is adequate, not only to the threats and security measures necessary at the time of the implementation of the system, but also to adapt and adapt to new threats that may appear. without **degrading** the level of cybersecurity.

Provider

The provider must take the following organizational steps to ensure that the high-risk AI system has an adequate level of cybersecurity.

- Globally plan the level of cybersecurity applied to AI during design and development, following the points detailed in this guide.
- Involve the data protection officer (DPO), from the design of the AI system and for cybersecurity planning, as an interlocutor within the working group established to develop the planning. It must also be present when specific cybersecurity decisions are made to check its involvement with data protection.
- The instructions for use of the artificial intelligence system must be accompanied by high-level recommendations on AI-focused cybersecurity measures, specific to the AI system and its intended purpose, to be taken into account by the deployer. In

addition, it is recommended that the AI system, through the interaction interface for the end user, has recommendations for use in terms of cybersecurity, so that they are accessible to the end user of the system. It is also possible to include an interaction mechanism (such as a confirmation window or similar) each time it is accessed to ensure that the cybersecurity instructions have been read.

- Within the design process, those responsible for monitoring the cybersecurity measures applied to the AI system must be established. Define a dashboard for monitoring the operation. The mandatory cybersecurity indicators for AI to be included in the dashboard for monitoring the operation of the AI system will be established. The minimum frequency at which a security audit of the AI system will be carried out will be established, regardless of whether it is internal or external.
- The cybersecurity measures applicable throughout the life cycle of the artificial intelligence system, which are detailed in this guide, will all be applicable by the provider, if it provides the deployer with the AI system as an MLSaaS (*Machine Learning as a Service*).) marketing and commissioning forms involving automatic installation by the deployer where there is no system configuration.
- If the AI system is to be delivered to the deployer in an *on-premise* or *in-cloud* format, managed by the deployer, the provider must provide appropriate instructions for performing system protection especially during the inference time of the AI system in production. The installation process of the AI system must have mechanisms that guarantee that the installation takes into account the instructions in a mandatory manner, either through automatic or semi-automatic script procedures, the obligation to have the instructions open for the process before continuing, the explicit request to read them or the reference to the specific reading to be known for each step of the process.
- High-risk AI system upgrades should be addressed with all applicable measures described in this guide.

In addition to these organisational measures, the provider must align the following technical measures:

- In the installation and/or configuration processes of the Artificial Intelligence system and in the instruction manual, the provider must include information regarding the specific cybersecurity risks of the system and how it is protected.
- During the life cycle of the high-risk AI system, tools that allow security testing to be automated should be used. The use of these tools, from the beginning of the development of the system, allows a security-oriented design, conceiving the attack tests of the system in parallel to its development and life cycle. There are both open-source variants on the market, a wide variety of tools that comply with the indications.
- Updates to the AI system must be developed by applying the technical measures described in the following sections so that the level of cybersecurity applicable to AI is consistent, continuous and does not degrade with the application of these measures.

Example - Attendance at work

To address the execution of cybersecurity measures in AI, within the analysis and design team of the high-risk AI system, the responsibilities and the people who will carry out the actions are established. The profiles present in the organization have been checked and the possibility of expanding its staff to cover the necessary efforts for the system's AI-focused cybersecurity has been assessed. The provider has concluded that, with their current staff resources associated with risks management, AI system development and analysis, as well as the implementation team, they can carry out the necessary aspects.

The DPO has been included in the work team from the design phase, ensuring their participation in all cybersecurity decisions. A comprehensive risk management plan has also been developed that specifically addresses threats related to AI systems and the software delivery security control process has been updated to include specific testing against these threats (this plan is reviewed and updated quarterly).

Automated vulnerability scanning tools have been implemented and run on each new *system release* and when new training data sets are added. These tools include specialized AI security scanners and continuous model monitoring systems. An audit program has been established with quarterly internal reviews and a mandatory annual external audit, documenting findings and corrective actions in the risks management plan.

After defining all the above points, the provider has scheduled training for all the personnel involved, which allows consolidating the knowledge and process of cybersecurity in AI within the organization, using this guide and the approach indicated in section 7.2 of this guide as a basis. Specific cybersecurity training for AI systems has been included in the provider's security plan, with biannual updates to the training program to incorporate new threats and countermeasures.

Deployer

The organisational measures in terms of cybersecurity applicable to the deployer depend on the level of involvement of the in the life cycle of the high-risk Artificial Intelligence system.

- You'll need to distribute the installation and configuration information across your organization.
- If the system is in its facilities, or is managed by it (on-premise or in-cloud managed by the deployer), it must establish, **within** its **general cybersecurity policy**, all the **assets** associated with the Artificial Intelligence system that could be applicable. These can be data, model, processes, environments/tools, and artifacts. In addition to the identification of the actors in your organization on each of these assets. Thus, the considerations indicated in this guide will be applied to the set of assets to the conditions of use of the AI system, according to its intended purpose, especially those focused on the **inference** phase with the AI system **in production**.

As a technical measure, it is important to have a centralised repository where the information can be managed, accessible to those actors identified in the development of the artificial intelligence system.

Example - Aid granting automatic system

Deployers can be the General State Administration and other Public Administrations, such as Autonomous Communities or local entities. In either case, they can opt for an *on-premise installation* of the system. The *provider* has provided, on the one hand, an **instruction manual** and on the other, the **configuration of the system** (which will be carried out on-premise).

Both documents specifically include cybersecurity information related to both processes. Two actions are carried out on it:

- The Public Administration staff in charge of managing the response of the AI system have received the instruction manual and specific training in cybersecurity for AI according to their scope.
- The personnel of systems administration and related IT technologies have received the **system configuration** manual and the aspects focused on cybersecurity. Particularly in Local Entities, it is considered to update the knowledge of personnel in the area of AI cybersecurity with training adapted to their technical knowledge. These specific personnel also receive this guide.

4.2 Systems resistance to alterations of use

As stated in the European Regulation on Artificial Intelligence in paragraph 5 of its article 15 on accuracy, robustness and cybersecurity:

AI Act

Art.15.5 – Accuracy, robustness and cybersecurity

High-risk AI systems **shall be resilient against attempts by unauthorised third parties** to alter their use, outputs or performance by exploiting system vulnerabilities.

The technical solutions aiming to **ensure the cybersecurity** of high-risk AI systems shall be **appropriate to the relevant circumstances and the risks.**

The measures described in the previous section are accompanied by the realization of inventories of the **assets** of the artificial intelligence system during their life cycle, and the **actors** that interact with these assets. Thanks to these inventories, two aspects indicated by the AI Act are covered, as indicated at the beginning of point 5 of Article 15:

- Establish the basis for protecting the system from attempts by unauthorized third parties to alter the use and operation of the artificial intelligence system. By having inventories of assets and actors, access policies can be applied (see Annex I).
- To be able to size technical solutions that guarantee cybersecurity on the identified assets, as described, identify the vulnerabilities on them and their controls (see in detail from subsection 4.3 data security vulnerabilities and controls to subsection 4.5 Attacks on AI system flaws).

4.2.1 Applicable measures

Provider

The provider's organizational measures related to asset and actor inventories are:

- All **actors** involved in the AI system design and development process should be inventoried. The actions of unauthorized third parties not only refer to external third parties but also to internal third parties with inadequate functions.
- Once the **actors** have been defined, establish the level of scope and permissions that each of them can have and define them appropriately. See 7.1 Annex I: Access policies. An approach to high-risk artificial intelligence systems.
- The documentation of the AI system should clearly reflect the roles involved in using the tool, and the permissions that each of them will need to have. The installation instructions or steps must reflect in detail the relationship between the actors and the elements and actions of the installation.
- Planning and carrying out the **inventory of assets** in the development of the Artificial Intelligence system, especially considering the **tools, data, processes and model.**
- On the asset inventory, the **associated vulnerabilities** in those assets will be identified. This identification is made in successive sections of the guide.
- The AI system's instructions should include a **list of the threats/vulnerabilities** inherent in the system during its use.
- Develop a rigorous threat model to understand all possible attack vectors and with the consideration of the risk for the violation of rights and freedoms of natural persons, damage to health, serious damage to property and/or the environment fully aligned with the risks analysis carried out (see guide to the risks system).
- The established threat model, for each actor/asset/vulnerability identification, recommends that you maintain a RACI (*Responsible, Accountable, Consulted, Informed) scheme*. In the following sections (from section 4.3 to section 4.5), the vulnerabilities will be defined so that the system provider can develop a rigorous threat model for application to its AI system and establish the necessary technical controls, adding the characteristics of a RACI matrix as proposed.

In this process of inventorying the **actors** involved and **actives of the system**, and although its more specific scope is indicated within the data guide, the DPO data protection officer or his team must be involved. In this way, it will be possible to establish an inventory as complete as possible on which the threat model can be established.

The technical measures that the provider must carry out with respect to these inventories are merely to support the organizational ones:

- The inventory of assets will have a computer support appropriate to the dimensions and capabilities of the organization. It can range from an office file or a simple database where AI asset data is stored to existing market tools and solutions for IT asset inventory management. Likewise, in order to adequately size the system, the level of concurrent accesses and seasonality (those times when the volume of accesses may be more pronounced) will be taken into account to guarantee the availability of the system and avoid blockages due to excessive connection attempts.
- Have the necessary technical means so that the established access policies can be executed (See 7.1 Annex I: Access policies. An approach to high-risk artificial intelligence systems).
- Document and asset management system indicated, must be centralized with updated and available inventories.

Example - Aid granting automatic system

The AI system provider performs the asset and actor inventory actions:

- Assets: training data, testing and validation; the data is especially sensitive and **will use completely anonymised data** provided by the General State Administration, these represent historical data on granting/refusing aid for the last 10 years. Selected model. Libraries and software that interact directly with the AI model. Tool for version control and development environment. Internal development and commercial documentation. External documentation.
- Actors: Data scientists, ML and AI experts, software developers, project managers and systems team. It has also identified within its sales team those who will be in charge of promoting the product.

The provider, as a *SME*, must specify the multiple roles that part of its staff performs on the assets, building a matrix of relationship between assets and actors. Thus, for each asset, the related actor has been identified, indicating whether access is available. This matrix has been made using a conventional office tool to record the data and has been shared in the organization through the collaborative internal Wiki.

With this matrix, a technical market solution has been found that allows the implementation of access policies.

Deployer

For the field of actors and assets, the deployer must simply know what directly applies to it.

Organizationally, it must consider:

- Integrate the documentation provided by the provider on the roles of the different actors with its internal organizational chart. Define and establish the functions and the people who will cover them. Establish mechanisms to replace these people when they are alone in charge of a function, so that unexpected departures or the departure of an employee keeps the system and processes without discontinuities. As long as the organization and the size of the company allow it, the different roles, as defined by the provider, will be assigned to different people.
- If the format of use and commercialization of the AI system contemplates that the deployer has the **AI system as an asset** of its own facilities, it will have to consider the AI system as an asset that must be protected in terms of cybersecurity.

There are no relevant technical measures in this area for the deployer beyond those necessary for the correct implementation of security policies (see Annex I).

4.3 Data security vulnerabilities and controls

The identification of vulnerabilities associated with the training data and the establishment of the necessary security controls to prevent its alteration or manipulation allow the artificial intelligence system to be protected from poisoning attacks[1].

The European Regulation on Artificial Intelligence places special focus on vulnerabilities associated with data when it establishes in its article 15.5 on accuracy, robustness and cybersecurity:

AI Act

Art.15.5 – Accuracy, robustness and cybersecurity

The technical solutions to address AI specific vulnerabilities shall include, where appropriate, **measures to prevent, detect, respond to, resolve and control** for attacks **trying to manipulate the training data set** (data poisoning)

The attacker's goal is to alter the training of the AI system with the intention of controlling its results or compromising its operation. Unintentional changes to the training data are addressed in the data guide. In the system user manual, the provider must list the vulnerabilities associated with the data. The security control applied must be identified for each vulnerability. All vulnerabilities and controls associated with data are mainly present

[1] The highlighted words and descriptions correspond to terms that are developed in the glossary.

in the design and training phase of the system, where the risk of malicious data introduction is very high and the impact very high.

4.3.1 Applicable measures

Provider

The provider must identify vulnerabilities applicable to the training data sets. The list must be defined, based on the intended purpose of the AI system. Below is a list of vulnerabilities associated with tampering with training data, validation, and testing:

- Model easy to poison.
- Insufficient data to increase resistance to poisoning.
- Inadequate data access rights management
- Inadequate data management
- Uncontrolled data sources
- Lack of control to detect poisoning in the system.
- Do not detect poisoned samples in the training set.
- Using incorrect data labelling source

It is outside the scope of this guide, but it is important for the provider to consider, within the lifecycle of the data, the secure deletion of the data, once it is no longer in production or the system enters an evolution. In the data guide, this aspect is dealt with in a special way.

The following table lists some of the most relevant vulnerabilities and the security controls that are in place to mitigate their exploitation, at the time of writing this guide. Given the dynamic nature and rapid evolution of cybersecurity threats, it is advisable to always be aware of the most up-to-date and reference technical information. Recommended resources include the OWASP Machine Learning Security Top Ten, the OWASP Top 10: LLM & Generative AI Security Risks, and the OWASP AI Top Ten, which provide up-to-date analysis and measures against emerging risks in AI systems.

Vulnerability	Security Controls
Model easy to poison	In the design phase, choose or define a more resilient model. **Use bagging, boosting, or TRIM techniques** during training to strengthen the robustness of the model. To mitigate the effect of randomness and ensure reproducibility, use defined initial seeds when necessary. ISO27001/2 - NIST 800-53: Conduct regular and regular audits on data management and develop action plans to correct deficiencies.
Insufficient data to increase resistance to poisoning.	Expand the dataset using data augmentation techniques, for sets that are too small or insufficient for the intended purpose
Inadequate data access rights management	Appropriate access policies, see Appendices. 7.1 Annex I: Access policies. An approach to high-risk artificial intelligence systems.
Uncontrolled data sources	All data sources should be kept under version control, so that you have a trace of changes to the data and the actors who have made them.
Uncontrolled Data Sources (continued)	ISO27001/2 - NIST 800-53: Various mechanisms are proposed. Classify data appropriately using statistical tools, considering the intended purpose and properly review the classification; data must always be encrypted in transit, use encryption with a guarantee of authenticity, encrypting data using ciphers in accordance with current standards; always store encryption at rest data sources, at the hard disk and individual file encryption level, also using current standards. Use data loss prevention (DLP) tools on sensitive and relevant data.
Lack of control to detect poisoning in the system	**Apply the STRIP technique**, which involves perturbing the data in a controlled manner and observing the variance between perturbed and undisturbed data to identify potential poisonings.

Vulnerability	Security Controls
Not detecting poisoned samples in the training set	Use data sanitization techniques to detect and eliminate outliers: Z-score, Local Outlier Factor, Isolation Forest. Before deleting any data, perform a thorough diagnosis to **avoid deleting valid data** at extreme intervals: 1. Previously verify the hypothesis about the distribution of the variable and, if the distribution is not normal, avoid the use of Z-score. 2. Diagnose the nature of the outlier to decide whether it should be removed or kept. To detect changes in the samples, certify the distribution of the training data by marginal (empirical statistical distribution of each variable in the database, whether numerical or not), and multivariate (by bivariate or more complex relationships) periodically verify that it remains stable. In the event of significant changes in the distribution of one of the variables or in the relationship between several, drill down to identify the changes that have occurred in the data. Analyse the impact of datasets on model accuracy, retraining or continuous learning processes: RONI (reject on negative impact) or tRONI (public-aware RONI)
labelling	ISO27001/2 - NIST 800-53: Management of all interconnections with external systems from which data is received and reviewed. Establish a formal process to conduct a review of all interconnections and regulate and monitor their use. Use label flipping techniques, based on k-NN (k-nearest neighbours) identification in the dataset. This technique can be used when the labelled datasets are unreliable, but there is no alternative in them.

Table 1 Poison Attacks: Security Vulnerabilities and Controls

All controls and vulnerabilities associated with data are mainly focused on the **design and training phase of the system**, where the risk of introducing a poisoning attack is very high and the impact is very high.

In the **AI system user manual**, the provider must list the vulnerabilities associated with the training data, identifying for each vulnerability the **security control** to be applied. In the event that the deployer of the high-risk AI system carries out part of the training before its implementation, the way to apply the security controls must be detailed.

Example - Attendance at work

The system will work with biometric data obtained from employees for training, testing and validation, in such a way that it is able to properly identify employees from the various locations of the cameras (or biometric sensors) to record their attendance at work.

The risks analysis has identified the possibility of a user acting maliciously to alter the labels of the biometric data. This could allow you to associate your own data with that of other employees, thus compromising the system's detection.

Since the main source of this threat is data poisoning, STRIP disturbance techniques are used during training, specifically intended to detect poison triggers (or Trojans) in the dataset.

In addition, the design team decides to protect the AI system against a possible Label Flipping attack (see Glossary) in which the attacker has maliciously changed image labels of different employees.

Deployer

The main organizational action that the deployer must take to address the protection of artificial intelligence systems against poisoning attacks is to read and compress the instruction manual. As a general rule, the deployer does not participate in the design, development and training phase of the system.

However, if the format of use and marketing of the AI system contemplates that the deployer uses the AI system by carrying out a **final training** of the same, or in continuous training systems, this being previously partially pre-trained by the provider, all the cybersecurity measures applied to AI of the provider in terms of data would be applicable. In this scenario, by using a general-purpose artificial intelligence (GPAI) system, the deployer **would have provider responsibilities** (organizational and technical) in all the aspects described in the previous section.

4.4 Protection against adversarial attacks

In this guide, adversarial attacks of the data poisoning type have been separated from those related to the model itself. In addition, the system is more exposed during the **inference** phase, once it is **in production** (after commissioning) to this type of attack. Unlike the exposure related to data.

The European Regulation on Artificial Intelligence in Article 15, on accuracy, robustness and cybersecurity, paragraph 4 states:

AI Act

Art.15.5 – Accuracy, robustness and cybersecurity

Measures to prevent, detect, respond to, resolve and control for **attacks** trying to manipulate *[…]* **inputs** designed to cause **the AI model to make a mistake (adversarial examples or model evasion) […]**

To this end, the necessary **security controls** must be established to prevent the entry of certain input data from causing the model to make an error or allowing information to be extracted from it. These are adversarial attacks of the oracle type, extraction, inversion or evasion attacks on the high-risk artificial intelligence system.

4.4.1 Applicable measures

Provider

The provider should identify vulnerabilities related to the manipulation of input data, which can cause the model to make mistakes or allow inference of the data used in training through queries to the model. The list of vulnerabilities should be assessed and prioritized based on the intended purpose of the AI system. Below is a list of vulnerabilities:

- Transfer of adversarial attacks.
- Inadequate management of access rights to the model Failure to consider the possible attacks to which the AI system may be exposed. Absence of a security process that maintains the level of protection of the components of the AI system. Weak mechanism of access control to the components of the AI system.
- Failure to detect anomalous data inputs, increasing exposure to membership attacks to infer data or reversal attacks.
- Not considering the possibility of evasion attacks in the design and implementation of the model.
- Not considering in the design and implementation of the model the possibility of adversarial attacks, for example: **misclassification, misclassification origin/target, random misclassification and confidence reduction.**
- Use of widely known models in the AI system, which allows full or partial replication and thus facilitates the transfer of adversarial attacks.
- Data input fully controlled by the attacker (without preprocessing or by violation of the input mechanism), allowing the creation of input data and output pairs. Exposure to investment attacks. Both *Membership Inference Attack*, which allows us to know whether input data has been part of the training data set, and *Property Inference* attacks that seek to know the data used in the model. This risk affects the privacy of the personal data used in training and, therefore, must be taken into

account as a cybersecurity risk, as established by the ISO7IEC 27001 standard. Both attacks can precede or be aimed at **reconstructing** the entire set of training data.

- Excess information available in the model. Excess information about the model in its outputs.
- Failure to consider the possibility of extraction attacks.
- Theft or compromise of the model.

In the system user manual, the provider must list the vulnerabilities associated with the manipulation of the data in which the system is exposed. Those measures that have been put in place during the training of the model should be indicated. If the system is not offered as an MLSaaS, or by its nature could be physically located on the deployer's premises, the instruction manual and its documentation will need to refer to the application vulnerabilities and how to apply those security controls that may be your responsibility.

The technical measures applicable in this case are the security controls that prevent or mitigate the exploitation of vulnerabilities. If any of the measures **must be carried out in the process of installation and use** of the system by the deployer, it must be indicated in the instruction manual.

Below is a detailed list of vulnerabilities:

Vulnerabilities	Security Controls
Transfer of Adversary	As far as possible, you should avoid using models in the AI system that are widely used and known, as these may contain known vulnerabilities that facilitate transfer attacks. It is also advisable to avoid starting from base models or frameworks in versions that have already been compromised, which allow the attack to be identified and exploited this vulnerability. All families of AI systems and models have known vulnerabilities or aspects where they are more susceptible to certain types of adversarial attacks. Therefore, it is essential to know, identify and protect the AI system from the design, implementation, verification and validation phase to its commissioning and operation phase. In addition, to reduce the risk associated with the use of widely known models: - It is recommended to apply obfuscation or masking techniques that make it difficult for an attacker to identify and exploit the model. - Proactively assess the robustness of the system against transferred attacks through specific tests and continuous evaluation methodologies.

Vulnerabilities	Security Controls
Inadequate management of access rights to the	Access policies, see section 7 Annexes. 7.1 Annex I: Access policies. An approach to high-risk artificial intelligence systems.
Lack of consideration of the possible attacks to which the AI system may be exposed	Integrate security specificities for AI systems into the awareness strategy and ensure that all parties involved receive it (see Annex II: Training in cybersecurity and high-risk artificial intelligence systems). Monitoring the state of the art of adversary attacks: https://csrc.nist.gov/publications/detail/nistir/8269/draft
Absence of a security process that maintains the level of protection of the AI system components	ISO27001/2 - NIST 800-53: Conduct periodic and regular audits on the model based on the measures detailed in this guide and make action plans after each case. Perform a blue team/red team approach; the blue team develops the model, and the red team tests it with various attack formats of the opponent type to it (white, grey, black).
Failure to detect anomalous data inputs	Implement specific tools to detect if an input value may be an adversarial example. An effective strategy is to add adversary detection subnets to the system, which operate separately from the main model. These subnets are specifically trained to identify patterns associated with adversarial attacks, analysing particular characteristics of suspicious entries and alerting to possible manipulation attempts. In addition, metrics and confidence thresholds should be implemented in the classifications and distribution of input data. Configuring confidence levels allows you to generate alerts when the results of the model show uncertainty or values far from the expected behaviour, indicating possible attacks or anomalies in the inputs. Similarly, setting thresholds on statistical distributions helps identify deviations in the data, such as outliers, that could be related to adversarial attack attempts. Other complementary measures include: - Mechanisms for detecting OoD (Out of Domain data) elements: Identify input data located at the borders of the classification domains or outside the solution space provided by the model. For non-classification-based systems, locate and analyse

Vulnerabilities	Security Controls
	outliers, taking into account their relationship with other relevant variables of the system. - Control system of call sources: Monitor and detect suspicious requests that attempt to explore or manipulate the model's solution space. Access from these sources must be blocked temporarily or permanently or slowed down by increasing response times. - Confidence levels and warning mechanisms: Establish a mechanism that associates each result of the model to a level of confidence. Low confidence values should trigger automatic alerts to warn of potentially anomalous entries. The system interface should include: - Records of abnormal entries: A traceability system that allows suspicious entries to be recorded and analysed. - Visual alerts and user warnings: Make it easier for users to be warned of possible anomalous entries in real time. To facilitate implementation, open-source or commercial tools specialized in the detection of anomalous inputs and adversarial examples can be integrated. Some examples of open-source tools are listed below: • IBM Adversarial Robustness Toolbox (ART): An open-source framework that provides methods for evaluating and improving the robustness of models against adversarial examples. • Foolbox: Open-source tool for robustness testing and generation of adversarial examples in machine learning models. • TensorFlow Privacy and PyTorch-based libraries: Solutions that help apply detection and mitigation techniques, such as trust analysis and outlet monitoring.
Not considering the possibility of evasion	Design a resilient model against evasion attacks through input randomization, **input gradient regularization, and**

Vulnerabilities	Security Controls
attacks in the design and implementation	**defensive distillation**. Use reference resources for the state of the art of attacks by reference adversaries of which there are different formats, many of them open source. These toolkits provide appropriate sets of techniques to verify the model properly against adversarial attacks.
Not considering the possibility of evasion attacks in the design and implementation of the model	MITRE ATLAS (https://atlas.mitre.org/) is an extensive knowledge base with up-to-date threat information and case studies based on real-world observations and operational AI systems.
Misclassification and random misclassification	Attackers generate a sample that is not in the input class of the target classifier but is classified by the model as that particular input class. In the random variant the result can be anything other than the correct classification. Use of the Highly Confident Near Neighbour (HCNN) framework. Explanations of individual decisions based on attributions made by the AI system (Attribution-driven Causal Analysis).
Origin/Target Misclassification	The goal of the attack is to get a specific result (or ranking) for a given entry. To mitigate this: Feature Denoising A technique that especially mitigates this type of attack is to include known adversary examples in the training set so that the system is trained with adversary examples built for this purpose. Feature squeezing.
Reduced trust	This type of attack is intended to reduce the system's confidence in correct classification, so that in this way the perception that acts in the monitoring process is influenced. It can be mitigated by controlling entries, with the decrease in entries admitted to the system in a specific format or coming from a suspicious IP address, source, or demarcation. These entrances may be blocked in case of risk.

Vulnerabilities	Security Controls
Use of widely known models in the AI system facilitating the transfer of adversarial attacks	Use models with less transfer risk. During the process of choosing the model, know and analyse the transfer risks or, in general, its popularity and widespread use, in order to select the one that has a lower risk of transfer.
Use of widely known models in the AI system facilitating the transfer of adversarial attacks	Do not reuse models directly from the internet, without checking and validating them. As far as possible and balancing the intended purpose and protection, select models in which the exposure of adversary attacks, being known, has established security controls.
Input data wholly or partially controlled by the attacker	Apply modifications to the input data, performing preprocessing operations: to the extent that it does not compromise the accuracy of the system, perform a normalization of the data, eliminate noise from the data. Removing values outside the expected ranges can mitigate the attacker's control of input data. NLP systems can be very vulnerable to changes in the input data. The generation of adversarial texts during training using, for example, CAT-Gen. Maintain a history of input data that allows the detection of extremely close input data and raise an alarm in the system by levels (recording, notification and blocking). If data is fed directly into the model by the potential attacker, this data entry must be controlled by appropriate access policies (see Appendices 7.1 Annex I: Access policies. An approach to high-risk artificial intelligence systems). Access to enter data directly into the model should not be public unless strictly necessary, and in that case, they should be controlled as indicated.
Vulnerability to Membership Attacks	Research and articles published in this field indicate that the use of differential privacy techniques is an effective mitigation of these types of attacks. It is also possible to create private datasets, differentially derived from unprotected data. To do this, tools that allow these actions can be used.

Vulnerabilities	Security Controls
	The use of Neural dropout or Model stacking can increase the resilience of the AI system against this attack.
Property Inference Attacks	Establish a limit per origin (IP, user, device) and have alarm and origin blocking mechanisms. All inputs from accessible data to the model must have a validation mechanism.
Not considering the possibility of extraction attacks	Like transfer attacks, the information returned by the system must be controlled, as well as sources with *too many* queries, as it is an attack vector for extraction attacks. There are tools that, within their capabilities, allow defences to be generated by modifying the output of the model by rounding or adding statistical noise, thus making it difficult for the attacker to know the model's reasoning. The use of these tools is recommended for this type of security control.
Model theft	Model versioning, to prevent maliciously modifying the model at any point in the lifecycle from conception and design, training, and validation to production. As indicated in the case of transfer attacks, it is advisable that the exposure of the model (even the publication of its source code) does not pose a vulnerability in itself. This is achieved by applying the criteria of protection against other vulnerabilities throughout the system, in such a way that even at a level of exposure of the model it does not pose a vulnerability.

Table 2 Model Adversarial Attacks: Security Vulnerabilities and Controls

Example - Aid granting automatic system

During the development of the AI system, the provider has established a team to develop the system and another to test it, a process based on blue team/red team, in such a way that each iteration of the model (blue team) is tested (red team).

To this end, the red team has designed specific tools that allow security tests to be automated, a series of tests to the AI system that are verified every time the blue team makes a change to the model. These automated tests replicate attacks on known vulnerabilities and must be continuously updated when new vulnerabilities are known that could potentially be exploited.

The focus has also been on the control of input data, since it is the applicants who enter the data into the system. An OoD (Out of Domain data) control mechanism is available when receiving the data. This procedure has been selected because the data entered are within sequences of structured and tabulated values, with numerical values or discrete series and allow controls to be established on suspicious entries.

Given that the system has been trained with data provided by the General State Administration, based on anonymized data histories over the last 10 years, the vulnerability of membership inference attacks has been identified as a relevant risk despite anonymization. To protect the system, differential privacy techniques have been used. The initial dataset has been extended using tools that allow the data to be augmented by injecting noise into them. These tools are selected as being suitable when it comes to avoiding the risk of investment attacks.

Considering the misclassification origin/target vulnerability, which would allow an applicant to obtain help with a malicious input pattern, the Feature Denoising technique has been applied to the input data. The model has also been trained with adversarial examples.

Deployer

In this area of attacks, the **marketing model** of the high-risk artificial intelligence system **determines** the scope of the cybersecurity-related actions that the deployer must carry out.

In any case, and according to the marketing model of the AI system, these will be **clearly** defined in the **instruction manual**. Organizationally, the deployer must know and analyse the vulnerabilities and controls that may be his responsibility and assign human and technical resources to cover them.

4.5 Attacks on AI system flaws

Paragraph 5, Article 15 on accuracy, robustness and cybersecurity of the AI Act states that model defects must be taken into account when considering:

> **AI Act**
>
> Art.15.5 – Accuracy, robustness and cybersecurity
>
> ... prevent, detect, respond to, resolve and control for **attacks** trying to manipulate [...], **inputs** designed to cause the **AI model to make a mistake** (**adversarial examples or model evasion**), confidentiality attacks or **model flaws**.

Attacks aimed at discovering and exploiting defects in the AI system take advantage of vulnerabilities present at two levels: the **defects of the model used**, and the **defects derived from its integration into the** software environment that surrounds it and as it has been **configured**.

- Defects of the model: These are vulnerabilities intrinsic to the design, training or architecture of the AI model. They can arise due to errors in model design, incomplete or biased training data, or inappropriate hyperparameter configurations. These flaws allow an attacker to manipulate model predictions, generate controlled results, or force systematic errors.
- Defects in model integration: They arise when the model is incorporated into the software or hardware ecosystem in which it operates, due to incorrect configurations, insecure dependencies, or failures in implementation and deployment. These flaws can be exploited to alter the functionality of the system, cause information leaks, or facilitate other types of attacks, such as evasion and extraction attacks.

These flaws can be detected and exploited through oracle attacks, which involve interacting with the system to infer its inner workings and discover its vulnerabilities. Oracle attacks, in turn, can serve as a starting point for more specific variants, such as the adversary attacks (evasion, extraction, or reversal) already described in this guide.

4.5.1 Applicable measures

Provider

Regarding the exploitation of model defects, the provider is responsible for identifying and mitigating vulnerabilities that the model may have, and for its integration with the technological ecosystem. This process must cover the entire life cycle of the system, from its design and development to its implementation and operation, ensuring a continuous review of possible defects. The set of measures implemented not only strengthens security, but also meets the requirements of Article 15 of the European Regulation on Artificial Intelligence.

Below is a list of vulnerabilities that can occur in this area.

- Intrinsic defects of the model.
- Existence of biases in the model or in the data.
- Commitment of the framework used for the AI system model. I understand this as the set of libraries, elements or software artifacts defined together and maintained by a third party, intended for the implementation of AI systems.
- Existence of rear doors in the model.
- Exposure and access to the model code.
- Unknown vulnerabilities in relation to potential model defects.

The specific vulnerabilities associated with adversary attacks are specific to the model, but given their nature, they have been discussed in detail in the previous sections.

In the following table we can see the security controls applicable to the above vulnerabilities.

Vulnerabilities	Security Controls
Intrinsic defects of the model	When selecting a model for the AI system, the provider must be aware of and investigate its intrinsic defects, both those specific within the family to which the model belongs (CNN, Random Forest, KNN, etc.) and those applicable to the intended purpose. These defects must be inventoried in the threat model and the measure applied to mitigate their exploitation as vulnerabilities must be documented. The information collected will depend on the model used and its known defects.
Existence of bias in the model, or in the data	While the existence of biases may be a flaw exploitable by the attacker, it is not the goal of this guide and is described in other guides in more detail (data and accuracy guide).
Model Framework Commitment	Models are within frameworks used for the development and deployment of AI systems. To mitigate the risk of compromise, it is essential that the frameworks used are up to date and that published vulnerabilities are known and monitored, such as those registered in CVE (Common Vulnerabilities and Exposures) or CWE (Common Weakness Enumeration). In addition, it is recommended to integrate automated vulnerability review and analysis processes into the system development cycle, using specialized tools such as:

Vulnerabilities	Security Controls
	• OWASP Dependency-check: An open-source tool that allows detecting known vulnerabilities in dependencies and libraries used in the system. • Dependency-Tracj: A solution that facilitates continuous component analysis and risk management associated with third-party dependencies. These tools allow you to proactively identify and alert on vulnerable dependencies, automating monitoring and facilitating the update or replacement of insecure components. This strengthens the security of the framework and significantly reduces the risk of exploitation.

Existence of rear doors in the model	To mitigate the risk of backdoors in AI models, regular testing and security analysis should be performed using specific techniques:
	• Periodic analysis of **SAST** (Static Application Security Testing) code. To **identify vulnerabilities** in the **libraries** and tools used that could be an attack vector for any of the vulnerabilities described in this guide. In addition, it will make it possible to **identify** a vulnerable **model** or an **outdated or compromised AI** framework.
	• Periodic analysis of **DAST** (Dynamic Application Security Testing) code. The approach of a DAST analysis focused on artificial intelligence must take special consideration of the scenarios of access to the model, control of input data and the risk of privilege escalation in security policies that allow manipulation or intervention in the input data.
	• Penetration **testing** oriented to Artificial Intelligence and Machine Learning. In this case, the tests must be aimed at carrying out adversary-type attacks: evasion, inversion and extraction. To facilitate the execution of these tests, the use of specific frameworks for AI security is recommended, such as:
	  ○ MLSecTools: A tool designed to perform penetration tests and evaluate the robustness of machine learning models against adversarial attacks.
	  ○ Adversarial Robustness Toolbox (ART): IBM's open-source framework that provides a set of tools for robustness testing and adversary attacks on AI models.

Vulnerabilities	Security Controls
	The integration of these tools into the development and assessment process ensures more effective detection of vulnerabilities related to backdoors and other attack vectors, improving the model's resilience to malicious manipulations.
Exposure and access to model code	Protecting the model code with access policies is critical to prevent mishandling or malicious tampering, and the security measures applicable in this area are the same as those that should be implemented in any computer system. These risks are not exclusive to AI systems, but they are especially relevant given the sensitivity and complexity of the models used. Access policies should restrict access to the code to authorized and qualified personnel only. This includes implementing robust identity and authentication controls, such as the use of MFA, certificates, etc. (multi-factor authentication), and tracking all interactions using detailed record (logs). The equipment of the staff involved in the development of the AI system must be adequately protected against exfiltration, for example by not allowing the use of removable media (USB, external drive, etc.) or access to public networks. In those scenarios working with the central repository, the communication channel must be adequately secured and encrypted (such as VPN, etc.) to avoid the possibility of manipulation of the model in transit through Man-In-The-Middle attacks. These measures ensure the integrity of the code and must be applied consistently to both traditional systems and those that implement artificial intelligence, ensuring the security of the entire environment in which the system operates.
Unknown vulnerabilities in relation to potential model defects	Continuous analysis of the cybersecurity indicators of the scorecard for the monitoring of the operation of the AI system. Analysis of the audits of the AI system carried out.

Table 3 Security Vulnerabilities and Controls for Intrinsic Defects

Example - Attendance at work

In the risks analysis carried out, it has been identified that the risk of being able to falsify an identification for the work attendance system, thus making it possible to generate attendance records that have not really occurred. Misclassification **and random misclassification vulnerabilities** have been considered intrinsic defects of the model and directly related to this risk.

An attacker could use adversarial patterns, for example, on a T-shirt, which could cause a misidentification of both another user and a failure to locate the user in their database (see for example https://arxiv.org/pdf/2211.07383.pdf).

As a security control when developing the model, to mitigate this vulnerability, the design team uses the Highly Confident Near Neighbour (HCNN) framework in combination with an Attribution-driven Causal Analysis.

In addition, the model framework used, and all associated ML libraries follow periodic **SAST**-like scans to locate vulnerabilities in the *perimeter of the AI system* that serve as a gateway.

In the production environment, it is periodically subjected to a **DAST**-type analysis focused on an escalation of privileges and access on the model. In this environment and also periodically, AI-oriented **penetration** tests are carried out, carrying out adversarial attacks on the system.

For defects that may appear in the AI system, the following vulnerabilities can be identified as a result of configuration, integration, or deployment:

- The artificial intelligence system allows private information to be extracted.
- Too much information provided in the model in the output information.
- Too much publicly available information about the model.
- The attacks to which the AI intelligence system may be exposed have not been considered.
- There is no security process that maintains the level of security of the AI system components.
- Use of vulnerable components.
- Mismatch between development/test/production environments.
- Indicators of the correct functioning of the system have not been defined, making it difficult to identify the compromise of the system.
- Bad practices due to lack of cybersecurity awareness.
- Contracts with third parties that do not have an adequate level of security.
- Unknown vulnerabilities as a result of configuration, integration, or deployment.

Vulnerabilities	Security Controls
Inadequate access rights management	Inadequate or uncontrolled access can lead to defects by unqualified personnel.
The artificial intelligence system allows private information to be extracted	The AI system has a differential privacy mechanism; the design of the model has been carried out using PATE (Private Aggregation of Teacher Ensembles). The techniques used to protect against investment attacks are equally applicable as security controls for this vulnerability.
Using Vulnerable Components	In this case, ISO27001/2 and NIST 800-53 recommend: • Have an inventory of the components used and constantly monitor them and their versions to be aware of the vulnerabilities found in them using the reference databases for them (CVE, https://cve.mitre.org/ or CWE, https://cwe.mitre.org/). Automatic tools, such as OWASP Dependency-check, can be used to check the components used and perform checks against reference databases. • Manage obsolete components and plan their removal from the production chain. **Without being specific to AI,** perform frequent vulnerability scans on all systems surrounding the AI system.

Vulnerabilities	Security Controls
Indicators of the correct functioning of the system have not been defined, making it difficult to identify the compromise of the system.	The AI system must have a series of indicators defined and associated with its nature and intended purpose. These indicators must be presented on the interface of the AI system, through a dashboard or accessible information panel, where the deployer can visualize and know in real time the status of the system's performance and detect the possible existence of failures or possible compromise situation. To fulfil their purpose, these indicators must include clearly defined thresholds that allow warning about anomalies in the behaviour of the system. An out-of-range threshold can be indicative of a failure, an attack attempt, or a performance degradation, facilitating early response and risk mitigation. The indicators must refer to at least the following metrics: • Metrics of accuracy and robustness of the AI system, to evaluate its performance in expected and adverse conditions. • Distinctive indicators of the training data databases used, such as data distribution or bias detection, to guide the correct use of the system and warn of possible deviations. The combination of indicators with defined thresholds ensures that the system can be proactively monitored, making it easier to identify problems in real-time and improving resilience in the face of unexpected situations.
Mismatch between development/test/production environments	Systems trained in simulation environments can make real-world errors due to mismatches between training and production environments, or simulation and production environments. Both environments should be defined by container descriptors when possible or identical hardware devices in the same environments (IoT, edge computing). In commercialization scenarios other than MLSaaS, the instructions must be clear (even with examples of the software containers if necessary) and the hardware specifications will be based on standards, in the case of IoT or in which edge computing is performed.
Contracts with third parties that do not have an adequate level of security	Occasionally, a third party may develop an element or section of the AI system. If the provider delegates the full implementation of the system to a third party, all security measures described in this guide should be considered and complied with.

Vulnerabilities	Security Controls
	In the case of partial or specific development of the AI system (such as models, data pre-processing or supporting infrastructure), it is essential to integrate security requirements into contracts with third parties. The recommendations of ISO 27001/2 or NIST 800-53 include integrating specific clauses that establish clear and verifiable security controls. Examples of these clauses are: • Periodic audit requirements: Mandate security audits of the provider's development and processes, with access to the results for the customer. • Security certifications: Require the third party to prove compliance with standards such as ISO 27001, ISO 27017 (for cloud security), or equivalent certifications. • Vulnerability review and management: Incorporate the obligation to perform periodic security tests (SAST/DAST), with the remediation of critical vulnerabilities in specific deadlines. • Confidentiality and protection of information: Include strict data protection and intellectual property policies in the development of the system. • Contingency and incident response plans: Establish the obligation to have an action plan in place in case of cyber incidents that may affect the AI system. In addition, it is recommended to continuously monitor and review compliance with these requirements through performance indicators and periodic reviews of the services provided.
Unknown vulnerabilities as a result of configuration, integration, or deployment.	Continuous analysis of the cybersecurity indicators of the scorecard for the monitoring of the operation of the AI system. Analysis of the audits of the AI system carried out. It is recommended to implement AI-based real-time monitoring tools that detect anomalous patterns in system behaviour. Additionally, periodic audits should include simulations of adversarial attacks to uncover potential undocumented defects

Table 4 Security vulnerabilities and controls for integration/configuration defects

Example - Aid granting automatic system

This system is deployed in the facilities of the public administration that makes use of it, especially in the Autonomous Communities and Local Entities. In this process, it has been identified that there may be a **mismatch between the development, testing and production environments**. These mismatches include differences in hardware configurations, versions of libraries or frameworks used, and system-specific dependencies.

To mitigate this risk, and facilitate the work of the systems administration and maintenance team in the infrastructure of the Autonomous Communities and Local Entities, the AI system is encapsulated as a set of services (data inputs, model, storage system, etc.) through a market system based on container description (e.g., Docker or Kubernetes), in such a way that, with the same description mechanism, the environment can be replicated on any infrastructure, with the same conditions and software artifacts.

Example - Attendance at work

In the construction of the AI system, a user interface has been included that allows the behaviour of the system to be visualized in an aggregated way and broken down by sections and departments. This dashboard uses advanced metrics such as Highly Confident Near Neighbour (HCNN) in combination with an Attribution-driven Causal Analysis to identify anomalous behaviours related to adversary control mechanisms.

In this panel, historical information of this relationship is displayed with data from the last days, week and month. The progress of these metric values is graphically displayed. In this panel, you can filter by sections of the facility and by department, offering the same data and visualization as the aggregate.

Thus, the system has a panel where general **anomalous behaviour can be detected, by demarcation and even by zone**. The panel has a configuration mechanism to allow you to define alarm and notification levels. This solution facilitates rapid detection and response to anomalies, helping to maintain the robustness of the system as required by Article 15 of the AI Act.

These vulnerabilities, especially those corresponding to the integration and configuration of the system, are on the **border** between cybersecurity for high-risk artificial intelligence systems and the cybersecurity of the technological ecosystem that accompanies it. Its inclusion has been considered, in order to have an adequate perimeter that allows continuity in cybersecurity (beyond the scope of this guide). Likewise, references have been provided that allow the organization to work to have a more complete perspective on cybersecurity in a continuous and comprehensive way.

If the model is to be used by the user on-premise (in the cloud or on their premises), the **installation and configuration instructions must be guided**, clear and step by step,

indicating the functionality and relationship of each parameter and the configuration errors. The process of changing parameters must ask for their confirmation and wait for confirmation from the previously identified user (by the mechanisms available for the nature of the system).

Deployer

In this section, the organisational measures for the deployer involve reading and understanding the manual prepared by the provider, in relation to both the intrinsic defects of the system and the configuration mechanisms that are applied to it by AI within the scope of the intended purpose for this one. You must know the documentation relating to the indicators of the correct functioning of the system.

If the format for the use and commercialization of the AI system contemplates that the deployer has the AI system as **an asset of its own facilities** (either on-premise, or in any infrastructure or facility managed by the deployer), it must apply the indications provided by the AI cybersecurity provider listed in the manual.

The deployer must have the technical means that are applicable to him so that the mechanisms (alerts, notifications) or interfaces (web, desktop, terminals) available to him are accessible to the personnel who must interact with the high-risk AI system.

5. Technical documentation

This cybersecurity guide is accompanied by a specific technical documentation that establishes how a high-risk AI system should be documented, in relation to the indications provided by the European Regulation on Artificial Intelligence. As the aforementioned guide serving as the reference document for the completion of the technical documentation, in this section we are going to indicate in aggregate form those points that must be taken into account when preparing the technical documentation on cybersecurity for AI of the high-risk system, and which have been mentioned throughout this guide. For a complete reference, please refer to the technical documentation guide.

Annex IV of the European Regulation on Artificial Intelligence establishes the minimum content that technical documentation must include. In the field of cybersecurity, 1.h) refers to instructions for use, which must detail the main cybersecurity risks and controls of the system in order to inform the person responsible for its deployment. Likewise, 2.h) stipulates that the cybersecurity measures applied must be documented.

Below, we indicate an initial reference on how to address these requirements.

Instruction Manual

The cybersecurity measures put in place for this AI system, which may be applicable to the deployer, especially those involved in the installation processes of the system in the event that they are carried out, to prevent the appearance of attacks that exploit configuration flaws (see section 4.5).

The instruction manual must also contain information on the risks of adversarial attacks that the system may suffer, as well as the security controls put in place to mitigate their effect (see section 4.4). Similarly, as explained in the use case examples, the interfaces present in the system, which show indicators of correct operation, must be detailed in the instructions, with clear indications of their interpretation easily understood by non-specialized personnel.

Parameters used to measure cybersecurity.

The system's technical documentation must contain the parameters used to measure, monitor and update the system's cybersecurity. These parameters must be aligned with risks mitigation strategies and linked to the identified vulnerabilities. At least the following must be documented:

- The established AI cybersecurity plan, including its scope and objectives.
- The assets and actors identified in the cybersecurity analysis.
- The vulnerabilities identified and the controls applied to the data for poisoning attacks.
- The vulnerabilities identified and the controls applied to protect against adversarial attacks.

The vulnerabilities and controls associated with the model's defects, including the prevention and detection measures implemented.

In addition, this documentation must comply with the procedures and indications provided in this guide, specifying the techniques used to assess and mitigate risks, their direct relationship to the risks mitigated and vulnerabilities addressed, and a clear approach for their maintenance and updating in the system lifecycle.

6. Self-assessment questionnaire

To carry out a self-assessment of compliance with the requirements of the European Regulation on Artificial Intelligence referred to in this guide, a global self-assessment questionnaire has been generated with a series of questions with the key points to be taken into account with respect to the obligations dictated by the articles of the AI Act mentioned in this guide.

It will be necessary to refer to this document in order to carry out the section of the self-assessment questionnaire corresponding to this guide.

7. Annexes

7.1 Annex I: Access policies. An approach to high-risk artificial intelligence systems

We can say that the first and real line of defence with respect to many of the specific AI measures is based on properly controlling access to all assets during the life cycle.

Access policies, even if they are not a specific AI issue, are close enough to all assets and processes throughout the lifecycle to dedicate an approach that allows the organization to trace a continuity that ensures protection.

7.1.1 For the provider

Access policies, understood as the rules of access to assets, as well as the rules of action on them are very important when applying specific cybersecurity measures to high-risk AI systems. These policies set out **what** can and can be done and who is authorized to do so with the AI system's assets and lifecycle processes. Protected assets are crucial in the cybersecurity of the system. Leaving them exposed is an unacceptable risk for any AI system.

It is recommended that access policies be based on the RBAC (Role-Based Access System) paradigm, which allows assigning roles to those responsible for the deployment and actions allowed to these roles.

The approach to access policies should be **least privilege,** and a robust authentication mechanism, password rotation, and/or certificate-based access should be established depending on the system. It is recommended that a **multi-factor authentication** system be included, which prevents spoofing, by a compromise of a password. Access **revocation** policies must be part of the company's exit mechanisms to prevent access for people who are no longer in the organization.

It is recommended that it be compatible with market identity management solutions, widespread in organizations, such as Active Directory or equivalent. This integration will be **documented** in the **instruction manual** to facilitate the process for users of the system.

Who does it apply to within the organization?

The AI system access policies apply to all personnel who interact with the system at any level since its development, implementation, and commercialization. It is very important to convey to each of the groups of actors indicated the minimum possible privileges.

These groups may be:

- Data Actors: Profiles that act on the data that feeds the AI system's learning (Data Analysts, Data Scientists, AI Experts, etc.)
- Actors on the configuration and operation of the AI system (AI Experts, AI system administrators, System Analysts, ML Experts, etc.)

- Developers of integration and/or interaction with the AI system (including MLOps, if applicable).
- System administration actors that host and operate the AI system (AI system administrators, system analysts, etc.)
- Actors on the design, implementation and maintenance of the software that materializes the AI system itself (including, where appropriate, the use of firmware or similar) (Developers, Analysts, AI Experts, etc.).
- Actors on the architecture, administration and maintenance of the AI system in its life cycle, both during development and once it is launched (AI Experts, AI system administrators, ML experts, etc.).
- The general staff of the organization, applying the policy of least privileges, should not have access to the AI system, the development of the model or the data. In those cases, outside the actors identified in the previous points, access must be analysed and the need must be recorded, if it is temporary, it must be revoked appropriately.

See also ISO/IEC 22989:2022 Information technology – Artificial Intelligence – Artificial intelligence concepts and terminology.

What aspects should it cover?

The access policies established for any of the assets described in this section must be reviewed periodically, with the following points being updated:

- People and roles assigned, updating and changes.
- Defined roles and actions.
- Permit revocation and removal process.
- Permissions without assigning to a user.
- Access recording and usage monitoring system.

Actions on **training data** must be protected by security policies, considering at least:

- Incorporation of new data.
- Deletion of existing data.
- Editing existing data.
- Use of data.

The segmentation of actions on data is essential to have control and protection over the risks **of data poisoning**.

The **model** as a fundamental part of the artificial intelligence system must be protected by security policies. At least the following actions must be controlled:

- Access (display) to the model code.
- Editing the model code.

- Use of model for training.
- Training process.
- Validation and deployment (release) of the model to production.
- Access, editing and deletion of any technical and design documentation relating to the model and descriptive of the nature of the system and its components.

The segmentation of actions on the model itself allows it to be protected from inadvertent manipulation of it, both intentional carried out by malicious attackers, and erroneous by unqualified people (without the established permission) introducing errors in the model. It prevents exfiltration of this for transfer attacks and controls the training process for protection against **adversarial attacks.**

7.1.2 For the deployer

Deployers must also have established and defined access policies for the AI system. Access to high-risk Artificial Intelligence systems by the personnel of the deployer should never be anonymous and without controlled access.

In addition, and in accordance with the commercialization model of the artificial intelligence system, it will be necessary to establish other access policies.

The approach to access policies should be least privilege, and a robust password rotation mechanism should be in place.

Who does it apply to within the organization?

The staff of the user organisation who have access to the system and its interaction. If, due to the marketing model, it is this organization that manages the system (whether in-cloud, on-premise or any variant), all related personnel must be included in the access policies.

The accesses, their activity and the authorized persons must be registered and audited periodically, especially the revocations of permits, which must take effect immediately.

The AI system shall be integrated, in those marketing scenarios, where necessary, with the identity mechanism of the user organisation for access to the AI system, following the indications and recommendations of the **instruction manual** provided by the provider.

What aspects should it cover?

The main asset protected by security policies in the case of the deployer is **the AI system itself**. Access to the system, its control panel, interface or equivalent, must be protected following the provider's instructions, and integrated with the identification mechanisms of the deployer's organization in those cases in which it applies.

It is **the responsibility of the deployer to** manage this access and have appropriate policies in place.

7.2 Annex II: Training in cybersecurity and high-risk artificial intelligence systems

As mentioned in the introduction, the training makes it possible to strengthen the security chain that protects the high-risk artificial intelligence system. Again, this annex covers an aspect that is not central to AI cybersecurity, but that is located in a border region and that **needs** to be covered.

7.2.1 For the provider

Cybersecurity for AI applies to the provider throughout the life cycle of the AI system and throughout all assets (data, model, etc.) of that system. For cybersecurity policies to be effective, they must be communicated to all the personnel involved. Cybersecurity training for AI systems has a two-fold objective depending on the scope of cybersecurity in the provider's organization:

- If the provider **already has** a cybersecurity process, this will have training as one of its pillars. This training **should be expanded** to cover the topics described here and to reach the people considered.
- If the provider **does not have** a specifically designed cybersecurity process, it **must organize** training plans that cover the aspects described in this section.

The provider-deployer training must be clearly defined. In those models of marketing and/or putting into service of the artificial intelligence system that require the participation of the staff of the deployer, beyond the use of the system itself, the basic aspects to be covered by the cybersecurity training plan must be described in the instruction manual.

Who does it apply to within the organization?

The recipients of cybersecurity training may be:

- Data Actors: Profiles that act on the data that feeds the AI system's learning (Data Analysts, Data Scientists, AI Experts, etc.)
- Actors on the configuration and operation of the AI system (AI Experts, AI system administrators, System Analysts, ML Experts, etc.)
- Developers of integration and/or interaction with the AI system (including MLOps, if applicable).
- System administration actors that host and operate the AI system (AI system administrators, system analysts, etc.)
- Actors on the design, implementation and maintenance of the software that materializes the AI system itself (including, where appropriate, the use of firmware or similar) (Developers, Analysts, AI Experts, etc.).
- Actors on the architecture, administration and maintenance of the AI system in its life cycle, both during development and once it is launched (AI Experts, AI system administrators, ML experts, etc.).

- The general staff of the organization, applying the policy of least privileges, should not have access to the AI system, the development of the model or the data. In those cases, outside the actors identified in the previous points, access must be analysed, and the need must be recorded, if it is temporary, it must be revoked appropriately.

See also ISO/IEC 22989:2022 Information technology – Artificial intelligence – Artificial intelligence concepts and terminology.

What aspects should it cover?

The aspects described here must be treated in different depths, depending on the personnel receiving them and their link with the artificial intelligence system. This scope may range from awareness for all staff to greater complexity for those directly related to the AI system.

This training must cover, at least, the following topics of cybersecurity applied to artificial intelligence systems:

- Why is cybersecurity important in AI? Cybersecurity during the AI system lifecycle.
- The cybersecurity of training data: risks of poisoning of training data.
- Protection of the AI system. Risks of information and code leakage of the AI system or tampering.
- What are adversarial attacks? Definition. Different types. Adversarial attacks for the Artificial Intelligence system under consideration.
- Exploitation of model defects. Definition of transfer attacks. Need to use secure models/frameworks. What are intrinsic model defects? Secure configuration and integration. Avoid errors in the model due to the use of bad practices.
- The importance of role-based access control (RBAC) for access to both data and the authentication system model and robustness, including reinforcements such as two-factor authentication.

The training must be updated in relation to the changes in the cybersecurity model and paradigm in artificial intelligence, which is a changing and continuously evolving area.

It may require that the receipt of training by certain profiles (data scientists, ML/AI experts) requires the performance of tests and/or certifications.

For the deployer of the AI system, an **indication of the training material** must be developed, explaining the level of cybersecurity required for the use of the Artificial Intelligence system: FAQs, explanatory documentation, step-by-step descriptions of the operation and critical points. All of this adapted to the level of commercialization of the AI system (SaaS, On-premise, in-cloud, etc.)

7.2.2 For the deployer

The cybersecurity training applied to the deployer will depend on the approach of the commercialization model of the artificial intelligence system considered. In those scenarios in which the system is installed on the deployer's infrastructure (on-premise or in-house models), the deployer must adapt their cybersecurity training plans to cover the aspects indicated in the system's instruction manual.

Who does it apply to within the organization?

The staff of the organisation using the artificial intelligence system, who will carry out the direct interaction with it. Depending on the marketing approach:

- Systems administration staff.
- Staff responsible for the configuration of the AI system.
- Staff responsible for monitoring the AI system.

What aspects should it cover?

The deployer must use the material generated by the provider in terms of cybersecurity applied to its AI system to guide and cover the necessary knowledge and/or profiles.

Regardless of the marketing format in which the system is received, all its personnel in direct interaction with the AI system must receive training in awareness, vulnerabilities of AI systems, poisoning, adversarial attacks...

Personnel who interact with and are tasked with working with the AI system should be trained specifically in adversarial attacks for the AI system and its intended purpose.

According to the marketing formats, if the configuration or administration of the AI system is carried out by the staff of the deployer (in-house system administration, for example), the training must include:

- Exploitation of model defects. Definition of transfer attacks. What are intrinsic model defects? Secure configuration and integration.

The training must be periodic and updated according to the changes and updates provided by the provider.

7.3 Glossary

Term	Definition
Adversary Attack	Attacks that are carried out against an artificial intelligence system are considered adversarial attacks. In relation to the level of knowledge that the attacker of the adversary attacks can be: - White Box: Access to the model architecture, training data, parameters and hyper parameters. - Black Box: inputs and outputs. - Gray Box: a mixture of both. In relation to the mode of operation, attacks can be (see in detail the definitions in the glossary): - Poisoning. - Evasion, which are sometimes considered directly in a generic way to adversarial attacks, although it is really a specific type of them. - Extraction. - Inversion.
Evasion Attack	A type of attack that occurs during the inference phase with the AI system in production. The attacker does not have access to the training data, he only communicates with the system via interface (physical, visual or API-type). In general, this type of attack seeks to make a prediction of the model wrong.
Attack Extraction	Type of attack during the inference phase with the AI system in production. Without access to the training data, it only communicates with the system via interface (physical, visual or API-type). The attacker provides inputs to the model by accessing it, and records the outputs, to infer a total or partial replica of the model through their responses.
Attack Investment	It is a type of attack that occurs during the inference phase with the AI system in production. In these attacks, the user does not have direct access to the data, but they can access the model through its interface (whether physical, visual or API). Investment attacks aim to gain insight from model data. As indicated in the guide, they can be: - Membership Inference Attack to find out if a sample belongs to the training data. - Property Inference Attack, which seeks to know the data used by the model during training to make predictions. - There is a third type that is reconstruction, which is more complicated to execute, because its intention is to reconstruct all the input data.
Attribution-Driven Causal Analysis	This concept is related to the fact that there is a connection between resistance to adversarial perturbations and the attribution-based explanation of individual decisions generated by machine learning models. Adversarial inputs are not robust in the attribution space, i.e., masking a few traits with high attribution leads to changing the

Term	Definition
	indecision of the machine learning model in adversarial examples. Instead, natural inputs are robust in the attribution space. The concept has been developed in an article, which can be found here https://arxiv.org/abs/1903.05821
Bagging	Use of models trained in parallel to constitute the artificial intelligence system. The goal is to take advantage of the independence between separately trained models on different subsets of the original set of training data (for the same intended purpose for the system) to average their output. This type of technique is a protection mechanism against data poisoning. It is also an efficient technique against evasion attacks, as it does not depend on a single model that may be vulnerable to the attack under consideration. For classifications, the different trained models perform a *vote* on class membership, for those operations that do not perform a classification (regressions, estimates, etc.) an arithmetic mean is performed. It is also sometimes listed as *Bootstrap Aggregation*.
Boosting	Reinforcement is a set technique that seeks to change the training data and adjust the weight of the observations based on the previous classification. Unlike the bagging approach, boosting involves reliance on weak models. Weak learners take into account the results of the previous weak model and adjust the weights of the data points, making it a strong model. Boosting changes the weight associated with an observation that was incorrectly classified by trying to increase the weight associated with it. Reinforcement tends to decrease bias error but can sometimes lead to an overfitting of the training data set.
CAT-Gen	When developing NLP models, we will be able to perform different tasks automatically, such as translating texts, summarizing them, etc. The problem with this type of model is that they are very unrobust against small modifications in the input data. However, in the use of this technique he shows how by introducing adversarial learning, in this type of models, the robustness against changes in the input data improves significantly This method can generate more diverse and fluid adversarial texts. The use of these adversarial texts allows the models to be improved through adversary training. The development of the technique is detailed by the authors can be found in https://arxiv.org/abs/2010.02338
CVE	Common Vulnerabilities and Exposures (CVEs) is a list of publicly disclosed vulnerabilities and information security exposures. A list is available on https://www.cve.org/

Term	Definition
CWE	Common Weakness Enumeration (CWE) is a system of categories for software weaknesses and vulnerabilities. It can be consulted in https://cwe.mitre.org/
Defensive Distillation	Defensive distillation is an adversarial training technique that adds flexibility to an algorithm's classification process so that the model is less susceptible to exploitation. In distillation training, one model is trained to predict the output probabilities of another model that was trained on a previous reference standard to emphasize accuracy. The first model is trained with "hard" labels to achieve maximum accuracy, for example, by demanding a probability threshold of 100%. If an attacker learns what features and parameters the system is looking for, they can send input data with only the relevant features, resulting in a false positive match. The first model then provides "soft" labels with a probability of 95%. This uncertainty is used to train the second model to act as an additional filter. Since there is now an element of randomness to achieve a perfect match, the second algorithm or "distilled" is much more robust and can more easily detect impersonation attempts.
DLP (tools)	They are tools meant to prevent data loss. Its relationship with cybersecurity for artificial intelligence is dictated by the need for control, stability, and tracking of training data sets. In this way, the control established by these tools allows you to add security control over data poisoning attacks. There are several market tools and open-source solutions.
Poisoning	Poisoning is a type of attack, in the design and training phase, that tries to manipulate the set of training data, with the aim of controlling the predictions of the artificial intelligence system in such a way that the model transforms, in the inference phase once deployed in production, malicious inputs into correct results.
Feature Denoising	This concept is based on the fact that adversarial perturbations made in images cause noise in the features built by neural networks. It is an architecture that increases resistance to adversarial attacks by eliminating noise in the features. This architecture has been proposed in an article that can be found in https://arxiv.org/abs/1812.03411
Feature Squeezing	This technique can be used to protect models by detecting adverse examples. Feature *squeezing* decreases the search space available to an adversary by merging samples that correspond to many different feature vectors in the original space into a single sample. By comparing the prediction of a DNN model on the original input with that of the compressed input, this technique detects the adversary examples with high accuracy and few false positives. If the original and compressed examples produce substantially different results from the model, the

Term	Definition
	input is likely to be adverse. The authors of this technique have published their results in https://evademl.org/squeezing/
GPAI (general purpose AI)	According to the European Regulation on Artificial Intelligence, in **Article 3 Definitions (1b):** *"an AI system that – regardless of how it is marketed or put into service, including as open-source software – is intended by the provider to perform general-purpose functions, such as image and speech recognition, audio and video generation, pattern detection, question answering, translation, and others; a general-purpose AI system can be used in a plurality of contexts and integrated into a plurality of other AI systems."* Given that the use of this type of system in the development of an artificial intelligence system may be possible, it should be clarified here, and an extract from the following recital (12c) should be added to the above definition: *"[...] Providers of general-purpose AI systems, regardless of whether they can be used as high-risk AI systems as such by other providers or as components of high-risk AI systems, should cooperate, as appropriate, with providers of the respective high-risk AI systems to enable their compliance with the relevant obligations under this Regulation and with the competent authorities established under this Regulation.';*
Highly Confident Near Neighbour (HCNN)	It is a framework for working in neural networks that combines trusted information and the search for the nearest neighbour, to reinforce the adversarial robustness of a base model. This can help distinguish between correct and incorrect model predictions in a vicinity of a sampled point of the training distribution used. The technique is developed in an article that can be accessed at the following link https://arxiv.org/abs/1711.08001
Input gradient regularisation	It consists of imposing restrictions on the entry gradient, or regularization. The regularization technique in general consists of restricting the estimates of the coefficients to zero. This technique allows to avoid overfitting *the* training set and is fundamentally a technique used to provide resistance to opponent attacks, or to mitigate them. Regularization in the input gradient prevents the input gradient from being too large, making neural networks vulnerable to adversarial attacks.
Isolation Forest	This method is based on the decision tree algorithm. It isolates outliers by randomly selecting a feature from the given feature set, and then randomly selecting a split value between the maximum and minimum values of that feature. This random partitioning of features will produce shorter paths in the trees for anomalous data points, thus distinguishing

Term	Definition
	them from the rest of the data. This algorithm is based on the principle that anomalies are observations that are few and different, which should make them easier to identify. Isolation Forest uses a set of isolation trees for the given data points to isolate anomalies. In this way, both the outliers present in the dataset, and those that may have been introduced to generate poisoning, can be detected and the impact mitigated
Label Flipping	This is a type of attack that, for data captured from unsecured sources, and where it is not possible to obtain the data in any other way, there is a risk that an attacker may have altered the labelling of the data in a part of the training data set. Applied to high-risk systems, it is important to consider that this concept should be taken into account, even as it occurred *naturally or not maliciously*, but that in any case it could affect the performance of the system. In that scenario, techniques to detect data relabelling should be applied to mitigate this effect. As described in https://www.researchgate.net/publication/323550082_Label_Sanitization_against_Label_Flipping_Poisoning_Attacks, the procedure uses k-NN to assign the tag to each instance of the training set. The aim is to reinforce the homogeneity of the label between the instances that are nearby, especially in the regions that are far from the decision limit
Local Outlier Factor	It is a method that allows both natural outliers, i.e. those that may be naturally present in a sample (given measurement, storage errors, etc.) and those that may have been inserted as an intention of an envenomation attack. This is an unsupervised anomaly detection method that calculates the deviation of the local density of a given data point from its neighbours. It considers samples that have a substantially lower density than their neighbours to be outliers.
MITRE ATLAS (tools/fonts)	It is a state-of-the-art source of information on adversarial attack techniques, based on systems that are in production. It is a very relevant source of information, which must be consulted both in the design and development phases. During the inference phase, the system may require updates in relation to use cases found in other systems that allow transfer attacks. All information is accessible through the address https://atlas.mitre.org/
Model stacking	Stacking is a joint machine learning algorithm that best combines the predictions of multiple well-performing machine learning models. The advantage of this technique is that it can take advantage of the capabilities of a number of well-performing models in a classification or regression task and make predictions that perform better than any single model in the set. The following link provides an example of how to apply the technique

Term	Definition
	using Python https://machinelearningmastery.com/stacking-ensemble-for-deep-learning-neural-networks/
Neural Dropout	Dropout is a regularization method that approximates the training of a large number of neural networks with different architectures in parallel. During training, a certain number of layer outputs are randomly ignored or "discarded". This has the effect of making the layer look and be treated as a layer with a different number of nodes and connectivity to the previous layer. In effect, each update of a layer during training is performed with a different "view" from the configured layer. This technique is implemented on a per-layer basis in a neural network. It can be used with most types of layers, such as fully connected dense layers, convolutional layers, and recurring layers, such as the short-term memory network layer. (see https://arxiv.org/pdf/1806.01246v2.pdf)
OoD Data (Out of Domain Data)	This is input data, received by the inference system once the AI system is put into production, that is not within the established training domain and related to the intended purpose of the system. This is an indicator of possible adversarial attack and belongs to the techniques of control of input parameters. In the case of structured and tabulated data, or numerical data, discrete series, etc., it is relatively easy to detect those values that are not within the domain and restrict or prohibit their entry into the artificial intelligence system. When it comes to systems based on natural language processing, it is especially important to detect those elements that are not part of the domain (see https://arxiv.org/pdf/1909.03862.pdf).
Differential Privacy	Differential privacy seeks to protect the values of personal data by incorporating statistical "noise" into the analysis process. The calculations required to form noise are complex, but the principle is quite intuitive: noise ensures that data aggregations remain statistically consistent with actual data values, allowing for random variation, but making it difficult to work with individual values in aggregated data. In addition, the noise is different for each analysis, so the results are non-deterministic; In other words, two analyses that perform the same aggregation may generate slightly different results. https://becominghuman.ai/what-is-differential-privacy-1fd7bf507049 https://learn.microsoft.com/es-es/training/modules/explore-differential-privacy/2-understand-differential-privacy

Term	Definition
RONI	The premise of this type of mitigation in attacks focuses on simulating the effect of including training input data, **before it** is incorporated into the process. To do this, the data is rejected in the event that its impact on the model is negative (Reject On Negative Impact) RONI. This type of defence is especially interesting if the system has a learning phase parallel to the inference with the system in production, or if the source of training data is very extensive and difficult to control its state (for example, email texts) and therefore prevent them from being candidates to poison the model. It sets a threshold by looking at the average negative impact of each instance on the training set and flags an instance when its performance impact exceeds the threshold. A job detailing the process can be found in https://people.eecs.berkeley.edu/~tygar/papers/SML/misleading.learners.pdf
STRIP	The STRIP technique is intended to be able to mitigate the data poisoning attacks they generate or allow the creation of a *backdoor* in a deep neural network. Generally, these attacks have a *trigger* in the data that allows the response of the model with which it has been poisoned to be activated and generate an erroneous classification. STRIP is short for **STRong Intentional Perturbation,** this technique for detecting these responses and is focused on artificial vision systems. The technique has been developed in the following paper that can be found here https://arxiv.org/abs/1902.06531
Transfer of Adversary Attacks	This vulnerability represents the exposure of the artificial intelligence system to receive adversarial attacks for other systems, given the similarity it may have with these. This type of attack can be performed once the attacked system has been identified to a greater or lesser extent and known attacks for the model type or family can be used.
TRIM	It is a technique that, especially for the case of regressions. The TRIM method estimates regression parameters iteratively, using a trimmed loss function to eliminate points with large residuals. After a few iterations, TRIM is able to isolate most poison points and learn a robust regression model. Greater detail can be found in https://www.researchgate.net/figure/Several-iterations-of-the-TRIM-algorithm-Initial-poisoned-data-is-in-blue-in-top-left_fig2_324166859
TRONI	If the attack we are exposed to is targeted, poisoned data instances may not cause a significant drop in performance. These variant leverages prior knowledge about a test-time misclassification to determine which instances of training might have caused it. While RONI estimates the negative impact of an instance as a whole, tRONI considers its effect exclusively on the classification of the target. The applicable procedures

Term	Definition
	for performing this technique can be found in https://arxiv.org/pdf/1803.06975.pdf
Z-score	It is a data normalization technique, or applied in the terminology of Artificial Intelligence, it is a scaling of characteristics. In this technique, the values are normalized based on the mean and standard deviation of the data. The essence of this technique is the transformation of the data by converting the values to a common scale in which the mean is equal to zero and the standard deviation is one.

8. References, Standards & Norms

Various sources consulted have been used to prepare the cybersecurity guide. It is especially recommended that providers and deployers read two reports prepared by the European Union Agency for Cybersecurity (ENISA).

This European agency prepares relevant documents in the field of cybersecurity and will undoubtedly be a leading player in the aspects of cybersecurity in AI related to the European Regulation on Artificial Intelligence.

8.1 Standards

The following standards have been consulted to reflect in some cases the *intermediate* aspects between generic and AI-specific security aspects. The points of the guide have indicated those recommendations that could be applicable.

- ISO/IEC 27001 Information technology – Security techniques – Information security management systems – Requirements
- NIST 800-53 Security and Privacy Controls for Information Systems and Organizations
- NISTIR 8269, A Taxonomy and Terminology of Adversarial Machine Learning.

8.2 ENISA

8.2.1 Artificial Intelligence cybersecurity challenges

Published in December 2020, this paper maps the AI cybersecurity ecosystem and its threat environment. It has been carried out by the working group on cybersecurity for artificial intelligence.

It can be consulted in full at:

- https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges

8.2.2 Securing Machine Learning

Published in December 2021, this document has been carried out by systematically reviewing the existing literature on machine learning. The document makes a special presentation of a threat analysis for machine learning systems, providing, as indicated in this guide, a relationship to security controls.

It can be consulted in full at:

- https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms

8.2.3 Cybersecurity of AI and standardization

This document, published in March 2023, presents the standards (existing, under development and planned) related to cybersecurity in artificial intelligence. It presents its scope, and in its conclusions offers those possible gaps present in the planned standardization. Beyond the AI sandbox framework, tracking these types of cybersecurity standardization efforts for AI is very important for the provider and deployer of high-risk AI systems.

It can be consulted in full at:

- https://www.enisa.europa.eu/publications/cybersecurity-of-ai-and-standardisation

8.3 Other references

- https://www.querypie.com/blog/machine-learning-data-poisoning-and-how-to-prevent-it/
- https://arxiv.org/abs/2009.07008
- https://elie.net/blog/ai/attacks-against-machine-learning-an-overview/#toc-10
- https://arxiv.org/pdf/1706.03691.pdf

Bagging or weight bagging

- https://www.simplilearn.com/tutorials/machine-learning-tutorial/bagging-in-machine-learning
- https://www.projectpro.io/article/bagging-vs-boosting-in-machine-learning/579

Trim Algorithm

- https://www.researchgate.net/figure/Several-iterations-of-the-TRIM-algorithm-Initial-poisoned-data-is-in-blue-in-top-left_fig2_324166859
- https://secml.github.io/class11/

Data augmentation

- https://arxiv.org/pdf/2010.01862.pdf
- https://research.aimultiple.com/data-augmentation-techniques/
- https://research.aimultiple.com/data-augmentation/

Data sanitization

- https://towardsdatascience.com/5-outlier-detection-methods-that-every-data-enthusiast-must-know-f917bf439210

- https://towardsdatascience.com/poisoning-attacks-on-machine-learning-1ff247c254db

STRIP

- https://arxiv.org/abs/1902.06531

IMPROPER LABELING

- https://www.researchgate.net/publication/323550082_Label_Sanitization_against_Label_Flipping_Poisoning_Attacks

Adversarial Examples:

- Adversarial attacks on medical machine learning: Adversarial attacks on medical machine learning – MIT Media Lab

Oracle vulnerability protection

- Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures https://dl.acm.org/doi/pdf/10.1145/2810103.2813677
- Membership Inference Attacks against Machine Learning Models https://arxiv.org/pdf/1610.05820.pdf.
- Secure, Robust and transparent application of AI https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Secure_robust_and_transparent_application_of_AI.pdf
- A Taxonomy and Terminology of Adversarial Machine Learning https://csrc.nist.gov/publications/detail/nistir/8269/draft
- SoK: Security and privacy in machine learning https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8406613

Protection of vulnerabilities and evasion

- Adversarial attack and defence in reinforcement learning-from AI security view https://cybersecurity.springeropen.com/track/pdf/10.1186/s42400-019-0027-x.pdf
- On detecting Adversarial perturbation https://arxiv.org/pdf/1702.04267.pdf
- More resilient model:
- Algorithm stability on adversarial training https://proceedings.neurips.cc/paper/2021/file/df1f1d20ee86704251795841e6a9405a-Paper.pdf
- Input gradient regularisation https://towardsdatascience.com/the-many-uses-of-input-gradient-regularization-e2af244e6950
- Defensive distillation: https://deepai.org/machine-learning-glossary-and-terms/defensive-distillation
- Generating Adversarial Examples with Adversarial Networks: https://arxiv.org/pdf/1801.02610.pdf

- Adversarial Attacks and Defenses in Deep Learning, https://www.sciencedirect.com/science/article/pii/S209580991930503X
- Distillation as a defense to adversarial perturbations against DNN https://arxiv.org/pdf/1511.04508.pdf
- Towards Deep Learning Models Resistant to Adversarial Attacks https://arxiv.org/pdf/1706.06083.pdf
- Improving the Adversarial Robustness and Interpretability of Deep Neural Network by Regularizing their input Gradients https://arxiv.org/pdf/1711.09404.pdf

Differential privacy

- PATE model (paper https://arxiv.org/abs/1610.05755, explanation http://www.cleverhans.io/privacy/2018/04/29/privacy-and-machine-learning.html)
- PATE Model (https://blog.openmined.org/build-pate-differential-privacy-in-pytorch/)
- Security and privacy in ML https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8406613
- Scalable Private Learning with Pate https://arxiv.org/pdf/1802.08908.pdf
- Membership inference attacks
- Article referenced https://arxiv.org/pdf/1806.01246v2.pdf.
- Also https://www.cs.cmu.edu/~mfredrik/papers/fjr2015ccs.pdf.
- Using neural dropout and model stacking helps mitigate risk.
- Neural Dropout https://towardsdatascience.com/dropout-in-neural-networks-47a162d621d9
- Model stacking https://towardsdatascience.com/simple-model-stacking-explained-and-automated-1b54e4357916

CAT-Gen

- https://arxiv.org/abs/2010.02338
- Reinforcing Adversarial Robustness using Model Confidence Induced by Adversarial Training - Xi Wu, Uyeong Jang, Jiefeng Chen, Lingjiao Chen, Somesh Jha
- Attribution-driven Causal Analysis for Detection of Adversarial Examples, Susmit Jha, Sunny Raj, Steven Fernandes, Sumit Kumar Jha, Somesh Jha, Gunjan Verma, Brian Jalaian, Ananthram Swami

Financiado por
la Unión Europea
NextGenerationEU

GOBIERNO DE ESPAÑA
MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

Plan de
Recuperación,
Transformación
y Resiliencia

España | digital 2026