5

Companies developing compliance with requirements

Guide 5. Risk Management

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 relevant technical references currently under development and applicable, particular note should be made of **prEN 18228, *"Artificial Intelligence – Risk Management,"*** and **ISO/IEC 23894, *"Information technology – Artificial intelligence – Guidance on risk management,"*** which together will serve as the basis for risk management and risk assessment throughout the life cycle of artificial intelligence systems, aligned with compliance with the European Regulation on Artificial Intelligence

Revision date: 10, December 2025

General content

Detailed Index

1. Preamble

1.1 Purpose of the document

This guide presents the organisational and technical measures that will help providers and those responsible for deployment to comply with the article "Risk management system" of **the European Regulation on Artificial Intelligence (*AI Act*).**

This article regulates the risk management system that must be incorporated by all high-risk AI systems (HRAIS) and certain general-purpose AI systems (article "Requirements for general-purpose AI systems and obligations for providers of these systems").

In this sense, throughout the guide we will generally refer to these systems as "AI system" with the aim of simplifying the discourse.

1.2 How to read this guide?

If the reader does not know how to develop a risk management system:

It is recommended that the reader first read the guide with the help of the examples incorporated, which will help them gain an understanding of the aspects that must be covered for the implementation of a risk management system in the context of AI.

Next, a second reading is recommended for the reader, following the detail provided in the Excel mentioned in section 1.4, which includes the process of developing a risk management system for 2 use cases.

If the reader knows how to develop a risk management system:

It is recommended that the reader complete at least one full reading of the guide. Although the concepts in section 2 and section 4, which describe the main elements for the development of a risk management system, may be familiar to them, it is nevertheless recommended that special attention be paid to:

- The examples presented throughout these sections that aim to apply these practices to the AI environment.
- The catalogues referenced in these sections (present in the annexes):
    - Annex A: Most relevant elements of the internal and external context around AI
    - Annex B: Most common components of AI systems (every risk management system revolves around the identification and analysis of risk and these risk are inherent to the components of the system under analysis).
    - Annex C: Common types of risk in the field of AI (will allow the reader to identify the new risk inherent to AI).
    - Annex D: Examples of controls in the field of AI.
    - Annex E: Effectiveness indicators.

- Section 5 is developed to cover additional aspects of the article's requirements.

1.3 Who is it for?

The requirements described in the article "Risk management system" must be fulfilled by the person in charge of developing the system, i.e. the provider.

This article does not specify requirements for the system deployer. If the deployer is involved in the development of the system, he or she must implement the measures developed for the provider. Nevertheless, the deployer is expected to always make responsible and ethical use of the system.

In addition, implementing the requirements of the article "Risk management system" is not among the obligations defined in the article "Obligations of deployers of high-risk AI systems".

The measures detailed throughout this guide are serve as guides for the provider. They are both organisational and technical in nature. Although in view of the nature of the article "Risk management system", most of the measures are of an organizational nature.

1.4 Use cases and examples throughout the guide

To **facilitate the understanding of the guide**, an Excel document is provided alongside the process of developing a **risk management system** for **different use cases**.

To this end, the steps described in section 4 of this guide have been followed, which outline the elements to be implemented for the development of an adequate risk management system.

These examples are developed based on the **use cases described** in the **Cross-Cutting Information and Concepts Guide**.

The Excel document is supplemented with various explanations, to link each phase of the development of the risk management system with the elements described in the guide.

Additionally, to facilitate reading the guide, small examples have been incorporated between the different sections of it.

Finally, it should be clarified that these examples are merely illustrative. The provider and the deployer should consider implementing all measures outlined in this guide as appropriate.

In addition, the examples presented are specific to the use cases. This implies that the proposals are specific to the models considered as examples, and not a general solution for other types of models, or even models of the same type.

Each organization must, in accordance with this guide, establish the appropriate measures for its type of AI system and its intended purpose.

2. Introduction

2.1 What is a risk management system and what are the main elements?

A risk management system is a management system whose objective is the identification and analysis of risk and the implementation of mitigating measures.

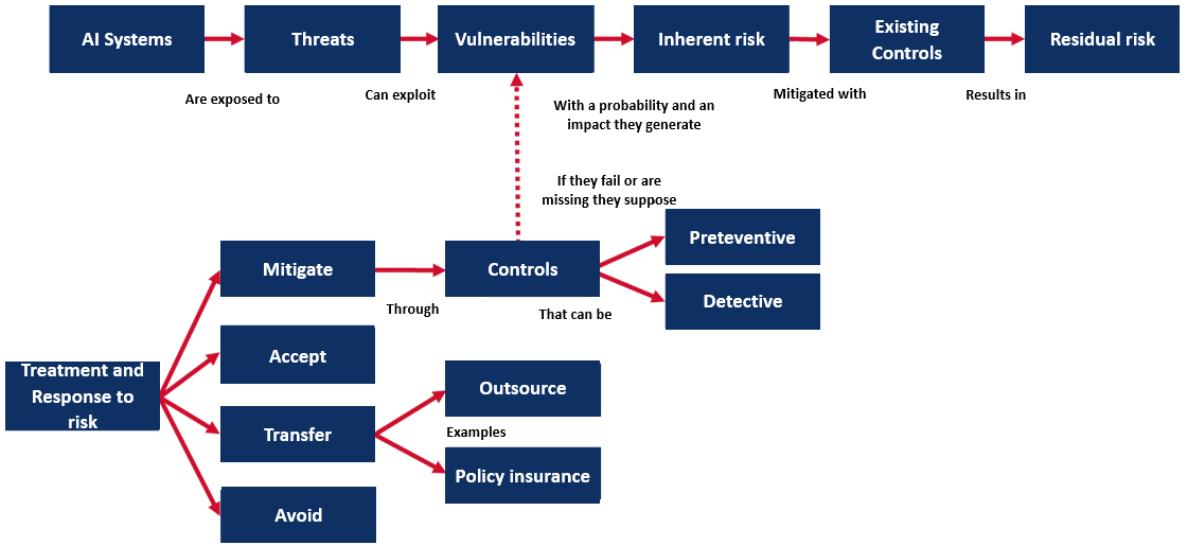
By risk, we mean any event with a probability of occurring and an impact in case it happens. Risk, as we will see later in detail, is calculated as the product of risk factors, their probability and impact.

In the context of the European Regulation on Artificial Intelligence, the risk management system to be developed should pay particular attention to **risk that may affect the health, safety and fundamental rights**.

The focus should be on the identification and analysis of any risk that may have a special impact on the previously mentioned elements. In addition, appropriate measures must be implemented to allow its mitigation.

The risk management process must be addressed at all stages of the AI system's lifecycle, from design and development to commercialization and post-commercialization.

The fundamental concepts that revolve around the risk management ecosystem are detailed below and arranged in a diagram that represents the relationships between them:



- Our **AI system** may be exposed to different **threats** that could end up posing a **risk** to our system and, consequently, to **people's health, safety and fundamental rights.**

Example

In the employee promotion use case, let's say we are an organization that decides to develop and market an **AI system** that analyses worker profiles and performance and helps determine who is most deserving of a promotion.

A possible **threat** associated with this system would be, for example, an **external malicious agent that tries to contaminate the training data** that the system analyses to disrupt the recommendations it provides. This may lead to a discriminatory promotion decision for some employees.

- These **threats** can materialize into **risk** through the exploitation of **vulnerabilities** in components of our system. If we do not have the **appropriate control measures** in place, we may be **vulnerable** to these **threats**.

Example

A possible **control measure** would be, for example, the implementation of an adverse data identification tool in the AI system. This tool analyses our training data and tries to determine if there is data that has been modified or entered by an external agent in an unwanted way.

- These **threats** that can exploit a **vulnerability** of system components can materialize with a certain **probability** and produce a certain **impact**.
Risk is determined by the value of the impact multiplied by the probability of it occurring (risk = impact*probability). We'll look at this in more detail in section 4.4.

Example

Following our example, we will need to determine the **impact** on **the health, safety and fundamental rights** of **employees** of the fact that an **external agent disrupts the data (threat)** with which the system makes decisions about who deserves a promotion the most.

We will also have to analyse the **probability** of this happening. This type of **threat** could be a discriminatory promotion of one employee over another.

In addition to mitigating risk, there are **other ways to deal with risk**.

These can be **assumed** (*accepting their consequences*), **avoided** (*deciding not to start or discontinue the activity that generates the risk*) or **transferred** *(for example, by taking out an insurance policy*). This will be discussed in more detail in section 4.5.

These **elements in more detail and additional elements** that belong to the risk management ecosystem from a more complete perspective, are the ones we will analyse in section 4.

2.2 Proportionality in risk management

The measures described throughout this document are intended to serve as a guide for providers of AI systems to comply with the requirements of the article "Risk management system".

In this sense, for the correct implementation of a risk management system, it is recommended that every organization address the phases set out in this guide.

While the comprehensiveness and depth with which each phase is addressed may depend on the needs and resources of each organization, the risk to **health, safety, and fundamental rights of individuals** must be given special consideration.

Also, the control measures implemented should also be carefully selected: each organization will determine which measures are appropriate and necessary to comply with the requirements of the AI Act to guarantee **the health, safety and fundamental rights of people**.

The measures described in this guide may inspire organizations to define and implement new measures, or the organization may determine the implementation of different measures.

The measures described in this guide may be implemented by each organization in proportion to the risk to be analysed and mitigated, their impact on the organization and the cost of their implementation.

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 risk management. Those requirements aim to ensure that high-risk AI systems available in the Union, or whose output is used in the Union do not pose unacceptable risk to important public interests recognised and protected by Union law.

This section includes the articles referring to the generation of risk management of Regulation 2024/1689 of the European Parliament and of the Council, of 13 June 2024 (European Regulation on Artificial Intelligence) 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

The AI Act addresses risk management systems mainly in **Article 9**, which regulates the **implementation and maintenance** of these for high-risk artificial intelligence (AI) systems, establishing an **iterative** and continuous approach that covers the entire life cycle of the system.

Its main objective is to **identify, analyse, assess and mitigate potential risk** related to health, safety and fundamental rights, both in intended use and in reasonably foreseeable uses. The article also includes monitoring mechanisms, testing, and systematic updates, with an emphasis on technical mitigation, proper design, and training of deployers.

It establishes the obligation **to carry out regular tests and check-ups**, prioritising the minimisation of residual risk and adapting the measures according to the purpose and context of use, with special attention to minors and vulnerable groups.

3.2 Article content

AI Act

Art.9 – Risk management system

1. A risk management system shall be established, implemented, documented and maintained in relation to high-risk AI systems.

2. The risk management system shall be understood as a continuous iterative process planned and run throughout the entire lifecycle of a high-risk AI system, requiring regular systematic review and updating. It shall comprise the following steps:

> (a) the identification and analysis of the known and the reasonably foreseeable risk that the high-risk AI system can pose to health, safety or fundamental rights when the high-risk AI system is used in accordance with its intended purpose;
>
> (b) the estimation and evaluation of the risk that may emerge when the high-risk AI system is used in accordance with its intended purpose, and under conditions of reasonably foreseeable misuse;
>
> (c) the evaluation of other risk possibly arising, based on the analysis of data gathered from the post-market monitoring system referred to in Article 72;
>
> (d) the adoption of appropriate and targeted risk management measures designed to address the risk identified pursuant to point (a).

3. The risk referred to in this Article shall concern only those which may be reasonably mitigated or eliminated through the development or design of the high-risk AI system, or the provision of adequate technical information.

4. The risk management measures referred to in paragraph 2, point (d), shall give due consideration to the effects and possible interaction resulting from the combined application of the requirements set out in this Section, with a view to minimising risk more effectively while achieving an appropriate balance in implementing the measures to fulfil those requirements.

5. The risk management measures referred to in paragraph 2, point (d), shall be such that the relevant residual risk associated with each hazard, as well as the overall residual risk of the high-risk AI systems is judged to be acceptable. In identifying the most appropriate risk management measures, the following shall be ensured:

> (a) elimination or reduction of risk identified and evaluated pursuant to paragraph 2 in as far as technically feasible through adequate design and development of the high-risk AI system;
>
> (b) where appropriate, implementation of adequate mitigation and control measures addressing risk that cannot be eliminated;

(c) provision of information required pursuant to Article 13 and, where appropriate, training to deployers.

With a view to eliminating or reducing risk related to the use of the high-risk AI system, due consideration shall be given to the technical knowledge, experience, education, the training to be expected by the deployer, and the presumable context in which the system is intended to be used.

6. High-risk AI systems shall be tested for the purpose of identifying the most appropriate and targeted risk management measures. Testing shall ensure that high-risk AI systems perform consistently for their intended purpose and that they are in compliance with the requirements set out in this Section.

7. Testing procedures may include testing in real-world conditions in accordance with Article 60.

8. The testing of high-risk AI systems shall be performed, as appropriate, at any time throughout the development process, and, in any event, prior to their being placed on the market or put into service. Testing shall be carried out against prior defined metrics and probabilistic thresholds that are appropriate to the intended purpose of the high-risk AI system.

9. When implementing the risk management system as provided for in paragraphs 1 to 7, providers shall give consideration to whether in view of its intended purpose the high-risk AI system is likely to have an adverse impact on persons under the age of 18 and, as appropriate, other vulnerable groups.

10. For providers of high-risk AI systems that are subject to requirements regarding internal risk management processes under other relevant provisions of Union law, the aspects provided in paragraphs 1 to 9 may be part of, or combined with, the risk management procedures established pursuant to that law.

3.3 Correspondence of the articles 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 requirement	Section
9.1	What elements should I implement and how should I do it to develop an adequate risk management system?	Section 4
9.2.a	Identification of risk Risk analysis and assessment	Sections 4.3 and 4.4
9.2.b	Identification of risk Risk analysis and assessment	Sections 4.3 and 4.4
9.2.c	Assessment of risk related to the post-market monitoring system	Section 5.2
9.2.d	Risk response	Section 5.5
9.3		
9.4		
9.5.a		
9.5.b		
9.5.c		
9.6	Tests to verify that HRAIS function as intended purpose and comply with risk management requirements	Sections 5.1.1 and 5.1.2
9.7	Testing in real-world conditions	Section 5.1.3
9.8	Right time for testing	Section 5.1.4
9.9	Access and impact of the system by children under 18 years of age	Section 5.1
9.10	Entities subject to sectoral legislation	Section 5.4

4. What elements should I implement and how should I do it to develop an adequate risk management system?

A **risk management system** is based on the implementation of **processes** that allow **the identification and evaluation of risk** and the **definition and implementation** of **timely treatment measures** to **manage** their impact.

As we have already indicated, the main risk that we must manage are those that may affect **people's health, safety and fundamental rights.**

In the diagram below, the **processes** that make up a **risk management system are detailed** and then, we will try to explain what they consist of and how to implement them.

4.1 Determining risk appetite

What is it?

Generally speaking, it is the level of risk that one is willing to accept.

In the context of the European Regulation on Artificial Intelligence, it is the level of risk that we will be willing to accept in relation to the risk that the AI system may pose to **the health, safety and fundamental rights of individuals.**

How should I approach it?

To do this, we should establish risk appetite. That is, if our AI system can have a critical impact on a person's life (for example, the AI system that delivers insulin to patients) the appetite for risk should be very low. On the other hand, if our AI system has no impact on people's **health, safety and fundamental rights** (for example, an AI system used to

recommend films or series, based on our tastes and preferences) the appetite for risk could be much higher.

Risk appetite should be defined quantitatively. First, a scale is established and then the level of risk on that scale is selected.

Generally, the scale is defined by setting a minimum value (usually equal to 1) and a maximum value. The maximum value is determined by the maximum value of the risk which, as we have seen in section 2.1, is equal to the product of the impact by the probability.

As mentioned there, the risk assessment process will be discussed in detail in section 4.4. However, suppose we select a 3-level scale for the probability of the risk happening and 5 levels for the impact it could generate. In this case, the maximum value is equal to 15 and on this scale, we must determine what the risk appetite is (with a value around 12 high and a value around 3 low).

Example

In our example in the previous section, we said that the inappropriate promotion of some employees over others could pose a **risk** to their **fundamental rights**, in particular if the AI system were to lead to promotions based on **discriminatory decisions**. Suppose we have determined (we will do this in section 4.4) that this risk, if it happens, would result in a level 9 out of 15 (probability equal to 3 and impact equal to 3). If we have defined a **risk appetite threshold** equal to 4 (since we are not willing to accept much risk given the potential impact of the system on the **fundamental rights** of workers) it means that we are not willing to assume the level of risk that it entails, so we will have to determine the corresponding **control** measures until that level is no longer above the defined threshold (this whole process will be addressed in more detail in section 4.5).

Additionally, for more details on how this exercise is approached and where and how it is reflected, it is advisable to consult the Excel document indicated in section 1.4.

Why is it important?

It is one of the foundations of risk management, as it is the way we have to quantify the level of risk that we will accept after analysing and evaluating the risk that we identify in later phases. It is the element that will allow us to assess the risk treatment measures that we need and the point at which these are sufficient to guarantee a level of risk that we are willing to accept. Without determining risk appetite, risk management would be reduced to the stage of identifying risk, because if we cannot determine how, a risk's impact should be assessed, the whole phase of analysis and evaluation of risk would lose meaning and risk treatment measures would not be established either.

4.2 Context of the AI system

What is it?

It is the external and internal environment, where the AI system is designed, developed, and used. In Annex A.I. A list of some relevant examples is provided (*among them are economic and regulatory factors, the technological context of the organization, the level of*

maturity and complexity of AI systems in the organization or the culture surrounding data and its use in the organization).

However, the most important elements of this context in which the AI system is designed, developed and used are those related to guaranteeing **the health, safety and fundamental rights of people.**

For this reason, we have included in Annex A.II a list of the main rights of the Charter of Fundamental Rights of the European Union. In addition, a general description of the Charter and a brief example for those rights most relevant from the perspective of the European Regulation on Artificial Intelligence are included in accordance with recital 28.

How should I approach it?

What we will do in this phase will be to inventory in a document (an Excel sheet, for example) those elements of the context that could impact the risk analysis, paying special attention to those related to **health, safety and the fundamental rights of people**. It is important to bear in mind that this process is a mostly qualitative and subjective process and whose accuracy and completeness will depend on our knowledge of these elements of the environment.

Once these elements have been documented, we will have to see how they can affect the risk analysis, that is, we will have to ask ourselves the following question: *Can these elements pose an additional risk that I should incorporate into my risk identification process?* (Notice, we will see this process more in detail in Section 4.3)

Example

Continuing with the example developed in the previous sections, we must identify the most relevant elements of the context in which the **AI system for employee promotion** is designed, developed and used. As we have seen, we must pay special attention to those that may be related to **guaranteeing the health, safety and fundamental rights of people.**

In this sense, one of the most important elements of the context of an AI system used to promote employees is compliance with the **fundamental right to non-discrimination,** so is compliance with the fundamental right related to the **protection of personal data**. These elements must be identified and inventoried in the documentation prepared for the risk management system and are the basis that will help us to properly identify the main threats as we will see in the following sections.

Additionally, for more details on how this exercise is conducted and where and how it is documented, it is advisable to consult the Excel document indicated in section 1.4.

Why is it important?

Understanding the environment and context in which the AI system is developed is the basis for the identification of risk that could affect **people's health, safety and fundamental rights.** Therefore, it is essential to analyse this element with due accuracy and

depth, since an inadequate evaluation of the context and the environment will result in an incomplete risk identification.

4.3 Identification of risk

What is it?

It is the process of discovering, recognizing and documenting the different risk that can affect our AI system and, consequently, the **health, safety and fundamental rights of individuals.**

How should I approach it?

The objective of this phase will be to complete an inventory (for example, in an Excel document) with all those possible risks that we consider may affect our AI system. To do this, we will follow a procedure as described below (brief examples are included in each stage described, although, as indicated at the end of the section, the complete exercise is set out in the Excel indicated):

1. First, we must **identify the components of our AI system**. These components define each system and differentiate it from the rest.
   To facilitate the identification of the components of an AI system, a list of some of the most common components of AI systems has been included in Annex B.
2. The second step is to **identify the threats associated with these components** through internal and external context analysis. The risk affects our AI system through its components (see example below).
3. These **threats can materialize into risk through the exploitation of vulnerabilities in components of** our system.

Example

In our example, we identified as **elements of the context** the need to comply with the **fundamental right to non-discrimination** and compliance with the **fundamental right** related to the **protection of personal data.**

As we indicated in the previous section, context is the **basis** for the identification of threats and consequently, risk. Thus, in this example, we are going to identify each of the elements of the aforementioned context with a component of the AI system and with an associated risk:

System	Component	Threat	Risk
Employee Promotion System	Training Data	Over- or under-representation of a set in the training data database	Discrimination of some employees against others in promotion
	Data Owner	Intentional or unintentional data breach by the owner	Disclosure of employee personal data

It is important to note, once again, that the European Regulation on Artificial Intelligence places special emphasis on **risk that may affect individuals' health, safety and fundamental rights.** In this regard, special attention must be paid to the identification of any risk that may affect the aforementioned elements. For example, in an AI system responsible for automatically administering insulin to a diabetic patient, the risk associated with a failure to read the parameters that allow determining the amount of insulin that the patient needs at a given time has a direct impact on the health and life of that patient. On the other hand, a risk of that same system associated with a failure to send a monthly report of the doses administered throughout the month to the patient's mobile application does not have a direct impact as representative as the previous one on the patient's health or life.

Additionally, for more details on how this exercise is approached and where and how it is reflected, it is advisable to consult the Excel document indicated in section 1.4.

Why is it important?

The identification of risk is important since only the identified risk can be assessed and receive the appropriate response. When an organization fails to identify risk properly, it is completely outside the risk management system.

4.4 Risk analysis and assessment

What is it?

It is the process of analysing and assessing the risk identified in the predecessor stage by determining, for each of them, the probability of the threat materialising and the magnitude of the impact on the component.

How should I approach it?

The objective of this phase is to determine the impact and probability of occurrence of each risk identified in the previous phase. Therefore, we will take our document where we are recording the risk management process and we will incorporate two new attributes, impact and probability (it is recommended to consult and follow the example developed in detail in the Excel document indicated in section 1.4).

Example

In the previous sections, we have used the **AI system for employee promotion and** as an example and we have identified as possible risk **the discrimination of some employees against others in the promotion and disclosure of employees' personal data**.

What we need to do now is decide what **magnitude of impact we give to this risk**, for this we will determine a scale in a similar way to how we did in section 4.1 when we defined risk appetite. And we'll do the same with probability.

At this point, several questions may arise, such as, under what criteria do I decide the level of impact that a risk has? how can I determine the probability that it will materialize? how will I establish the value of the final risk? how is it related to probability and impact? and to risk appetite? Below, we will try to answer these questions.

It should be remembered that risk assessment is a qualitative and subjective exercise, which must make it easier to know the AI systems that are in place and the main risk that could affect them and, therefore, **the health, safety and fundamental rights of people.**

The value or level of risk will be calculated as the product of the impact on the component times the probability of the threat occurring. It is not a matter of quantifying the impact and probability of a threat materializing with numerical accuracy, but of assessing them on a conceptual scale, and being consistent in their application to all cases. Thus, for example, we can define a scale from 1 to 5 to categorise risk into 5 levels of impact (in ascending order, very low or non-existent, low, medium, high, very high or critical) and a scale from 1 to 3 to categorise the probability of these happening (in ascending order, unlikely, probable, very likely). That said, the maximum risk level I could encounter would be equal to 15. We must consider that this scale must be the same as the one defined for risk appetite, and what we will do next is compare them.

Once the impact and probability levels have been established and the final value of the risk has been calculated, we must compare it with the risk appetite initially determined.

Example

Suppose that, for the **risk** related to **the disclosure of employees' personal data**, we have determined that it has a **probability** of happening of 3 on the defined scale of 1 to 3. Why 3? for example, because we do not have **any access control** to the database that contains the set of training data **or cybersecurity controls** that protect it from external malicious actors.

If this happens, we determine that it would have a **serious impact on employees**, as it would violate one of the **fundamental rights**, the **protection of personal data**. Let's say, then, that an impact 4 on the defined scale of 1 to 5.

The **result of the analysis of this risk** would be equal to 12 (product of 3 times 4). If we had defined a **risk appetite** equal to 4, we would be above the threshold and therefore would not be willing to assume this risk.

What should we do now? Decide what type of risk response is the best fit for this scenario (avoid, transfer, mitigate, accept). For example, mitigate it by establishing measures (such as access controls or cybersecurity) that reduce it until it drops below the defined appetite threshold. In section 4.5 we will delve into the different risk measures and treatments that we can address.

Additionally, for more details on how this exercise is approached and where and how it is reflected, it is advisable to consult the Excel document indicated in section 1.4.

Why is it important?

The analysis and assessment of risk makes it possible to quantify the level of risk to our system and, consequently, **to the health, safety and fundamental rights of people,** the risk identified in the previous stage.

Without this process of assessing the risk, we would not have the capacity to determine, quantitatively, whether we accept them or, on the contrary, we need to define and implement treatment measures to manage them properly.

4.5 Risk response

What is it?

It is the process of selecting and implementing measures to address the risk identified and analysed in the previous stages.

How should I approach it?

In this phase we are going to define the risk response measures that we will incorporate in our document where we are recording the risk management process. To do this, we will follow a procedure as described below (brief examples are included in each stage described, although, as indicated at the end of the section, the complete exercise is set out in the Excel indicated in section 1.4):

1. We must first **determine what type of risk treatment measures** we will select to address each of the risk identified and analysed. The main options are as follows:

    a. **Implement control measures to mitigate risk:**
    Normally, the first option we consider when dealing with a risk whose value exceeds our risk appetite will be to implement additional controls to mitigate its probability or impact. Some examples of AI-specific controls are set out in Annex D.

Example

Continuing with our example of the **AI system** for **employee promotion** and the **identified risk** associated with the **disclosure of employee personal data**, one measure that would mitigate this risk is the implementation of an RBAC (*Role Based Access Control*) system) or access control system and data rights management.

With this control we will be able to reduce the probability of the risk occurring. In the more detailed example in section 1.4 we will see how the value of the risk is reduced by implementing these controls, but let's assume for now that the value is reduced from the original value 12 to a value equal to 4, on the same scale from 1 to 15.

b. **Assume the risk**, accepting the consequences that it would have if it happened.

Example

In the previous section we saw how, after implementing a **risk mitigation measure, the risk** was reduced from a value of 12 to a value of 4. Initially we had defined a **risk appetite** equal to 4, now we are below that threshold.

In this scenario, we could implement more control measures to further reduce risk. But we could also determine that, since the current **risk value** is already **below** our **risk appetite, we should not implement additional control measures** and accept the current risk level.

In the more detailed example in section 1.4 we will see cases representing this scenario and how it is reflected in the document that includes our risk management system.

c. **Avoid risk** by deciding not to start or discontinue the activity that generates the risk.

Example

Continuing with our example, after analysing the risk associated with our AI system for employee promotion, we may find ourselves in a **scenario where the number of risk**, their probability of happening and their respective impacts on **people's health, safety and fundamental rights** are so high that they make us **consider** that we may not be interested in **continuing to use the system** or discontinuing it.

d. **Transfer risk**, for example, through contracts, such as the purchase of insurance.

Example

In the context of the AI Act, it should be noted that risk management, as we have already discussed throughout the guide, focuses on the management of those risk that may affect **the health, safety and fundamental rights of individuals**.

The transfer of risk is a measure, by its nature, applicable in a context of risk management for the organization. For example, suppose that the **risk** associated with the **possible loss of the organization's information** is **so high** that we decide **to hire** the services of an **external company** that will ensure **that** our **data is not manipulated** and that we reach an agreement with said company according to which in the event that they **are not manipulated. If they suffer unwanted manipulation**, the company **will compensate** us with an agreed financial amount.

2. Secondly, and once the risk treatment measures have been defined, we must establish a **plan for their implementation**. To do this, for example, we can establish by means of a *Gantt chart* the sequence of implementation of the measures defined according to the priority and the commitment dates to implement them.
   The prioritisation of the implementation of measures can be established, for example, according to their mitigation impact, giving special priority to those that mitigate the most relevant risk.
3. Third, we will need **to document and report residual risk.** We must report those risk identified and analysed in the documentation associated with ours. For the proper preparation of this documentation, the transparency *and information provision guide to users* provided for this purpose must be consulted, as indicated in the AI Act, in section 5c of the article: "Risk Management System").
4. Finally, we will establish **periods for the review and monitoring** of the risk management system (all new risk that are identified must be incorporated, as well as evaluating and defining the appropriate treatment measures). We must also communicate the results and activities of risk management, as well as the updates derived throughout the organization (especially to stakeholders and/or impacted).

Additionally, for more details on how this exercise is approached and where and how it is reflected, it is advisable to consult the Excel document indicated in section 1.4.

Why is it important?

It is the phase where, after identifying, analysing and assessing the risk that threaten our AI system, we establish the appropriate measures to manage them.

Without implementing risk response measures, the risk management system would remain in a phase of knowing the threatening risk and, in the worst case, of knowing that these exceed the level of risk we are willing to assume. This phase is what allows us, once all this has been determined, to face the risk and try to reduce or mitigate their impact, in short, to respond to them.

4.6 Technical documentation of the risk management system

For proper documentation of the risk management system, we must follow the steps described in section 4 and record them in a document that is the same or similar to the one provided as an example in section 1.4. In addition, pay special attention to the documentation of residual risk, as explained in section 4.5 section 3.

For more detail, see section 6.

4.7 Communication and consultation

The actors involved in the design, development and commercialization of the AI system should be aware of the AI system's risk management system and collaborate in ensuring its completeness and updates.

Communication aims to promote awareness and understanding of risk.

Consultation involves obtaining feedback and information to support decision-making.

Communication and consultation with appropriate stakeholders, both external and internal, should take place at all stages of the risk management process.

For organizations that develop or use AI systems, they must identify which parts of the organization are involved. The organisation must be aware that the application of AI technologies can have a greater impact than the application of other technologies and pay special attention to those impacts on people's **health, safety and fundamental rights.**

4.8 Monitoring and continuous improvement

Its purpose is to ensure and improve the quality and effectiveness of the risk management system. It is essential to develop a follow-up and establish certain periods for reviewing and updating the risk management system. In this way, the inventory of risk must be reviewed and any additional ones that can be identified must be incorporated. The analysis and

evaluation of new and existing risk should also be addressed and reviewed. Finally, timely mitigation measures will also need to be reviewed and updated accordingly.

As we mentioned at the beginning of the guide, the risk management process must be addressed at all stages of the AI system's lifecycle, from design and development to commercialization and post-commercialization. Therefore, the monitoring and updating of risk must also be carried out at each stage.

4.9 Leadership and commitment

Risk management is the process by which an organization's management supervises, leads, and commits to the development of the risk management system and ensures the integration of the different risk management systems of the organization (e.g. the risk management of the AI systems with the rest).

Formalizing leadership and commitment to risk management is another fundamental pillars in the development of a risk management system.

Senior management and supervisory bodies, where applicable, should ensure that lifecycle risk management of AI systems is integrated with risk management in the rest of the organisation. They can do this by, for example:

a) **The publication of a statement or policy** that establishes an approach, plan or course of action for the risk management described in ANNEX G.
b) **Ensuring that there are sufficient resources** allocated to risk management.
c) **The assignment of authority, responsibility, and accountability** at appropriate levels within the organization.

5. Other elements to consider

5.1 Test procedures

The AI Act, in governing the risk management system, establishes that the system must be tested to ensure that it fulfils its purpose and that the requirements established for HRAIS are met. To this end, it allows testing in real-world conditions.

These tests must be designed according to parameters, metrics and thresholds defined in accordance with the intended purpose of the AI system.

5.1.1 Tests to verify that HRAIS function as intended

To ensure that the AI system works as intended purpose, the *Accuracy and Robustness guides* developed to facilitate compliance with the article "Accuracy, robustness and cybersecurity" of the Regulation should be consulted.

These guides define measures and metrics that allow for assessing the parameters necessary to ensure that the system is working as intended and will continue to do so over time.

5.1.2 Tests to verify that HRAIS comply with the established requirements

To ensure that the measures developed and implemented are adequate, indicators and metrics may be established that allow the measurement and validation of the effectiveness of these measures.

One way to address this aspect is by defining **effectiveness indicators** . An effectiveness indicator is a metric that makes it easier to measure the degree to which a given goal has been met or achieved.

In order to develop the tests that will help ensure that the requirements of Chapter 2 are properly met, the provisions of each of them must be understood and addressed. The guides developed for this purpose (*Data and Data Governance, Technical Documentation, Record Keeping, Transparency and Provision of Information to Users, Human Oversigth, Accuracy, Robustness and Cybersecurity) should be consulted*.

Additionally, these guides will incorporate an inventory of tests or indications that will help the reader verify that they have covered the aspects established in the corresponding article.

To guide the reader in the definition of these indicators, some examples of effectiveness indicators in relation to the risk management measures described in this guide and additional ones related to the rest of the requirements have been incorporated in Annex E.

5.1.3 Testing in real-world conditions

The test procedures that are designed may include testing in real-world conditions, if the organization deems it appropriate or necessary, following the provisions of the article "*Testing of high-risk AI systems in real-world conditions outside the AI regulatory sandboxes*."

This article sets out the conditions under which providers of HRAIS in Annex III of the European Regulation on Artificial Intelligence may carry out testing in real-world conditions.

The following points provide these conditions in a summarized manner:

- Testing in real-world conditions must always be carried out before placing the system on the market or putting into service.
- Any subject may withdraw from the test at any time by revoking their informed consent without having to give any justification.
- Any serious incident detected in the course of the tests shall be reported to the appropriate authorities referred to in paragraph 6 of the Article.
- The provider and the potential provider shall be liable under Union and Member State liability law for any damage caused to the subjects by their participation in the testing in real-world conditions.

5.1.4 Timing of testing

The tests designed and implemented for the development of AI systems can be carried out at any time during their development, but always before they are commercialized or put into service.

However, it should be considered that the tests must be planned during design, carried out during implementation and be aligned with the intended purpose and the risk identified before going into production.

5.2 Assessment of risk related to the post-market monitoring system

The development of the risk management system shall take into account the evaluation of other risk that may arise based on the analysis of the data collected in the post-market monitoring system referred to in the article "*Post-market monitoring by providers and post-market monitoring plan for high-risk AI systems*".

This article describes the monitoring that providers will have to carry out after the commercialisation of the IA system and the main elements that must be contained in the post-market monitoring plan that they will have to develop for those systems.

To facilitate a clear understanding of this process, the guide that describes the post-market monitoring plan to be prepared by the providers (*guide for the preparation of a post-market monitoring system*) has been developed.

Providers of AI systems shall assess other risk that may arise from the analysis of the data collected in the post-market monitoring system.

To carry out this risk assessment, the steps described in section 4 should be followed, so that:

a) **The identification process** includes the risk that may arise from the analysis of the data collected in the post-market monitoring system. You will need to consider and analyse the context and environments in which the system will be used.
Whoever markets an AI system responsible for automatically administering insulin to a diabetic patient must take into account the various risk that may arise during the operation and use of the system by the patient.
b) **The risk identified in the previous point are incorporated** into the risk analysis and assessment process.
c) **In the process of adopting the measures and controls** for the management of risk, the risk identified, analysed and evaluated in the previous points must be taken into account.

5.3 Access and impact of the system by children under 18 years of age

During the development of the risk management system, it should be determined whether it is likely that children under 18 years of age will access the system, or whether it will have an impact on them.

In this context, the possible derived risk that may have an impact on these minors under 18 years of age must be analysed and incorporated into the risk management system designed. In this sense, and in a similar way to the previous section, we must:

a) **In the process of identifying** risk, include those that could arise from the use or impact on minors under 18 years of age.
For example, if we have an AI system that helps in processes of granting social aid, we must consider whether there are children under 18 years of age among the group affected by this system. Another example of an AI system that could affect children under the age of 18 would be an AI system used to determine the access or admission of natural persons to educational programmes or centres.

As the guide explains, this risk identification process will be determined by the knowledge of the organization that develops the risk management system for its AI system of the context and environment in which it is developed.
b) **In the process of analysis and risk assessment**, incorporate the risk identified in the previous point.
c) **In the process of adopting measures and controls** for risk management, consider the risk identified, analysed and evaluated in the previous points.

5.4 Entities subject to sectoral legislation

The providers AI systems subjects to requirements of internal risk management processes under the relevant Union sectoral legislation may incorporate the measures described in this guide as part of the risk management procedures of that legislation.

6. Technical documentation

Article 9 establishes that artificial intelligence systems must undergo a continuous and iterative process of risk identification, analysis, assessment and mitigation throughout their entire life cycle. The documentation associated with this process must clearly and comprehensively reflect how these phases have been applied, providing the necessary information to demonstrate compliance with the requirements of the AI Act.

In accordance with the above, and in line with the relevant sections of Annex IV, the risk management system documentation must include the elements that are relevant to adequately describe the framework, the methodologies used, the decisions taken and the results obtained.

For the proper documentation of the risk management system, we must follow the steps described in section 4 and reflect them in a document identical or similar to the one provided as an example in section 1.4. In addition, special attention should be paid to the documentation of residual risks, as explained in section 4.5, paragraph 3.

In order to facilitate understanding of the development of a risk management system and its documentation, the document 'Illustrative example of the development of a risk management system.xlsx' is attached, which shows a practical example for various use cases. This document details the development for the use cases described in section 4.3, following each of the phases of the guide, from the definition of risk appetite to the selection of response measures.

The document includes explanations that link the guide and the phases described with the development of the risk management system in each example.

7. 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 European Regulation on Artificial Intelligence 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.

8. Annexes

8.1 Annex A - Most relevant elements of the internal and external context around AI

8.1.1 Annex A.I - General Elements

Below is a detailed list of the most relevant elements that complete this environment [1] [2]:

a) **Social, cultural, political, legal, regulatory, financial, technological, economic, and environmental factors**. It can be useful to use guides and guides on ethical issues in AI, such as the guides for ethical and trustworthy AI published by the EC in 2019.
b) **The main technological trends, advances in areas of AI and the social and political implications** of the deployment of these technologies that can affect our system and, consequently, **the health, safety and fundamental rights of people.**
c) **The main stakeholders, their perceptions, values, needs and expectations.** These can be affected by issues such as the lack of transparency of AI systems or biased AI systems.
d) **The complexity of networks and their dependencies**, which can increase with the use of AI technologies.
e) **The internal organizational factors** that revolve around the vision, mission, values, culture, strategy, governance model, policies, rules and procedures adopted, and contractual relationships and commitments.
f) **The culture of the organization**, as well as guides, models and standards adopted.
g) **The organization's capabilities, resources, and expertise** in relation to AI, it is important to take into consideration issues of transparency of AI systems, variation in resources needed, and specific knowledge requirements in AI and data science technologies, these can be causes of additional risk.
h) **The use of data and information flows**. AI systems can be used to automate, optimize, and improve data processing.
i) **Relationships with internal stakeholders**, taking into account their perceptions and values. Stakeholder perception can be affected by issues such as the lack of transparency of AI systems or biased AI systems. Stakeholders should be provided with information on the capabilities, failure modes, and failure mitigation of AI systems. Although this aspect includes the entire organization internally, it must be considered especially in the field of AI systems, the departments and professionals involved in the conception, implementation, exploitation, and evolution of AI systems.

8.1.2 Annex A.II - Elements related to the EU Charter of Fundamental Rights

The adverse consequences that a high-risk AI system can have on the rights protected by the Charter of Fundamental Rights of the European Union [11] can be really serious.

Therefore, during the life cycle of Artificial Intelligence systems, it is necessary to adopt an approach that seeks to ensure a high level of protection of these rights.

Specifically, recital 28 of the AI Act mentions those fundamental rights of the Charter that may be particularly affected by the development and deployment of Artificial Intelligence AI systems.

This section mentions each of the rights and principles recognised by the Charter of Fundamental Rights of the European Union that are expressly mentioned in Recital 28 of the AI Act.

In addition, an example of the possible impacts that the use of an artificial intelligence system can have on each of these fundamental rights is incorporated[1]:

1) Right to human dignity: Article 1 of the Charter.

The dignity of the human person is not only a fundamental right in itself but constitutes the real basis or presupposition of fundamental rights.

The exercise of rights cannot be used to injure the dignity of another person. Likewise, the necessary restrictions on fundamental rights must respect dignity.

High-risk AI systems can impact not only on specific fundamental rights, but also on several and even many of them. In these cases, a vision from dignity allows these varied risks to be brought together. Likewise, a vision of dignity allows us to elevate the analysis of risk to a vision that transcends the possible effects on the individual rights of specific people, to address the impact that the use of high-risk AI can imply in a joint vision of the rights of large groups of people and even of society as a whole.

2) Respect for private and family life. Article 7 of the Charter.

For decades, information and communication technologies have had a special impact on private and family life, at home and in communications. To a large extent, private life in general and, as we will see, data protection in particular, can be particularly affected when a high-risk AI system is used on specific people, which is the most common.

[1] For an exhaustive analysis of each of the fundamental rights recognised by the Charter of Fundamental Rights , it is recommended to visit the website of the European Union Agency for Fundamental Rights. http://fra.europa.eu/es

The legal texts mentioned in this section are as follows:

Charter of Fundamental Rights of the European Union. (2000)

Committee on the Rights of the Child's General Comment No. 25 on children's rights in relation to the digital environment. (2021)

Spanish Digital Bill of Rights. (2021)

Possible violation: Annex III. 1.a) of the Regulation

Especially conflictive are high-risk biometric identification systems that are not directly prohibited. And above all, HRAIS systems that incorporate an emotion recognition system to detect or deduce the mental states, emotions or intentions of natural persons from their biometric data; or systems that assign people to specific categories based on their, have an impact on private life. biometric data. Although they are not prohibited, the choice to use of these systems must be very identified and the risk analysis must justify very well the need and legitimacy of the use of this type of system as long as there is no other alternative that has less impact on private life.

The examples are very abundant. Before the regulation of the AI Act, several cases are already incompatible with various fundamental rights, others are doubtfully legal.

In the UK, since 2017 the South Wales Police have been running an "AFR Locate" project for events such as the Champions League final, international rugby matches, concerts, a Christmas Day on a busy shopping street in Cardiff, etc. The images were captured, processed automatically, and checked against people on watch lists. Its use was announced with posters and warned on social networks. In 2019 the courts gave it the go-ahead.

In Germany, on the occasion of a G20 meeting in Hamburg in 2017, a facial recognition system based on recordings was implemented for the detection and investigation of crimes. The Hamburg data authority deemed it inadmissible, but a court overturned the suspension.

In Sweden, in 2021, the data authority sanctioned police officers who decided on their own to use facial recognition software with images from social networks and websites for police and investigative purposes.

In Buenos Aires , a system with 300 active cameras began to be implemented in 2019 that allowed the identification of fugitives and generated some 10 million inquiries about non-fugitives. Guillermo Federico Ibarrola was erroneously identified as a fugitive and was detained for 6 days. On April 12, 2022, the system was suspended by a judge and has finally been annulled by a judgment of September 7, 2022.

In Spain, Mercadona was heavily sanctioned in 2021 for implementing a smart biometric system that controlled whether those who accessed some establishments were on its search lists for previous judicial reasons.

In Brazil, the São Paulo Metro implemented a security control system for 4 million daily users. On May 7, 2021, the São Paulo Court of Justice prohibited the concessionaire of the São Paulo Metro from using the "Digital Interactive Door System" (DID) with facial recognition. The system inferred people's emotions, gender, and age to personalize advertising. Finally, it was judicially suspended on March 22, 2022.

In the US, the Automated Virtual Agent for Real-Time Truth Assessment (AVATAR) analyses the non-verbal and verbal behaviour of travellers, and the system has apparently also been tested at **Bucharest** airport. The European Commission funded the "Intelligent Portable Control System" (iBorderCtrl) project, with deception detection and risk-based assessment tools that has generated a relative reaction from civil society. It has generated a relative

reaction from civil society, a European citizens' initiative and the reclaimyourface.eu campaign.

Intelligent biometric systems allow the recognition of emotions, categorize people, detect behaviours, thoughts or assess personality.

The truth is that these systems impact not only on people's privacy and data protection, but, as indicated below, on many other freedoms, especially due to their use in public spaces and the inhibiting effect they cause.

Possible Violation: Annex II of the Regulation

High-risk AI systems that are Annex II products or safety components, such as toys or recreational boats, may have the ability to capture data from their environment of a diverse nature (images, sounds, geolocation, etc.). In the risk analysis, special care must be taken to ensure that this capture is the minimum necessary for functionality and that, in any case, the information extracted is not used for any other purpose and especially for profiling or personality assessments.

Annex II devices and systems, such as toys, have already generated controversy due to their internet connection and data extraction, including disproportionate images[2].

The AEPD has indicated the basic guides that must be complied with in these cases[3].

INCIBE has also pointed out the basic security elements to be met by toys to avoid being hacked[4].

3) Protection of personal data. Article 8 of the Charter.

Based on the right to privacy and as one of its components, the right to the protection of personal data has been increasingly recognized and specifically developed over recent decades. All content derived from this right and its European and national regulation must be guaranteed by the high-risk AI system in question. Whenever a HRAIS system processes data of identified or identifiable individuals at any stage of its lifecycle, it will affect data protection. The technique of risk management and impact assessment is particularly developed in the field of data protection and can be referred to the many instruments to carry out this in general and in the case of AI systems in particular. It should be especially noted that, if the high-risk AI system processes personal data, it will very likely be mandatory to carry out not only a general risk analysis, but also a special and exhaustive impact study

[2] See:

https://www.lavozdegalicia.es/noticia/sociedad/2017/12/21/lista-juguetes-espia-crece-tras-analisis-advierte-risks-intolerables-dos-robots/0003_201712G21P28991.htm

[3] More information at: https://www.lavozdegalicia.es/noticia/sociedad/2017/12/21/lista-juguetes-espia-crece-tras-analisis-advierte-risks-intolerables-dos-robots/0003_201712G21P28991.htm

https://www.aepd.es/es/node/824

[4] More information at:

https://www.incibe.es/incibe/informacion-corporativa/con-quien-trabajamos/proyectos-europeos/is4k

in terms of data protection (Article 35 GDPR). Similarly, it is very possible that the specific safeguards relating to Article 22 GDPR on automated decisions will have to be specifically taken into account.

It should also be noted that data protection risk analyses in recent years incorporate not only the impact that a processing of personal data has on private or family life or intimacy, but also include risk analyses with respect to other rights at stake, such as non-discrimination in data processing. Thus, when conducting a risk analysis or impact assessment of a high-risk AI system, it will be especially important for the organization to consider the tools that have been developed for data protection compliance. Moreover, the best practice will be to jointly develop these risk analyses of impact on rights and, to this end, integrate, cooperate or coordinate with the subjects in the organisation that have special powers in the matter, such as the data protection officer.

Possible Violation: Annex II and III of the Regulation

The examples mentioned above in the right to respect for private and family life are fully valid.

(4) Freedoms of expression and information, assembly and association. Articles 11 and 12 of the Charter.

AI systems can impact freedom of expression and information in many ways. It can be seen more clearly in autonomous systems for the management, control or moderation of platform content. In many cases, the impact can be produced in conjunction with the freedoms of assembly and association.

Possible violation: Annex III. 4.a) and Annex III. 3. b)

This impact can generally occur due to the interference that the use of high-risk AI systems can cause in their exercise. In many cases, the mere existence of the high-risk AI system can generate the inhibiting effect of the exercise of these freedoms, that is, the person who knows that there is a system that can capture and evaluate their thoughts, manifestations and expressions of them is very likely to decide not to express themselves or participate in associative activities or meetings. trade union or demands.

This can happen, for example, in the case of using emotion recognition or personality assessment systems incorporated into a high-risk AI system for education assessment or job selection. This will happen particularly in certain contexts and places where these systems are used or in relation to certain subjects (opinion makers in networks, union leaders, journalists, teachers, researchers, etc.). For this reason, the possible impact will have to be specifically analysed in these contexts or with these subjects and, if the high-risk AI system is to be used, it will be especially possible to control and mitigate and reduce these risks with the appropriate guarantees and techniques.

5) Non-discrimination. Article 21 of the Charter.

Any use of AI systems usually generates errors and biases, this is an anomaly in the output of the system due among other reasons to: prejudices or erroneous assumptions made during the system design process, prejudices in the training data, the autonomous development itself that has derived from the deployment of the high-performance AI

system. risk and its interaction with the environment where it is implemented. In some cases, these biases can lead to different treatment of people who should be treated equally. Moreover, errors or biases may lead to treating groups of people who are particularly prohibited from discriminating in a way: birth, racial or ethnic origin, sex, religion, conviction or opinion, age, disability, sexual orientation or identity, gender expression, disease or health condition, HIV status and/or genetic predisposition to suffer pathologies and disorders, language, socioeconomic situation (art. 2.1°, Law 15/2022).

Possible Violation: Annex III of the Regulations.

The assumptions of algorithmic discrimination are the most varied, many of them involuntary. U.S. airport body scanners (TSA) have flagged transgender travellers as more likely suspects to conduct a particular screening. In **the field of health**, a system applied approximately 200 million Americans indirectly penalized black people, who were left behind by the system with respect to whites with similar diseases and needs. Under the COMPAS "Correctional Offender Management Profiling for Alternative Sanctions" system, Black people are nearly twice as likely as whites to be labelled as at higher risk of reoffending.

In the **field of education**, in France the **Parcoursup admission system** distributes students in higher education establishments based on criteria provided for by law. Courts and authorities have demanded special transparency. In 2020 in the United Kingdom, due to Covid, the A-Level university entrance exams were not held, the regulatory body (Ofqual) generated a system based on an algorithm that affected almost one million people. The most common criticism is that public school students were penalized.

6) Consumer protection. Article 38 of the Charter.

The Charter of Fundamental Rights of the EU states that "*The policies of the Union shall ensure a high level of consumer protection*", which results in very broad regulations in the EU and the Member States in this area. High-risk AI systems, especially linked to products in Annex II of the European Regulation on Artificial Intelligence, can affect consumer rights to a large extent. Similarly, it is worth taking into account the obligations imposed especially by the Digital Markets Act, the European Regulation on Artificial Intelligence that also affects the use of AI systems in relation to consumers.

Possible Violation: Annex III 4.a) of the Regulation

Especially high-risk Annex II AI systems that are safety components of products or that are themselves products can affect consumers.

Also, the high-risk AI systems of Annex III relating to education services, solvency of natural persons or **establishing** their credit rating, pricing, offering of services, work or products through personalised advertisements or life and health insurance. In these contexts, special account must be taken of the rights and guarantees of consumers.

Personalized ads can give your recipients an advantage by informing them about certain services, products, or job openings. However, these same announcements can affect other groups that have normally been discriminated against. For example, one study has shown that **personalized job ads** on Facebook **can reinforce racial and gender stereotypes and**

biases at work. Thus, for supermarket cashier positions, they were shown to an audience made up of 85% women.[5]

(7) Workers' rights: Articles 27 to 33 of the Charter

Workers' rights are expressly recognized in Articles 27 to 33 of the Charter [11]. The content of these rights includes: the right of workers to be informed and consulted of the most relevant decisions taken by the company and that affect them, the right to collective bargaining and action, the right to strike, the right to conciliation at work and in person, the right to work in fair and equitable conditions, the right to social security or the right to obtain adequate protection in the event of unjustified dismissal.

Possible violation: Annex III. 4.b) of the Regulation

A potential breach of Article 27 of the Charter (Information and consultation) and implementing legislation would be that the artificial intelligence systems used by companies were not transparent enough to enable employers to adequately inform workers or their representatives of the rules and instructions on which the algorithms are based when they are used for the dismissal of a worker.

Possible violation: Annex III. 4b) of the Regulations.

A potential violation of Article 31 of the Charter (Fair and Equitable Conditions) by the developer of the AI system is the value or score it sets for the various variables of its algorithm. In a real case, a court considered that there was discrimination on the part of a company because the algorithm it used similarly scored a worker's absence or lack of punctuality with a lack of attendance due to strike, illness or care of dependent children [11]. In this case, the algorithm follows the instructions set by the company and implemented in the algorithm. Instructions that cause a discriminatory situation to these workers.

Possible violation: Annex III. 4.b) of the Regulation

A possible violation of Article 31 of the Charter (Fair and Equitable Conditions) is that the company establishes a system of constant monitoring of the worker that forces the latter to adopt behaviours that are not natural in the workplace (smiling continuously, being active at all times, being attentive, etc.). Especially when that permanent observation can be combined with a dismissal, a labour sanction, a salary reduction, etc.

(8) The rights of persons with disabilities: Article 26 of the Charter

To a large extent, it is worth referring to the fundamental rights of disabled people and the possible impacts produced by high-risk AI systems.

Possible violation. Annex III. 1. a) and Annex III. 5(a) of the Regulations

A possible violation of Article 26 of the Charter can occur if an organization uses an AI system in which, during its design, it has not been taken into account that this system may

[5] More information at:

https://arxiv.org/abs/1904.02095

yield more inaccuracies for certain people. For example, an organization uses an AI system to conduct video job interviews that analyses applicants' speech patterns to reach conclusions of their future performance at work. Those applicants for that job who have a certain speech disability, the system will likely give them a lower or unacceptable grade for the position.

In these cases, the violation does not have to occur during the design of the application by the provider of the AI system, it is much more likely that the violation comes from the hand of the organization that uses the system by not taking into account that this system can unjustifiably affect certain people with certain disabilities. Human oversight of these processes by the user organization can compensate for this initial situation. The provider of the AI system should also warn of the limitations that this system has with respect to certain people with certain disabilities.

(9) The right to effective judicial protection and to an impartial judge. Article 47 of the Charter

Rights and guarantees in the judicial field are the most compromised by the use of high-risk AI systems due to their relevance. Among other guarantees, this right recognises in a generic way access to justice and the right of every person to have their claims taken into account by the courts. All this under conditions of equality.

Possible violation. Annex III. 8(a) of the Rules of Procedure

A possible violation of the right to effective judicial protection by an artificial intelligence system could happen if a judge uses an artificial intelligence system to apply it to a specific court case where the system has to interpret a series of facts and propose a solution to the case. If the system is not transparent enough to facilitate the subsequent reasoning of the judicial decision to be adopted by the judge, it is possible that such a violation exists.

Possible violation. Annex III. 8(a) of the Rules of Procedure

A possible violation of the right to effective judicial protection by an artificial intelligence system could happen: this system is introduced into the Administration of Justice for use by all judges in a country whose objective is to interpret a series of facts and propose a solution to the case. If the system has been trained with very old court rulings or rulings that only collect rulings from some provinces or regions and do not take into account the set of rulings from across the country, it may not adequately generalize the environment where this system will later be used.

(10) The rights of the defence and the presumption of innocence: Article 28 of the Charter.

As with the rights mentioned in the previous paragraph, the rights of defence and presumption of innocence can be clearly affected by the use of high-risk AI systems. The infringement of these rights entails significant risk for the judicial and police system itself.

Possible violation. Annex III. 6(a) of the Rules of Procedure

Any AI system that aims to predict that a person is at high risk of committing a criminal offence seriously affects the rights of defence and presumption of innocence. These rights

may be violated to the extent that the implementation of such an AI system is used by the competent authorities and sufficient accountability measures have not been foreseen, as well as a certain degree of transparency regarding the decision taken.

(11) The right to good administration: Article 41 of the Charter.

To a large extent, many of the rights and guarantees of due process are integrated into the right to good administration for the sphere of public action.

Possible violation. Annex III. 3. a) and Annex III. 5(a) of the Regulations

A possible violation of the right to good administration would occur, for example, if the use of high-risk artificial intelligence systems by Public Administrations does not allow a natural language motivation of the decisions on which it is based and, therefore, the adequate control of the administrative activity based on these decisions. This may be due to a lack of transparency or because there are no adequate accountability mechanisms.

There may be certain variables that during the design of an AI system a priori do not affect certain groups, but once the system is put into operation, this impact is appreciated. This has happened in the United States with a system that assesses a prisoner's risk of committing crimes again and, depending on the result, obtaining benefits to get out of prison early[6]

(12) The right to good administration: Article 41 of the Charter.

To a large extent, many of the rights and guarantees of due process are integrated into the right to good administration for the scope of public action [11].

On the one hand, it constitutes a duty and requirement of the Public Administrations that must be present in their actions. So, they must act with due diligence and in time.

On the other hand, a whole series of rights in favour of citizens derive from this principle (hearing, resolution within the deadline, motivation, effective and equitable treatment of cases, good faith) that must be effectively and diligently enshrined.

Possible violation. Annex III. 5(a) of the Regulations

A possible violation of the right to good administration would occur if the use of high-risk artificial intelligence systems by Public Administrations does not allow a motivation, when required by law, in natural language of the decisions on which it is based and, therefore, the adequate control of the administrative activity based on these decisions. This may be due to a lack of transparency or because there are no adequate accountability mechanisms. In the U.S., *K.W. V. Armstrong* 89 F.3D 962, 976 (9TH Cir. 2015) revised the *Department of Health and Welfare*'s algorithmic system that provided social security benefits to people with special needs in the State of Idaho. In the procedure carried out by this Public Administration to budget the public aid, there was not a minimum motivation on the

[6] More information at: https://www.npr.org/2022/04/19/1093538706/justice-department-works-to-curb-racial-bias-in-deciding-whos-released-from-pris

reasons or causes why the system had reached such a budget, specifically when it was reduced.[7]

13) Rights of minors

According to Recital 28 of the Regulation, when assessing the impact that a high-risk AI system may have on children, the rights of children as provided for in Article 24 of the EU Charter of Fundamental Rights [11] and General Comment No 25 of the Committee on the Rights of the Child on children's rights in relation to the environment should be taken into account digital.

Possible violation. Annex III. 3. a) of the Regulation

A possible violation of the rights of minors may be present when an artificial intelligence system is used to determine admission to an educational centre and certain variables to predict a result cannot be entered because minors do not have this data. For example, in the United Kingdom, due to the coronavirus pandemic, the public authorities decided not to carry out the exams. Specifically, the qualifications that mark the end of compulsory school education, which is required for most access to university courses. Those ratings were awarded based on a weighted grade of those students' teachers and an algorithm's predictions. Among other problems, it was detected that some of the variables that had to be taken into account to favour the accuracy of the system could not be incorporated because there was no certain information about various students, for example, a list of previous grades[8].

(14) The fundamental right to a high level of environmental protection. Article 37 of the Charter

AI systems, as well as high-risk ones, must respect this right. As stated in the Spanish Charter of Digital Rights (Article XXII), "*The development of technology and digital environments must pursue environmental sustainability and commitment to future generations, and that is why the public authorities will promote policies aimed at achieving these objectives with particular attention to sustainability, durability, repairability and backward compatibility of devices and systems avoiding policies of integral replacement and planned obsolescence*." Similarly, "energy efficiency in the digital environment, favouring the minimisation of energy consumption and the use of renewable and clean energies" must be promoted. A system of risk assessment of the system must take these elements into account since a very evident lack of knowledge of them could lead to the violation of this right.

Possible violation. Annex III. 2.a) of the Regulation

A possible violation of the right to environmental protection can occur when an AI system that is used as a safety component of a gas supply network has a critical error that endangers said installation, generating a significant leakage of these resources. The impact

[7] More information at: https://casetext.com/case/kw-ex-rel-dw-v-armstrong-5?q=

[8] More information at:

https://blog.container-solutions.com/what-can-we-learn-from-the-ofqual-algorithm-debacle

https://www.engadget.com/uk-algorithm-a-levels-gcse-results-143503870.html

on the environment would come both from the loss of this resource, which is not infinite, and from the impacts that the presence of this resource can generate on the environment.

8.2 ANNEX B - Most common components of AI systems

This section incorporates some of the most common components of AI systems categorized into the following groups [4]:

1. Main actors:
    a. **Data owner:** They are responsible for the organization of the data, they are in charge of its definition, classification, protection, use and quality.
    b. **System owner:** He is responsible in the organization that requests the study of an AI solution, he is responsible for the AI system.
    c. **Data scientists:** Professionals who apply statistics, machine learning, and analytical approaches to analyse different datasets of different sizes and shapes and solve complex and critical problems.
    d. **Data engineers:** Professionals who prepare computational infrastructure and focus on the design, management, and optimization of data flow.
    e. **End users:** Those within an organization who use and benefit from the results provided by the AI system.
    f. **Data Provider:** Third parties that provide data for use in the development of the AI system.
    g. **Cloud Provider:** Third parties that offer computing platforms, and even in some cases tend to offer some data analytics or "machine learning as a service" capabilities.
    h. **System Provider:** See definition in global glossary (AI Act definition).
    i. **Responsible for the deployment:** See definition in global glossary (AI Act definition).
2. Facts:
    a. **Raw Data:** Information collected for AI analysis purposes, possibly after cleansing, but before it is transformed or analysed in any way.
    b. **Labelled Data:** A set of scalar or multidimensional data elements labelled with one or more informational labels, to train an AI system for supervised learning.
    c. **Public Data:** Information that can be freely used, reused, and redistributed by anyone without any local, national, or international legal restrictions on access or use.
    d. **Training data:** Initial data used to develop an AI system, from which the system adapts its internal parameters to refine its rules.
    e. **Augmented data:** A set of data (usually labelled) that has been augmented by adding data produced by transformations or generative systems. In image recognition, data augmentation techniques include cropping, filling, and horizontal flipping.
    f. **Testing data:** A dataset used to provide an unbiased evaluation of an AI system matched to the training data set. The testing data is used to test the system.

g. **Validation data:** Labelled datasets, which differ from ordinary labelled data only in their use and, usually, in their circumstances of collection. To evaluate an AI system in training, validation data is used.
h. **Evaluation Data:** Evaluation data is used to evaluate the predictive quality of the trained system. The AI system evaluates predictive performance by comparing predictions in the evaluation dataset with actual values (known as ground truth) using a number of metrics.
i. **Pre-processed data:** The data pre-processed before it is fed into the AI system.
j. **Data for metrics:** The type of numbers we collect when we measure something. Metric data can be proportion scale, interval scale, whole number scale, and cardinal number.
k. **AI system parameters:** An AI system parameter is a configuration variable that is internal to the AI system and whose value can be estimated from the given data.
l. **Training parameters:** The training parameters of the AI system are quantities adjusted by the learning process by applying training algorithms based on the training data. The values of the training parameters determine the actual classification, prediction or detection function calculated by the AI system.
m. **Hyperparameters:** Hyperparameters define high-level concepts about AI systems, such as the frequency of adjustment of internal parameters by the training algorithm. They cannot be learned from input data but must be established by trial and error using AI system space search techniques.

3. <u>Environments and tools:</u>
   a. **Access control lists:** An access control list (ACL) is a table that represents what access rights each user has to a particular resource, such as a file directory or an individual file.
   b. **Cloud:** It is the on-demand availability of the resources of the computer system, especially data storage (cloud storage) and computing power, without direct active management by the user.
   c. **Communication networks:** Networks with Internet connectivity for communication purposes.
   d. **Libraries:** Prescribed programs that implement ready-to-use systems, for: scientific calculation, tabular data, time series analysis, data modelling and preprocessing, deep learning, among others.
   e. **Data ingestion platforms:** This is the platform where data ingestion is performed.
   f. **Distributed File System:** File system distribution is a method of storing and accessing files, which allows multiple users to access and share files from multiple machines, or multiple hosts, over a computer network.
   g. **Communication protocols:** The communication protocol is a system of rules that allows two or more entities in a communications system to transmit information through any type of variation of a physical quantity.
   h. **Database Management System (DBMS):** It is the software that manages the storage, retrieval, and updating of data in a computer system.

i. **Data Exploration Tools**: The tools used for data exploration. Tools such as visualization and charting are frequently used to create a more direct view of data sets than simply examining thousands of individual numbers or names.
j. **Monitoring tools:** The tools used to track the status of the system in use, so that failures, defects, or problems are alerted and improved sooner.
k. **Operating system:** Manages computer hardware, software resources, and provides common services for computer programs.
l. **Optimization techniques:** Techniques used for optimization in system tuning, such as grid search, random search, and Bayesian optimization.
m. **Machine Learning Platforms:** Provides an ecosystem of tools, libraries, and resources that support the development of machine learning applications.
n. **Processors:** A processor is the part of a computer that interprets commands and performs the processes that the user has requested.
o. **Data retention and erasure tools:** Retention tools that provide support for the implementation of secure data retention, archiving, locking, anonymization, and erasure according to defined periods.
p. **System Retention and Erasure Tools:** Retention tools that provide support for the implementation of secure system retention, archiving, locking, anonymization, and deletion based on defined time periods.

4. <u>Processes:</u>
   a. **Learning transfer:** Ability to reuse previously learned knowledge to solve new problems more quickly.
   b. **Data pre-processing:** Understanding, preparing, and cleaning the data.
   c. **Data storage:** Data can be stored locally, in a distributed file system, or in the cloud.
   d. **Data understanding:** Knowledge about data, data assets, the needs that the data will satisfy, its content, and its location.
   e. **Feature selection:** During this process, the number of dimensions or features of the input vector is reduced, identifying those that are most significant to the AI system.
   f. **Data Ingestion:** The process of transporting data from multiple sources to compose multidimensional data points. The data can be placed on a storage medium where it can be accessed, used, and analysed, or the data stream can be used directly by the AI system.
   g. **Data labelling:** This is the process of detecting and labelling data samples. The process can be manual and time-consuming and software-assisted.
   h. **Data augmentation:** Techniques used to augment the amount of data by adding slightly modified copies of existing data or newly created synthetic data from existing data. It helps reduce overfitting when training machine learning.
   i. **System tuning:** Tuning focuses on setting special parameters, often called hyperparameters. This process can be done manually or automatically by searching the system parameter space, using so-called hyperparameter optimization.
   j. **Discretization technique:** It is the process of converting a numerical attribute into a symbolic attribute by partitioning the domain of the attribute.

k. **System maintenance:** After implementation, it is necessary to monitor the accuracy of the prediction to detect possible changes or deviations from the concepts. A decline in system performance could be overcome by retraining it with recent data and then redeploying it to production.
l. **Data retention and deletion:** The process of defining and implementing data retention and deletion periods according to their type.
m. **Systems retention and deletion**: The process of defining and implementing the retention and deletion periods of systems according to their type.

8.3 ANNEX C - Common types of risk in the field of AI

Here are some of the most common types of risk in the field of AI [2] [3]:

1. **Lack of transparency:**

Transparency involves communicating an organization's activities and decisions (policies, procedures, etc.) and appropriate information about an AI system (capabilities, performance, limitations, design choices, algorithms, training data, etc.) to stakeholders. If organizations are unable to provide adequate information to stakeholders, it will have a negative impact on the trustworthiness and accountability of the organization and the AI system. This could lead, for example, to risk related to poor and ineffective identification of responsibilities. For example, in the employee promotion system, if we fail to properly establish and communicate the system's limitations, it could result in excessive trust by stakeholders in the decision-making process, potentially leading to erroneous promotion decisions, harming employees, and reducing confidence in the system.

2. **Lack of explainability:**

Explainability is the property of an AI system that the factors that influence a decision can be expressed in a way that humans can understand. If these factors cannot be explained, the validation of the AI system and trust in the system are negatively affected, since it is not clear why the system has made a decision and whether it will make the right decision in all cases. This uncertainty can lead to many risks and have a strong impact on overall objectives such as reliability and accountability and specific objectives such as safety, security, fairness, and robustness. For example, in the employee promotion system, if faced with a decision by the system to promote one employee over another, the injured employee requests a review of the decision process, we must be able to give an explanation of how the system has made the decision, for example, identifying those variables of the system that have been the most decisive in the process.

3. **Level of automation:**

AI systems can operate with different levels of automation. This level of automation can be very low, in the case where an operator controls the system, or very high, such as autonomous action systems. Depending on the specific use case, automated decisions from such systems can impact on various areas of concern, such as security or fairness. For example, in the AI system for employee promotion, if there is no person responsible for

reviewing the decisions proposed by the system and promotions are executed directly, there is a risk that if there is an error in the input data, an erroneous promotion will occur, harming other employees and losing confidence in the system.

4. **Machine learning-related risk sources:**

The behaviour of AI systems depends not only on the algorithms in use, but also on the data with which the systems are trained. There are several risks associated with the use of data. For example:

- Inadequate data quality could affect several goals such as fairness, security, and robustness.
- Data may no longer be representative of the application domain, leading to risk to business objectives.
- The collection and storage of data may incur significant ethical and legal risk.

In the example of the employee promotion system, a legal (and also ethical) risk would be, as we have analysed throughout the guide, a discriminatory promotion of some employees over others, thus failing to comply with the fundamental right to non-discrimination.

5. **System hardware issues:**

Sources of risk related to hardware problems include, for example, errors based on faulty components (short circuits, interruptions, faulty bus lines, etc.). The development of AI systems could be limited due to the different hardware capabilities of the systems in terms of processing power, memory, and the availability of dedicated AI hardware accelerators.

In an example such as employee promotion, perhaps a hardware limitation and a delay in the system's decision process due to the resulting unavailability does not pose a high risk, but on the other hand, in an AI system responsible for automatically administering insulin to a diabetic patient, a hardware failure could pose a critical risk to the patient's life.

6. **System Lifecycle Issues:**

Methods, processes, and also the inappropriate or insufficient use of an AI system throughout its lifecycle can lead to risk. For example, a faulty design process may fail to anticipate the contexts in which the AI system will be used, causing it to fail unexpectedly when used in these contexts.

In the example of the automatic insulin delivery system, we must analyse and anticipate the contexts in which the system will be used throughout its life cycle. Define the update and revision cycles to which it must undergo to mitigate possible failures in its operation.

7. **Technological preparation:**

Technology readiness indicates how mature a given technology is in a given application context.

Less mature technologies used in the development and application of AI systems can add risk that are unknown to the organization or difficult to assess.

For mature technologies, a greater variety of experience data may be available, making it easier to identify and assess risk. For example, in the employee promotion system, a risk associated with technological readiness could be related to the lack of experience in the operation of these systems, which could lead to delays and increased costs if at any given time it is necessary to update or modify the parameters of the system and the necessary knowledge is not available.

8. **Enviroment complexity**

The complexity of an AI system's environment determines the range of situations that an AI system can withstand in its operational context. One of the most relevant risks, for example, is related to the degree of understanding of the AI system's environment.

A partial understanding of the environment will result in a level of uncertainty that is a particularly relevant source of risk in the design phase of AI systems.

In the example of the employee promotion system, we must analyse the complexity of the environment and the different situations in which the system will be used to adequately identify the possible risk derived from it and mitigate possible failures in its operation to the greatest extent.

9. **Other potential sources of risk:**
    a. Difficulty in identifying responsibilities and accountability.
    b. Improper, incorrect or fraudulent use of the AI system.
    c. Overconfidence in AI system decisions (automation of decisions without any oversight).
    d. Security and cybersecurity threats (it is especially recommended at this point to consult the cybersecurity guide that incorporates an additional inventory of risk and threats of AI systems in the context of cybersecurity).
    e. Potential threats to people's privacy if proper data governance is not developed (it is recommended to consult the data governance guide).
    f. Potential unwanted data disturbances or tampering (refer to the data governance guide and cybersecurity guide recommended).
    g. Other misuses resulting from poor AI system specification.
    h. Potential biases in the data can also pose a threat to the proper use of the AI system (it is recommended to consult the data governance guide).

8.4 ANNEX D - Examples of controls in the field of AI

Important note: *As indicated in section 4.5 of the guide, this Annex provides examples of possible controls to serve as a reference for defining control measures [2]. In this sense, the reader is not expected to implement all the control measures listed or to limit themselves only to these, this process will depend on the context and the risk identified, analysed and evaluated throughout the development of the risk management system.*

1. **Government measures:**
    a. Establish and define professional and ethical standards in the field of information technology to develop AI systems following the reference standards.
    b. Provide documentation for deployers that includes the appropriate context and known limitations of AI systems.
    c. Establish reparation mechanisms if people are negatively affected by the decisions of the AI system.
2. **Inclusion measures:**
    a. Incorporate the contribution of IT experts throughout the lifecycle of AI systems, with technical expertise that includes in-depth knowledge of artificial intelligence, data and data computing technologies.
    b. Include users and stakeholders, as far as possible, in the AI system development process (review of specifications, participation in testing, etc.).
    c. Collect, analyse and incorporate the opinions and feedback of users and stakeholders (e.g. through surveys).
    d. Consolidate, as far as possible, development and maintenance teams with a diversity of opinions, origins and thoughts.
    e. Assess interaction bias based on feedback collected from all stakeholders.
3. **Transparency and explainability:**
    a. Implement approximation techniques for the models incorporated in the AI system (such as the diagnostic explanation technique of the interpretable local model).
    b. Implement model diagnostic methods (such as principal component analysis).
    c. Implement model mimicking or distillation methods (e.g., knowledge transfer from neural networks to decision trees).
4. **Ability to control:**
    a. Establish control points throughout the AI system development lifecycle. The number of control points should be assessed based on risk level, available resources, and specific needs. Additionally, implement a knowledge transfer mechanism.Establish mechanisms for users to report any type of incident of the AI system.
    b. Implement mechanisms for users, as far as possible, to adjust the AI system's decision.
5. **Security and cybersecurity controls:**
    a. Implement mechanisms for the detection of possible *fuzzing* attacks (automated testing technique by which invalid, random or unexpected data is introduced into a computer system) or manipulation of voice, video and gesture input.
    b. Develop mechanisms for the timely identification of malicious training data (e.g., unauthorized change detection systems, intrusion detection systems, or AI system behavioural monitoring systems).
    c. Identify users who act in an anomalous behaviour (in a coordinated and different way than ordinary), as they can be a potential source of malicious attack.

d. Implement auditing or event tracking facilities to examine the decision, training, or detection states of AI systems.

6. **Data Privacy Controls**
   a. Implement mechanisms to ensure the proper use and processing of sensitive data.
   b. Provide training to users working with sensitive data to ensure that they do so appropriately, with discretion and caution to avoid, for example, disclosing critical AI system information to acquaintances or family members.
   c. Obtain consent from individuals or groups of individuals for the processing of their data in the AI system.
   d. Empower AI systems to identify biases or unwanted deviations.
7. **Data controls and measures**
   a. Appropriately select the AI system's training data and test sets so that they are consistent with the environment they intend to represent.
   b. Identify sensitive attributes that are required for proper AI system behaviour.
   c. Implement control and validation measures for datasets (e.g., data retention methods, K-fold cross validation, leave-one-out of cross validation (LOO), jackknife resampling, stratified sampling).
8. **Model Design and Deployment Performance**
   a. Develop a detailed specification of the design, functionalities, and implementation of the AI system.
   b. Detail the "feature engineering" techniques used, such as the selection, extraction and/or regulation of features.
   c. Specify the model, algorithms, hyperparameters, and topology by scenario used.
   d. Develop an API to verify the performance metrics of the AI system.
   e. Develop a counterfactual analysis to better understand the behaviour of the system and avoid undesired results.
9. **Robustness, reliability and resilience**
   a. Implement adversarial training techniques, such as injecting adversarial examples (generated by different attack strategies) into the training data).
   b. Implement mechanisms to reconstruct damaged inputs, i.e. remove noise (e.g. denoising autoencoder).
   c. Add a critical system to detect when the AI system is overconfident despite noisy input.
   d. Develop tests in simulated environments and/or field tests and do so periodically even after the deployment of the AI system.

8.5 ANNEX E – Examples of Effectiveness Indicators

8.5.1 ANNEX E.I – In relation to risk management measures

Measures to be evaluated	Effectiveness indicators
Analysis and definition of the internal and external context.	• Number of factors evaluated and considered in the development of the risk management system
Definition and updating of risk appetite.	• Elements considered in the definition • Defined refresh rate
Inventory of AI system components.	• Number of AI system components identified and inventoried
Identification of the main sources of risk.	• Number of sources of risk identified
Assessment of sources of risk.	• Number of sources of risk assessed
Identification and analysis of the impact of the effects and their probability of occurrence.	• % of risk categorized according to their impact and probability of occurrence
Reporting of the results.	• Maturity level of the defined results reporting methodology
Analysis of data collected post-marketing.	• Implemented methodology for collecting post-marketing information • Defined frequency of analysis of the information collected
Identification of new possible risk.	• Number of risk identified in the post-marketing process • % of these risk incorporated into the risk management system (with corresponding analysis, evaluation and implementation of controls)
Definition and selection of options for the treatment of risk.	• Number of treatment options assessed and evaluated • Distribution of the treatment measures applied according to the type of option
Plan and implement the treatment of the risk.	• Maturity level of the defined treatment plan • Number of stakeholders consulted for its preparation
Evaluate the effectiveness of each treatment.	• % of risk that remain acceptable (risk below the defined appetite) after the application of the treatments

Determine if the residual risk is acceptable.	• % of risk that have needed additional measures • Number of additional measures required by each of these risk
Document and report residual risk.	• % of risk remaining as residual • % of residual risk that have been adequately documented and reported

8.5.2 ANNEX E.II – In relation to the controls in Annex D

1. **Government measures**

Measures to be evaluated	Effectiveness indicators
Establish and define professional and ethical standards for the use of AI following the reference standards.	• Number of professional norms or standards established and applied
Provide documentation for users that includes the appropriate context and known limitations of AI systems.	• % of users who have read and applied the documentation
Define redress mechanisms if people are negatively affected by the decisions of the AI system.	• Number of defined and implemented mechanisms

2. **Inclusion measures**

Measures to be evaluated	Effectiveness indicators
Incorporate the contribution of different experts in the field of AI for the development of AI systems and throughout the entire life cycle.	• Number of experts involved
Include those responsible for the deployment and stakeholders, as far as possible, in the development process of the AI system (review of specifications, participation in testing, etc.).	• Number of deployment managers and Stakeholders Involved in Testing
Collect, analyse and incorporate the opinions and *feedback* of users and stakeholders (e.g. through surveys).	• Number of stakeholder reviews collected
Consolidate, as much as possible, development and maintenance teams	• Degree of diversity in team composition

with a diversity of opinions, backgrounds, and thoughts	
Assess the bias of the interaction following feedback collected from all stakeholders.	• Number of identified bias elements

3. Transparency and explainability

Measures to be evaluated	Effectiveness indicators
Implement approximation techniques for the models incorporated in the AI system (such as the diagnostic explanation technique of the interpretable local model).	• Accuracy of the Implemented Explanation Model • Number of different techniques implemented
Implement model diagnostic methods (such as principal component analysis).	• Number of diagnostic methods implemented
Implement model mimicking or distillation methods (e.g., knowledge transfer from neural networks to decision trees).	• Accuracy of the imitation model implemented • Number of different techniques implemented

4. Ability to control

Measures to be evaluated	Effectiveness indicators
Establish checkpoints in the AI system development lifecycle and a knowledge transfer mechanism.	• Number of checkpoints implemented.
Establish communication mechanisms (e.g., enabling an email address) for those responsible for the deployment to report any type of AI system incident.	• Number of reactions from deployants reported
Implement mechanisms for those responsible for the deployment, as far as possible, to adjust the decision system of the AI system.	• Number of adjustments addressed by deployers

5. Security and cybersecurity controls

Measures to be evaluated	Effectiveness indicators
Implement mechanisms for the detection of possible fuzzing attacks (automated testing technique by which invalid, random or unexpected data is introduced into a computer system) or manipulation of voice, video and gesture input.	• Number of fuzzing attacks detected at different input points.
Develop mechanisms for the timely identification of malicious training data (e.g., unauthorized change detection systems, intrusion detection systems, or AI system behavioural monitoring systems).	• Number of detections of malicious training data.
Identify users who act in an anomalous way (in a coordinated and different way than ordinary), as they can be a potential source of malicious attack.	• Number of malicious users detected
Implement auditing or event tracking facilities to examine the decision, training, or detection states of AI systems.	• Number of auditing or event tracing facilities deployed

6. Data Privacy Controls

Measures to be evaluated	Effectiveness indicators
Implement mechanisms to ensure the proper use and processing of sensitive data.	• Number of sensitive attributes identified, and measures implemented.
Provide training to users working with sensitive data to ensure that they do so appropriately, with discretion and caution to avoid, for example, disclosing critical AI system information to acquaintances or family members.	• % of workers aware of the discretion and precaution requirements established
Obtain consent from individuals or groups of individuals for the processing of their data in the AI system.	• % of data processing with approved consent
Empower AI systems to identify biases or unwanted deviations.	• Number of bias sources identified

7. Data controls and measures

Measures to be evaluated	Effectiveness indicators
Appropriately select the AI system's training data and test sets so that they are consistent with the environment they intend to represent.	• Qualitative observation that the data represent the environment and reality that it intends to reproduce
Identify sensitive attributes that are required for proper AI system behaviour.	• Number of sensitive attributes identified
Implement control and validation measures for datasets (e.g., data retention methods, K-fold cross validation, leave-one-out of cross validation (LOO), jackknife resampling, stratified sampling).	• Qualitative observation of each of the data validation measures

8. Model Design and Deployment Performance

Measures to be evaluated	Effectiveness indicators
Develop a detailed specification of the design, functionalities, and implementation of the AI system.	• Degree of understanding and knowledge of the design, functionalities and operation by those responsible for the deployment.
Detail the "feature engineering" techniques used, such as the selection, extraction and/or regulation of features.	• Number of documented implemented techniques
Specify the model, algorithms, hyperparameters, and topology by scenario used.	• Maturity level of model documentation, algorithms, hyperparameters, and topology by scenario used.
Develop an API to verify the performance metrics of the AI system.	• % of deployment managers who know and use the tools to verify performance
Develop a counterfactual analysis to better understand the behaviour of the system and avoid undesired results.	• Number of future errors predicted and fixed

9. Robustness, reliability and resilience

Measures to be evaluated	Effectiveness indicators
Implement adversary training techniques, such as injecting contradictory examples (generated by different attack strategies) into the training data).	• Number of techniques implemented • Qualitative analysis of the adequacy of the measures implemented to ensure the

	robustness, reliability and resilience of the model
Implement mechanisms to reconstruct damaged inputs, i.e. remove noise (e.g. denoising autoencoder).	Number of reconstructed damaged inputs
Add a critical system to detect when the AI system is overconfident despite noisy input.	• Number of System Overconfidence Detections Identified
Develop tests in simulated environments and/or field tests and do so periodically even after the deployment of the AI system.	• Number of tests deployed

8.6 ANNEX F – Glossary of Terms

This guide has been developed with an approach that tries to explain each concept present in the guide when it is exposed, however, certain specific terms have been collected in this section as additional clarification:

1. **Threat:** dangers to which the system is exposed that can end up materializing in a risk. Threats to a system primarily originate from external attacks (e.g., cyberattacks), non-compliance with security policies (e.g., connecting unauthorized devices to the network or using weak passwords), and unexpected events (e.g., fires or physical thefts).
2. **Vulnerability:** weakness of a system that allows it to be attacked and receive damage. Vulnerabilities are commonly caused by low protection against external attacks.
3. **Risk:** The likelihood that a system will suffer an incident and that a threat will materialize causing damage. Risk is therefore the probability that the threat will materialize by exploiting an existing vulnerability.
4. **Control measures:** In the context of risk management, these are the measures that need to be taken to protect the system from threats, making it less vulnerable and reducing the likelihood that the risk will materialize or instead reducing the impact it would have on my system.
5. **Risk appetite:** the level of risk an organization is willing to accept in pursuit of its mission.
6. **Inherent risk:** it is the intrinsic risk of each activity, without considering the control measures that may be implemented.
7. **Residual Risk:** is that risk that remains, after having implemented controls.
8. **Fuzzing attack:** Automated testing technique by which invalid, random or unexpected data is introduced into a computer system.
9. **HRAIS:** High-risk systems based on Artificial Intelligence are those with a significant impact on people's lives and their fundamental rights.

These systems are used in critical areas such as biometrics, education, employment, law enforcement, critical infrastructure management, and other sectors where their misuse could cause considerable damage.
The AI Act in Article 6 and Annex III sets out the requirements for development, implementation, and monitoring of these systems to ensure their safety and reliability.

8.7 ANNEX G – AI Risk Management Policy

The purpose of this annex, as indicated in section 4.9, is to present a proposal for an Artificial Intelligence (AI) Risk Management Policy that serves as a practical reference for organisations in the process of defining and developing their own risk management framework. In this regard, the content described here seeks to offer general guide on the elements and principles that a policy of this nature should consider.

This policy is not intended to be adopted verbatim or in its entirety by the reader, but rather to serve as an example of structure and content that can be adapted to the context, nature and level of maturity of each organisation. Its application should be tailored to the specific risks identified, analysed and assessed throughout the AI systems risk management process, in accordance with the needs and particularities of each case.

1. Summary

This policy establishes a comprehensive framework for risk management and governance of Artificial Intelligence (AI) systems in the organisation, aligned with the European Regulation on Artificial Intelligence (EU 2024/1689) and the recommendations of ISO/IEC 23894. Its purpose is to provide the organisation with clear guide on how to manage AI systems ethically, safely and responsibly, especially those considered high risk. The policy defines the objectives and goals that guide the identification, assessment, mitigation and monitoring of risks, ensuring that decisions about AI are made in a transparent and consistent manner.

The audience to which it applies is established, including all personnel involved in any phase of the AI systems life cycle, as well as external suppliers and collaborators who may influence their performance or regulatory compliance. Key concepts are also included, providing a common language on high-risk systems, operational, ethical and legal risks, regulatory compliance and the systems life cycle.

The policy assigns specific roles and responsibilities to each area, ensuring that all actors are aware of their role in risk management and oversight of AI systems. It defines the risk strategy, which covers internal and third-party risk assessment, residual risk identification and ongoing monitoring, consolidating a robust, consistent and aligned governance framework that is consistent and aligned with European regulations.

2. AI usage policy

The AI usage policy complements risk management by providing practical guides for the responsible use of Artificial Intelligence systems within the organisation. It establishes training and awareness programmes for all staff involved with AI systems, ensuring that they understand the operational, legal and ethical risks and promoting informed and responsible use of the technology.

Internal usage guides are defined, indicating acceptable uses of AI and the restrictions necessary to protect security, privacy and fundamental rights. In addition, pre-approval processes are included, so that any new system or significant modification undergoes a formal risk review prior to implementation. Finally, the policy establishes an inventory of AI systems, maintaining an up-to-date record of all systems used, those responsible for them, their level of risk and compliance status, ensuring traceability, control and continuous monitoring throughout the organisation.

9. References, Standards and Norms

For the development of this guide, the following norms and standards have been consulted and used:

[1] ISO 31000:2018 - Risk management – Guideline
[2] ISO/IEC 23894 - Information technology – Artificial intelligence – Guidance on risk management
[3] ISO/IEC 42001 Information technology – Artificial intelligence – Management system
[4] ENISA Report – AI Cybersecurity Challenges - Threat Landscape
[5] ISO/IEC 27001:2022 - Information security, cybersecurity, and privacy protection – Information security management systems – Requirements
[6] ISO/IEC 22989:2022 - Information technology – Artificial intelligence – Artificial intelligence concepts and terminology
[7] ISO/IEC 23053:2022 - Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)
[8] ISO/IEC 5259-1 - Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 1: Overview, terminology, and examples
[9] NIST - AI Risk Management Framework
[10] ENISA Report - Securing Machine Learning Algorithms
[11] Charter of Fundamental Rights of the European Union (2000)
[12] prEN 18228 AI Risk Managment

This guide is based on Regulation 2024/1689 of the European Parliament and of the Council of 13 June 2024 (European Regulation on Artificial Intelligence)

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