8

Guide 8. Transparency and User Information Provision

European
Artificial Intelligence Act

Companies developing compliance with

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 that are applicable, stand out the standard ***'prEN 18229-1 AI Trustworthiness Framework – Part 1: Logging, Transparency and Human Oversight',*** which is currently under development.

Revision date: 10 December 2025

General content

Detailed Index

1. Preamble

1.1. Objective of the document

The **European Regulation on Artificial Intelligence (*AI Act*)** dedicates its entire **article 13** to Transparency.

This document provides implementation measures for providers and users of AI systems **to facilitate compliance with** the obligations expressed in said article, dedicated to transparency.

1.2. How to read this guide?

The structure of this guide has a **first section** with the preamble

A **second** introductory section where what is meant by "transparency" is defined and its main characteristics are mentioned.

A **third section** focuses on the AI Act and the articles around the transparency requirement. A table with each of these articles and their references is also included within the sections of this guide to facilitate their location.

The **fourth section** delves into the transparency requirement according to the AI Act, **going through** all the sections of said article in order, answering the fundamental questions necessary to **facilitate compliance with the** obligations expressed in these sections.

The **fifth section** describes the necessary measures to be applied in order to comply with the principle of transparency. In each subsection, a brief introduction to the measure is made, it is identified to whom it applies, an example of a use case is opposed and it is related to the point or points of the articles of the AI Act to which it responds.

The **sixth section** covers the technical documentation related to the transparency of AI systems and the **seventh section** refers to the self-assessment to be carried out by AI systems.

Finally, the **eighth and ninth sections** include, respectively, a glossary of terms and references to norms and standards that have been consulted for the preparation of this guide.

1.3. Who is it for?

It is the responsibility of the **providers and deployers** of high-risk AI systems to implement appropriate measures to ensure the records retention and maintenance obligations mentioned in the AI Act. More specifically, the document is mainly aimed at:

- The technicians of the supplier entity in charge of making technical design decisions that will allow these requirements and conceptual design of the system to be met.
- Managers of the provider entity who conceptually design the AI system in accordance with the requirements of the user entity, who may consider the measures described in this document to create a transparent AI system based on the requirements described in the AI Act.
- The heads of the user entity, who must be aware of the transparency requirements that they will have depending on the use case and process that the AI system will support.

Throughout the document, language is used that is understandable by all of them, minimizing the technicalities necessary for its understanding.

1.4. Use cases and examples throughout the guide

In order **to facilitate** the **understanding of the guide,** different examples **are incorporated into it** that are intended to serve as **a reference** for the adequacy of the HRAIS for the **generation of records** in accordance with the requirements of the AI Act.

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

Finally, it should be noted that whenever an example is given, it will be done in an illustrative way. Provider and deployer should consider implementing all measures outlined in this guide, as appropriate. Each AI system, following the guide in this guide, should identify and implement the most appropriate measures according to the characteristics of its AI system and its specific purpose. 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 typology. Each organization must, in accordance with this guide, establish the appropriate measures for its type of AI system and its intended purpose.

The **selected examples** to be addressed in this Guide are:

- **Aid granting automatic system**
- **Chronic Disease Management - Smart Insulin Pump**

2. Introduction

2.1. What is transparency in Artificial Intelligence?

The concept of Transparency in Artificial Intelligence is defined as the quality of an AI system that can be fully **interpretable and understandable** by all the people who create it and interact with it **throughout its life cycle**. That is, from the moment it is conceptualized and designed, during its implementation, and when it is finally in operation. In this way it is possible to understand the reason for their reasoning and thus control it properly.

In order for this understanding and transparency to be achievable through design and development techniques, it is **sometimes necessary to resort to Explainability techniques**. This is the reason why the concept of Transparency is linked to concepts such as Explainability and Interpretability, these concepts being used in an integrated way in many of the approaches of the main providers of AI systems.

2.2. Why the need for transparency?

Artificial Intelligence is one of the most complex software technologies of all those we have assimilated. This is because their capacities (prediction making, decision-making and even emulation of cognitive abilities of the human being to make such predictions and decision-making), are similar to those of the human being. **The mechanisms necessary to massively automate these capabilities** through AI systems entail a **complexity that needs to be explained transparently to generate trust due to the criticality** of the predictions it can make or decisions it can make. This Transparency must be approached from two aspects:

- So that those responsible for the creation, operation and operation of the AI system can **understand how it works.**
- To enable end-users or deployers to **interpret the results** of the system, since the persons concerned have the right to know on the basis of which criteria and information a certain prediction has been made or a decision made on them, and under what circumstances such action could have been different if some of the information considered was modified.

In conclusion, any prediction or decision is difficult to have absolute consensus regardless of the context in which it occurs, but **detailing it as transparently as possible can at least build trust**, even in situations of disagreement. Therefore, **transparency is the basis on which trust in Artificial Intelligence is built**.

In this document, for each section of the article of the European AI Act (dedicated to transparency), measures are set out to provide Transparency to AI systems in accordance with the requirement set out in that section. **Each of these measures is exemplified in detail by the use cases described at the beginning of the document.**

To introduce the concept of transparency and the importance of its need, two simple questions are used through the use case of granting aid:

- What would happen if the entity providing the system did not provide the necessary mechanisms so that the person responsible for granting the aid in the user entity could

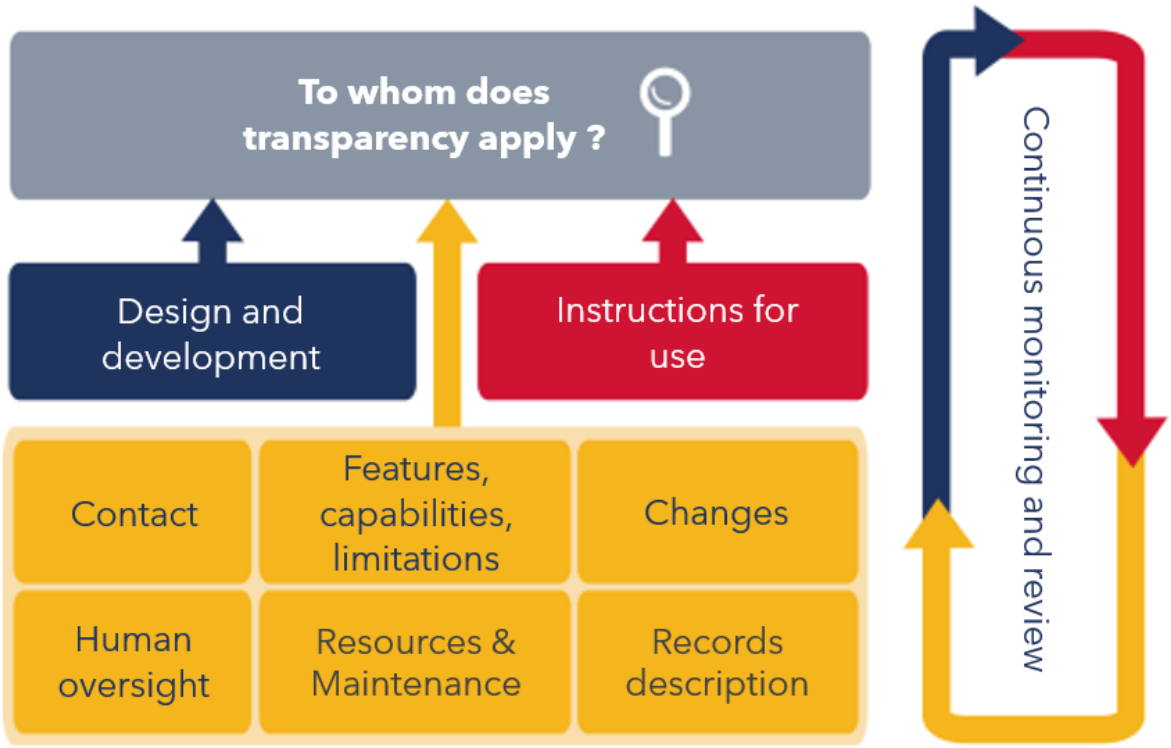
know on the basis of which specific criteria each of the grants has been granted or denied?

- What would happen if families who are denied aid received a brief "NO" as an answer, without detailing the reason for this refusal, or in what circumstances could they be given the aid?

The answer to both questions is the same: the AI system would not be transparent, and it could not be trusted to be integrated into a process as important as the granting of financial aid for families without resources.

These issues are especially relevant in AI systems compared to a "traditional" (programmed) computer system. The reason is that in a programmed computer system these responses have a response that can always be fully explained, since it is contained in the computer program (well-known technologies with a long history and of a stable nature with changes always linked to the intervention of a human programmer, not subject, for example, to learning processes that can change their behaviour). While in an AI system these transparency requirements are much more complex to manage and achieve for the reasons indicated above. It is therefore essential to be aware of this complexity and to take the necessary measures to achieve transparency.

A way to briefly visualize how to handle transparency as defined in the AI Act on AI systems is:



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

This section includes the articles referring to the generation of records 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. Previous analysis and relationship of the articles

The obligations on the generation of transparency are mainly found in an article of the European AI Act, Article 13 *"Transparency and provision of information to* deployers*"*.

To summarize, Article 13 of the European AI Act is divided into three sections:

- The **first section** sets out the general obligation to **design and develop** AI systems enabling users to understand and use the system appropriately.
- The **second section** sets out the need to provide **instructions for use** that include concise, complete, correct and clear information that is relevant, accessible and understandable to users.
- The **third section** specifies a set of **specific information** on which the global objectives set by the first two must be applied.

Each of the sections/subsections of the aforementioned Article 13 is detailed below, indicating for each of them the measures necessary to address the requirements.

3.2. Content of the articles in the AI Act

AI Act

Art.13 – Transparency and provision of information to deployers

1. High-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately. An appropriate type and degree of transparency shall be ensured with a view to achieving compliance with the relevant obligations of the provider and deployer set out in Section 3.

2. High-risk AI systems shall be accompanied by instructions for use in an appropriate digital format or otherwise that include concise, complete, correct

and clear information that is relevant, accessible and comprehensible to deployers.

3. The instructions for use shall contain at least the following information:

(a) the identity and the contact details of the provider and, where applicable, of its authorised representative;

(b) the characteristics, capabilities and limitations of performance of the high-risk AI system, including:

(i) its intended purpose;

(ii) the level of accuracy, including its metrics, robustness and cybersecurity referred to in Article 15 against which the high-risk AI system has been tested and validated and which can be expected, and any known and foreseeable circumstances that may have an impact on that expected level of accuracy, robustness and cybersecurity;

(iii) any known or foreseeable circumstance, related to the use of the high-risk AI system in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, which may lead to risks to the health and safety or fundamental rights referred to in Article 9(2);

(iv) where applicable, the technical capabilities and characteristics of the high-risk AI system to provide information that is relevant to explain its output;

(v) when appropriate, its performance regarding specific persons or groups of persons on which the system is intended to be used;

(vi) when appropriate, specifications for the input data, or any other relevant information in terms of the training, validation and testing data sets used, taking into account the intended purpose of the high-risk AI system;

(vii) where applicable, information to enable deployers to interpret the output of the high-risk AI system and use it appropriately;

(c) the changes to the high-risk AI system and its performance which have been pre-determined by the provider at the moment of the initial conformity assessment, if any;

(d) the human oversight measures referred to in Article 14, including the technical measures put in place to facilitate the interpretation of the outputs of the high-risk AI systems by the deployers;

(e) the computational and hardware resources needed, the expected lifetime of the high-risk AI system and any necessary maintenance and care measures, including their frequency, to ensure the proper functioning of that AI system, including as regards software updates;

(f) where relevant, a description of the mechanisms included within the high-risk AI system that allows deployers to properly collect, store and interpret the logs in accordance with Article 12.

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 the articles:

Article	AI Act requirement	Section
13.1	Design and development of high-risk AI systems	Section 4.1
13.2	Instructions for use for high-risk AI systems	Section 4.2
13.3	Specific information on instructions for use	Section 4.3
13.3.a	Contact	Section 4.3.1
12.3.b	Features, Capabilities, and Limitations	Section 4.3.2
12.3.b.i	Purpose	Section 4.3.2.1
13.3.b.ii	Level of accuracy	Section 4.3.2.2
13.3.b.iii	Risks by Intended Use	Section 4.3.2.3
13.3.b.iv	Specifications	Section 4.3.2.4
13.3.b.v	Impact on people	Section 4.3.2.5
13.3.b.vi	Input data	Section 4.3.2.6
13.3.b.vii	Output Information	Section 4.3.2.7
13.3.c	Changes	Section 4.3.8
13.3.d	Human oversight	Section 4.3.4
13.3.e	HW/SW Resources & Maintenance	Section 4.3.5
13.3.f	Records files	Section 4.3.6

4. Transparency requirements in the AI Act

4.1. Section 1. Design and development

AI Act

Art.13.1 - Transparency and provision of information to deployers

High-risk AI systems **shall be designed and developed** in such a way as to ensure that their operation is **sufficiently transparent** to enable deployers to **interpret a system's output and use it appropriately**. An appropriate type and degree of transparency shall be ensured with a view to achieving **compliance with the relevant obligations of the provider and deployer set out in Section 3.**

What we understand

That the AI system is designed and developed in general through measures that allow it to provide information about its operation in a transparent manner and that in this way users understand and use it appropriately, particularly in the obligations provided for in the third section of this article.

This objective of Transparency is aligned with the definition reflected in ISO 23894, section A.7 (*transparency and explainability*).

Measures to carry it out

Below are measures to be considered in the design and development of AI systems so that they work transparently (click to access the details):

- Attend to the domain of functionality
- Ensure the functional objective of the system
- Transparency of the data used
- Detailing from the most global to the most particular
- Adapting the language
- Manage complexity
- Use built-in metrics in the system lifecycle
- Apply prudence
- Using causation and minimizing correlations
- Use counter factuality

Since these measures must be considered in the design and development of AI systems, they can be supported by methodological tools for the construction of software and even by technologies that allow their automation.

The third section of the European AI Act details the measures that apply to each of the specific information described in that section.

4.2. Section 2. Instructions for use

AI Act

Art.13.2 - Transparency and provision of information to deployers

High-risk AI systems shall be accompanied by **instructions for use in an appropriate digital format** or otherwise that include **concise, complete, correct and clear information that is relevant, accessible and comprehensible to deployers.**

What we understand

In addition to the measures to be considered in the design and development of the system so that it can provide information about its operation in a transparent manner (first section detailed above), this second section indicates the need for a digital or other medium that collects the instructions for use of the system in a transparent way. *Concise, complete, correct and clear that is relevant, accessible and understandable for users.*". It is in the third section where the set of information that these instructions must contain is specified.

Measures to carry it out

- Enable a channel with system usage information

4.3. Section 3. Specific information

This section specifies **seven specific pieces of information** on which the measures applicable in the first two paragraphs must be applied.

AI Act

Art.13.3 - Transparency and provision of information to deployers

The instructions for use shall contain at least the following information:

(a) The **identity and the contact details of the provider** and, where applicable, of its authorised representative;

(b) The **characteristics, capabilities and limitations of performance** of the high-risk AI system, including:

(c) The **changes** to the high-risk AI system and its **performance which have been pre-determined** by the provider at the moment of the initial conformity assessment, if any;

(d) The **human oversight measures** referred to in Article 14, **including the technical measures** put in place to facilitate the interpretation of the outputs of the high-risk AI systems by the deployers;

(e) The **computational and hardware resources needed**, the expected lifetime of the high-risk AI system and any necessary maintenance and care measures, including their frequency, to ensure the proper functioning of that AI system, including as regards software updates;

(f) Where relevant, a description of the mechanisms included within the high-risk AI system that allows deployers to **properly collect, store and interpret the logs** in accordance with Article 12.

Each of these subsections is detailed below, indicating for each of them, the measures that allow them to be carried out.

As regards those measures, AI systems have the ability to **facilitate and even automatically generate** many of these instructions for use if **good design and development practices** aimed at providing transparency are applied, as required by paragraph 1 of this Article.

In this way, these instructions are constantly **aligned and synchronized** with the *software* without parallel and additional efforts, thus allowing all users who interact with the system to understand and properly use the system.

The measures described below can be **categorized into two types**:

- **Design measures** that facilitate the achievement of transparency requirements:
  - Provide contact with the provider.
  - Attend to the domain of functionality.
  - Ensure the functional objective of the system.
  - Apply prudence.
  - Use causality, minimize correlations.
  - Enable a channel with system usage information.

- Measures that can be **directly automated.** This document does not describe the underlying technologies that exist behind these measures, but describes the functionalities of these technologies, whether they are proprietary from a provider or even *open source* for free distribution:
  - Transparency about the data used.
  - Detailing from the most global to the most particular.
  - Adapting the language.
  - Use built-in metrics in the system lifecycle.
  - Use counter factuality.

4.3.1. Section 3a. Contact

AI Act

Art.13.3a - Transparency and provision of information to deployers

the **identity and the contact details of the provider** and, where applicable, of its authorised representative;

What we understand

Given that they are high-risk AI systems, it aims to ensure the support service provided by the AI system provider to the user entity of the same through a channel that allows the user to contact in case of doubt or complaint regarding the AI system.

Measures to carry it out

- Provide contact with the provider

4.3.2. Section 3b. Features, Capabilities, and Limitations

This section specifies six aspects to be taken into account.

AI Act

Art.13.3b - Transparency and provision of information to deployers

The characteristics, capabilities and limitations of performance of the high-risk AI system, including:

(i) its intended purpose;

(ii) the level of accuracy, including its metrics, robustness and cybersecurity referred to in Article 15 against which the high-risk AI system has been tested and validated and which can be expected, and any known and foreseeable circumstances that may have an impact on that expected level of accuracy, robustness and cybersecurity;

(iii) any known or foreseeable circumstance, related to the use of the high-risk AI system in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, which may lead to risks to the health and safety or fundamental rights referred to in Article 9(2);

(iv) where applicable, the technical capabilities and characteristics of the high-risk AI system to provide information that is relevant to explain its output;

(v) when appropriate, its performance regarding specific persons or groups of persons on which the system is intended to be used;

(vi) when appropriate, specifications for the input data, or any other relevant information in terms of the training, validation and testing data sets used, taking into account the intended purpose of the high-risk AI system;

(vii) where applicable, information to enable deployers to interpret the output of the high-risk AI system and use it appropriately;

Each of the subsections is detailed below, indicating for each of them, the reason for their need, as well as the measures to carry them out.

4.3.2.1. Section 3b.i. Purpose

AI Act

Art.13.3b.i - Transparency and provision of information to deployers

its **intended purpose**,

What we understand

To ensure the functional understanding of the entire system, an objective that must be applied from its conception and design, so that during its operation this understanding is viable. If not:

- The system will not be transparent, and therefore its supervision and control will not be possible in the necessary terms.
- It can lead to speculation about how it works, which can reduce confidence in it if its operation does not meet expectations.

Measures to carry it out

- Attend to the mastery of the functionality of the system to define the necessary Transparency measures in the specific use case in which the AI system will be used.

4.3.2.2. Section 3b.ii. Level of accuracy

AI Act

Art.13.3b.ii - Transparency and provision of information to deployers

The level of accuracy, including its metrics, robustness and cybersecurity referred to in Article 15 against which the high-risk AI system has been tested and validated and which can be expected, and any known and foreseeable circumstances that may have an impact on that expected level of accuracy, robustness and cybersecurity;

What we understand

Ensure understanding of the AI system's accuracy, robustness, and cybersecurity metrics, confirming a good fit between the provider's scope of use and the user's intent for adoption.

Measures to carry it out

The following are measures to be taken into account in the design and development of the AI system to facilitate the creation of such metrics that will facilitate Transparency:

- Manage the complexity of the AI system, opting for the simplest complexity that resolves the necessary level of accuracy.
- Use accuracy metrics embedded in the AI system lifecycle.
- Use causality and minimize correlations, as excessive use of correlations can complicate the level of accuracy.
- All the measures indicated in the guides of article 15 of the European AI Act (Accuracy, robustness and cybersecurity).

4.3.2.3. Section 3b.iii. Risks by Intended Use

AI Act

Art.13.3b.iii - Transparency and provision of information to deployers

any known or foreseeable circumstance, related to the use of the high-risk AI system in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, which may lead to **risks to the health and safety or fundamental rights** referred to in Article 9(2);

What we understand

It must be ensured that the AI system is used for the intended purpose in its creation, and not other alternatives (known or potential) that could distort the use for which the system was initially designed, posing a risk to health, safety or fundamental rights.

Measures to carry it out

- Ensure the functional objective of the system
- Matters relating to the Risks Management guide (Article 9(2))

4.3.2.4. Section 3b.iv. Explaining Your Output Results

AI Act

Art.13.3b.iv - Transparency and provision of information to deployers

where applicable, **the technical capabilities and characteristics** of the high-risk AI system to provide **information that is relevant to explain** its output;

What we understand

If required, the instructions for use of the AI system should provide detailed information on the features, parameters or technical documentation that the system uses to make decisions, allowing deployers to understand and evaluate the internal logic of the model.

This requirement seeks to guarantee transparency, facilitating traceability and responsibility in the use of AI.

Measures to carry it out

- Apply transparency measures on the data used, taking into account what the output of the system will be.
- Detail from the most global to the most particular that may occur in said outing.
- Adapt language to ensure understanding of your output results.
- Apply prudence and do not reveal sensitive information at the exit of the system.
- Use counterfactuality to detail in the output what, the reason for said action or in what circumstances said action could have been different if some of the information taken into account were modified.

4.3.2.5. Section 3b.v. Impact on people

AI Act

Art.13.3b.v - Transparency and provision of information to deployers

when appropriate, its performance **regarding specific persons or groups of persons** on which the system is intended to be used;

What we understand

Ensure that there are no unfair actions when the system's actions (predictions, decisions) are people-oriented.

Steps for implementation

- Analyse the data, ensuring the fairness of the system with respect to the group of people on whom the AI system may have an influence.
- To provide mechanisms that allow analysis from the most global to the most particular, thus ensuring the analysis of groups of people and specific people.
- Use counter factuality to be able to detail the reasons taken about people and groups of people.

4.3.2.6. Section 3b.vi. Input data

AI Act

> Art.13.3b.vi - Transparency and provision of information to deployers
>
> when appropriate, specifications for the input data, or any other relevant information in terms of the training, validation and testing data sets used, taking into account the intended purpose of the high-risk AI system;

What we understand

It aims to give visibility into the nature of the data used, so that the user can understand and assess whether the data sample is fair and representative for the purpose of the system.

Steps for implementation

- Transparency about the data used

4.3.2.7. Section 3b.vii. Output Information

AI Act

> Art.13.3b.vii - Transparency and provision of information to deployers
>
> Where applicable, information to enable deployers to interpret the output of the high-risk AI system and use it appropriately;

What we understand

Ensure understanding of the system's output information.

Steps for implementation

- Apply transparency measures on the data used, taking into account what the output of the system will be.
- Detail from the most global to the most particular that may occur in said outing.
- Apply prudence and do not reveal sensitive information at the exit of the system.
- Use counterfactuality to detail in the output what, the reason for said action or in what circumstances said action could have been different if some of the information taken into account were modified.

4.3.3. Section 3c. Changes

AI Act

Art.13.3c - Transparency and provision of information to deployers

The changes to the high-risk AI system and its performance which have been pre-determined by the provider at the moment of the initial conformity assessment, if any;

What we understand

Ensure that the user entity is informed of the changes made to the system by the provider, with the implications that this may have on its behaviour and/or accuracy. This is especially important, as the initial conformity assessment establishes an OK before the system is put into operation, which must be maintained during the operation and evolution of the system. In addition, it is the case that AI systems can degrade their performance over time due, for example, to new input data received by the system (*data drift*), or even due to changes in the system (*model drift*).

Steps for implementation

- Use built-in metrics in the system lifecycle
- To communicate these changes, it will be necessary to enable a communication channel between the provider and the user.

4.3.4. Section 3d. Human oversight

AI Act

Art.13.3d - Transparency and provision of information to deployers

the human oversight measures referred to in Article 14, including the technical measures put in place to facilitate the interpretation of the outputs of the high-risk AI systems by the deployers;

What we understand

Applying transparency measures to the AI system allows us to understand how it works and interpret its results. As a consequence, and in a bidirectional way, transparency and human oversight (developed in Article 14 of the regulation) are completely related, since for the system to be supervised by people it is essential that the system is transparent. Therefore, all the measures applicable to transparency and human oversight are those that resolve this article.

Steps for implementation

All the technical measures reflected in this document, as they all facilitate the interpretation of the output information of the AI system:

- Attend to the domain of functionality
- Ensure the functional objective of the system
- Transparency of the data used
- Detailing from the most global to the most particular
- Adapting the language
- Manage complexity
- Use built-in metrics in the system lifecycle
- Apply prudence
- Using causation and minimizing correlations
- Use counterfactuality

In addition, those indicated in the guide of article 14 (Human oversight).

4.3.5. Section 3e. HW/SW Resources & Maintenance

AI Act

Art.13.3e - Transparency and provision of information to deployers

the computational and hardware resources needed, the expected lifetime of the high-risk AI system and any necessary maintenance and care measures, including their frequency, to ensure the proper functioning of that AI system, including as regards software updates

What we understand

This section is partly related to 13.3.c, which also aims to inform the user entity of the changes made to the system as part of its maintenance and evolution by the provider, with the implications that this may have on its behaviour and/or accuracy and thus guarantee its correct functioning in the face of them.

Steps for implementation

- Use a continuous integration system on the AI system and its use (including user-processed data), applying integrated metrics in the life cycle of the system that allow measuring the use, performance and impact of updates on it.

4.3.6. Section 3f. Recording logs

AI Act

Art.13.3f - Transparency and provision of information to deployers

where relevant, a description of the mechanisms included within the high-risk AI system that allows deployers to properly collect, store and interpret the logs in accordance with Article 12.

What we understand

In order to facilitate the understanding and appropriate use of the systems by users, it is a matter of describing and implementing the mechanisms contained within the high-risk AI system itself to allow users of the system itself to collect, save and correctly interpret the logs (or records), whenever they are considered relevant. It aims to define the scope of what transparent AI should be. Specifically, it mentions that system users should be able to collect, save, and interpret system logs whenever they are relevant.

Steps for implementation

- Those indicated in the guide of Article 12 of the AI Act (Records).

5. Applicable measures to achieve Transparency

This chapter includes the design and development measures that allow an AI system to be Transparent (Article 13(1)), and which are also applicable to the specific information contained in Article 13(3) so that it can be transmitted in a Transparent manner.

5.1. Provide contact with the provider

The provider's work does not end with implementation but continues after the system is in production. This relationship, common in any *software* system, is especially important in AI systems, given their complexity, and particularly in high-risk ones given the importance of the processes they support. Therefore, within their organizational structure and governance model associated with the AI system provided, providers will enable a clear point of contact for their deployers . The functions of this contact shall be to at least:

- **Proactively** monitor that the system complies with regulations, taking into account that compliance is a constant activity over time for two reasons:
  - Like any other software system, it is susceptible to errors, is subject to update policies, etc.
  - Even if there are no errors in the software, its learning process and subsequent actions (predictions, decision-making, etc.) can end up degrading due to the appearance of new scenarios.
- **Reactively** respond to requests made by the entity deploying, either for incidents detected by it, or for requests aimed at understanding the operation of the AI system. Without prejudice to the provider's existing mechanisms for the attention and resolution of incidents of the AI system.

The contact channel must allow the management of requests and incidents of the entity deploying online by recording them, their management through work *workflows* by the functional and technical profiles of the provider entity necessary for their resolution, the monitoring of all requests and incidents managed, and the historical necessary for its resolution, etc. These processes can be supported by an IT demand channel, an industry standard.

Who does it apply to?

- The provider shall provide such contact and the channel through which it will occur.
- The deployer must also identify the persons responsible within their organization in charge of activating such contact with the provider in situations such as those described above.

Example – Aid granting automatic system

Those responsible for the entity responsible for deploying detect that the amount proposed for the aid of a family is not homogeneous with respect to others provided in the past to families with similar characteristics. These managers activate contact with the provider by opening a request in the demand management system.

Example - Insulin Pump

A medical manager of the entity responsible for deploying detects through an alarm provided by the system that the next dose prescribed to a patient is not the usual one despite having similar parameters in blood. This medical manager activates contact with the provider by opening a request in the demand management system, in addition to taking the appropriate measures based on their medical criteria on the dose, and contacting the patient to understand possible medical variables of the environment not contemplated by the AI system.

To which sections does this measure apply?

- Section 3a. Contact
- Section 3c. Changes

5.2. Attend to the domain of functionality

This measure is trivial in any computer system, but it is especially important as it is the application of a technology with the peculiarities of Artificial Intelligence, since different scenarios cannot be proposed that are not identified with other types of technologies.

Therefore, during the design of the AI system, it is essential to identify what are the transparency needs applicable to the domain to which it belongs (Insurance, Finance, Health, Retail, Transportation, Industry, Public Administration, etc.), identify the key characteristics of Transparency that must be taken into account in each use case and the reason for them since, for example, The transparency needs are different for a biometric identification system than for one dedicated to the management of critical infrastructures in the industrial sector. They are even different depending on the specific use case of both typologies, both categorized as high risk by the regulation.

All this without prejudice to the specific legislation on transparency.

Example – Aid granting automatic system

In our use case, these are the main requirements that must be supported by the system according to the domain of its functionality and that, in view of transparency, must be explicitly set out in the conditions of use of the system. In addition, it must be possible to monitor during the operation of the system that these criteria are still maintained and that they do not degenerate during its operation:

- The overall criterion that is being followed to predict the risk of social exclusion, or to grant aid, For example:
  - A family that has had an income of less than a certain monthly amount for a certain period of time.
  - That this amount is weighted by the number of members and their age, giving more weight to school-age members without the ability to contribute economic resources to the family in the short/medium term.
  - That you have recurring expenses (e.g.: mortgage, medical treatments not covered by social security, etc.).
  - That this criterion weighs with more weight those families that have dependent members or with some type of disability.
  - The degree of economic development of the geography in which the place of residence is located.
  - Since the system has been trained with historical data on family structures from previous decades and from traditional families that are still in the majority today, on which predictions are made, the same treatment is ensured for families with new family structures (e.g., single-parent/parental, made up of people of the same gender, etc.).
- The system must be able to compare decisions made on a set of families with similar characteristics based on the criteria indicated above, making it possible to verify that the decisions are homogeneous in that set.
- The system must be able to detail internally, and to the applicant family, the reason why the aid is granted/denied, based on the overall criteria for the operation of the system. In the event of refusal, it must be detailed, for example, under what conditions it would be granted, or from when.

Example - Insulin Pump

These are some of the requirements that must be supported by the system according to its business domain and that **must be explicitly set out** in the conditions of use of the system. In addition, it must be possible **to monitor** during the operation of the same that these criteria are still maintained and that they do not degenerate during its operation:

- The global criteria followed to predict a trend and perform insulin delivery automatically; detailing the variables in blood and their values, the patient's medical history, environmental conditions and, as a consequence, the dose administered.
- The system must be able to compare decisions made on a set of patients with similar characteristics based on the criteria indicated above, making it possible to ensure that decisions are homogeneous in this set.
- The system must be able to detail internally and to a specific patient the reason why the dose is varied, based on the overall criteria of operation of the system described above.

Who does it apply to?

This measure applies to the **entity deploying** since it is responsible for defining the functional requirements of the AI system that will support its use case, taking into account the business domain of said case. These requirements are independent of the technology used and the provider providing it. As part of these requirements, it must identify those related to the transparency of the same.

On the other hand, the **provider** must report in detail on the functional scope of the AI system that it finally makes available to the entity deploying, so that it can ensure alignment between its needs and those supported by said system, also specifying the possible risks due to unforeseen uses such as indicated.

To which sections does this measure apply?

- Section 1. Design and development
- Section 3b.i. Purpose
- Section 3d. Human

5.3. Ensure the functional objective of the system

To ensure the functional objective of the system, and to manage the risk that it will be used for a purpose other than that for which it was designed, it will be necessary to:

- Identify the foreseeable circumstances where such uses may occur.
- Define a risks plan for such uses, proactively monitoring that such scenarios do not occur during the operation of the system (described in the risks management guide, Article 9).

Who does it apply to?

This supervision measure applies to both the provider entity and the deployer, with the contact of the provider entity and its mirror in the deployer being the roles of said supervision.

Example – Aid granting automatic system

The system for granting aid uses data from the family unit. Such data must be **treated with special vigilance**, ensuring that **its use is limited only to the functional objective of the system** (the granting of aid), since minors under eighteen years of age can be part of said family unit, and this group of people is especially detailed in Article 9 of the European Regulation on Artificial Intelligence (risks management). indicating that "*special attention will be paid to the probability that people under eighteen years of age will be affected*". The potential use of such data for a purpose other than that intended must be specially managed, and also provided for in the risks plan of the AI system.

Example - Insulin Pump

The insulin delivery system uses data that is specifically applicable to a patient based on their pathology and current condition, based on which it determines the exact dose needed for the patient to carry, as this is one of the benefits of using this system. This information is obtained from the patient's medical history and the online data that the system collects from them.

The system must inform the patient in a transparent manner that the inoculated product is the one prescribed by their responsible doctor and not another variant, that the dose incorporated into the device by the patient is the one prescribed by the device based on the online analysis of their situation carried out by the device, and that this dose is validated by the doctor.

To which sections does this measure apply?

- Section 1. Design and development
- Section 3b.iii. Risks by Intended Use
- Section 3d. Human

5.4. Transparency about the data used

It is essential to know how the AI system works. But it is also essential to understand the data that this system handles (structured or unstructured, images or natural language, for example). This action should be applied both for data for the AI system's learning and for the usage data itself by the deployer. In this way, you can:

- To be able to detail with Transparency the results provided by the AI system.
- Understand, and assess whether the data sample is fair and representative for the purpose of the AI system.

To do this, it is important to:

- List the data sources used by the system,
- Perform an exploratory data analysis (EDA) of these sources, to know their essence, their associated meta-information, their critical or atypical values, etc., understanding their meaning since, sometimes, the characteristics of a given set of data are visible at first glance, but other times they are intertwined, and this can have implications for applying transparency to the system.

Who does it apply to?

- The deployer, since he is the final owner of the data used by said system and must be aware of the meaning of his data, its usefulness and the implication in the use of the same.
- The provider, since it is the one who must provide the tools that allow the descriptive and exploratory analysis of the data.

Example – Aid granting automatic system

Enumeration of the data sources used:

- Historical information on the evolution of families who had economic deprivation and finally needed some kind of help.
- Information with the general economic context of the geography to which they belonged when this situation occurred.
- Macroeconomic data on the current economic context and forecasts for it.

Exploratory data analysis (EDA) of such sources with the aim of ensuring that data sources include, for example:

- Families from different geographical locations.
- Families who, although in significant periods have had sufficient income, have finally ended up in a situation of social exclusion (an atypical but possible situation), with data that allow us to understand how they finally reached this situation.
- Families who do not even have an income record, which can exclude them from the process as it is a critical value not contemplated.
- Meta-information not visible at first glance, such as the relationship between a family in a situation of exclusion and members of the family with dependency needs or belonging to vulnerable groups of people (e.g., due to any type of disability).
- Current data, but also historical data that allow future risk to be detected based on patterns.
- That the macroeconomic data include all the professional sectors in which any family can focus its activity.

Example - Insulin Pump

Listing of the data sources used, such as:

- Historical information of the blood parameters evaluated by the system for the administration of the dose.
- Historical information on the doses given to all patients, as well as their reaction to them.
- Information on the compounds supplied

Exploratory data analysis (EDA) of such sources with the aim of ensuring that data sources include, for example:

- Patients from different geographies with different climatic conditions.
- Patients of different ages, sex, morphological characteristics (weight, height, etc.), level of physical activity, etc.
- Cases of adverse reactions to the doses administered, with data that allow us to understand what caused this reaction.
- Data from patients with no pathologies other than diabetes, as well as data from patients with other additional pathologies.

To which sections does this measure apply?

- Section 1. Design and development
- Section 3b.v. Impact on people
- Section 3b.vi. Input data
- Section 3b.vii. Output Information
- Section 3d. Human

5.5. Detailing from the most global to the most particular

The internal information provided by the AI system must be concise and as complete as possible, therefore, it is necessary to understand the operation of said system from the most general to the most particular, enabling technical measures that allow the greatest possible degree of detail to be addressed in:

- The global reasoning mechanism of the AI system.
- Predictions and decisions made on a subset of information with similar characteristics, to ensure the homogeneity of its operation.
- And even of each of these predictions and decisions individually.

Who does it apply to?

The provider must provide the system with the technical mechanisms that facilitate this need and thus ensure the user entity complete control and understanding of the system under a language understandable by all the actors that interact with it throughout its life cycle.

To which sections does this measure apply?

- Section 1. Design and development
- Section 3d. Human

Example – Aid granting automatic system

The system shall provide mechanisms that enable the deploying entity to know:

- The overall criterion used to grant the aid (detailed in the example of the measure Attending to the domain of functionality).
- Decisions taken (assessment of the risk of social exclusion of a family, time in which it would occur, amount allocated, etc.) on a group of families with similar characteristics (for example, based on the number of people in it, income level, geographical location), allowing deployers to verify that the decisions are homogeneous in that group.
- Decisions made about a specific family, based on the overall criterion of the functioning of the system. In addition, if the aid is not granted, it will detail under what values of the criteria used it would be assigned to them (following the counter factuality criteria).

Example – Insulin Pump

The system shall provide mechanisms that enable the deploying entity to know:

- The overall criterion followed for the decision of the dose to be administered (detailed in the example of the measure Attending to the domain of functionality).
- Composition of the doses administered, on a set of patients with similar medical characteristics (e.g., sex, age, physical condition, additional pathologies, etc.), allowing physicians to verify that the decisions are homogeneous in this group in a logical order of magnitude from a medical point of view.
- Decisions made about a specific patient, based on the overall criteria of operation of the system.

5.6. Adapting the language

The AI system must be designed so that it can provide information to all the profiles that interact with it and thus transparently ensure its complete understanding.

There are many types of profiles that interact with the AI system throughout its life cycle. Therefore, technical mechanisms must be set up to show this information in a transparent and understandable way for all of them, adapting the type of language to their level of dialogue:

- **Deployers of the AI system** need to understand through natural language, that the predictions and decisions that the system is providing are homogeneous in end users with similar characteristics, etc., with a focus oriented to the business objective of the system.
- **The provider's technicians**, whose language is technical, when they are building the AI system, need to explain to the users of the entity deploying, the reasoning of their models through a language understandable by them, and thus ensure between them

that the system complies with the European AI Act and the company's policies. And throughout the life of the AI system, they must know in technical language the reasoning that the system is following to ensure its correct functioning and even improve its accuracy.

- **Natural or legal persons affected by the operation of the AI system**, which in many cases does not necessarily have to have technical language of any kind, need to know, for example, the reasons why they have not been granted a product or service and to know under what conditions it would be granted applied to their reality.

To this end, in addition to the technical mechanisms necessary for the implementation of the system by the specialists, it is required for the rest of the profiles, to provide a user interface that provides the information in an understandable way, either visually, textually in natural language or by other means. In short, a quality human-AI system interface design to reinforce the transparency of AI systems by deployers and affected without specialized technical knowledge.

Who does it apply to?

- The provider must provide deployers with the technical mechanisms and user interfaces that facilitate this need and thus ensure complete control and understanding of the system under a language understandable by all the actors that interact with it throughout its life cycle.
- The deployer will make use of these elements to meet the transparency needs for their particular business solution.

To which sections does this measure apply?

- Section 1. Design and development
- Section 3d. Human

Example – Aid granting automatic system

The system will provide a user interface, aligned with the level of dialogue of each of the profiles that interact with it, and that provides the following information in an understandable way, visually and textually in natural language, especially to those non-technical profiles:

- **The managers of the granting of aid** must receive information from the system, **in natural language**, that allows them to ensure that it is complying with the policies for granting aid, the degree of accuracy it is having when predicting the social exclusions that finally occur and that give rise to the aid, such as **the homogeneity of** the predictions it is providing in families with similar characteristics, details about all the predictions and decisions made by the AI system. Likewise, those **responsible for granting aid** will have dashboards that allow them to obtain statistics on the results.
- **The technicians** who implement the system and are responsible for monitoring it while it is in operation, must know the same information as above, but from **a technical perspective.**
- **To the applicant families,** the system must be able to explain **in natural language** the reasons why they have been granted a certain amount and not another, why the application has been denied and under what conditions applied to their reality it would be accepted, etc.

Example – Insulin Pump

The system will provide a user interface, aligned with the level of dialogue of each of the profiles that interact with it, and that provides the following information in an easily understandable way, visually and textually in natural language, especially to those non-technical profiles:

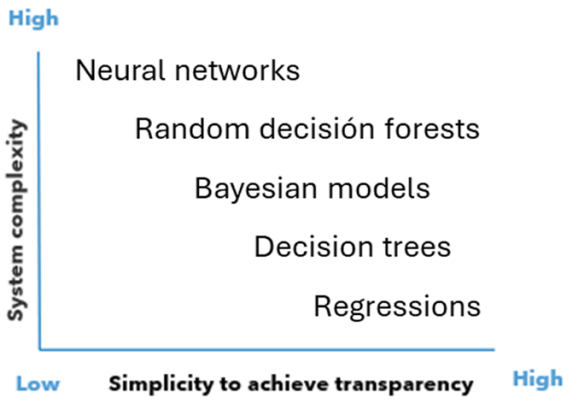
- **The physicians** responsible for administering the doses to their patients must receive information from the system, **in natural language** with the necessary technical detail from a medical point of view, which allows them to ensure that the doses administered are having the correct effect from a medical point of view, according to the desired levels of accuracy. They will also be able to have dashboards that allow them to know the evolution of the operation of the system, as well as that of the patients themselves.
- **The *software*** technicians who implement the system and monitor it while it is in operation should know the same information as above, but **from a technical perspective.**
- **The** system must be able to explain **to patients in natural language** the doses they are receiving, their effect, etc., in the same way that the doctor would explain it to them in one of their check-ups.

5.7. Manage complexity

Understanding the performance of AI systems in order to facilitate their transparency can involve the analysis of huge numbers of variables and mathematical operations. Therefore, a good design of the system understood from its simplicity can be key to achieving greater transparency, since sometimes the complexity of the AI system tends to increase to seek more accuracy, which can imply the loss of transparency about its operation. Therefore, it is necessary to measure the return on investment of complicating the model to obtain a little more accuracy at the cost of possibly losing transparency.

It is advisable to decide in the design phase on the simplest model that meets the business objectives, thus simplifying transparency and understanding of its operation. This makes it easier to identify what factors affect the system and to be able to explain it. In addition, this simplicity in the system results in lower energy consumption, which is very important in these AI systems, whose need for computing implies a high environmental impact.

The following graph represents the relationship between the technical complexity of some types of smart models vs. the simplicity to achieve transparency in them.



The special feature of black box/white box models

"White box" systems are often called "transparent," and they are the ones where it is easiest to understand their behaviour. Examples of this type of model are linear regressions and decision trees. On the other hand, "black box" models are those in which the input and output are known, but their inner workings are difficult to understand even sometimes for the technicians who implement it. Examples of "black box" systems are decision forests and neural networks.

Black box models offer high accuracy and can solve complex use cases, but **their transparency can be limited** due to their complicated construction. Their lack of transparency **may mean having to discard them** in order to resort to white-box models, which can be directly interpreted. **This design decision by the provider is especially relevant** in high-risk applications where it is necessary to ensure that their operation does not degenerate and where transparency is an unavoidable need.

Transparency in black box systems is a challenge that is being investigated. Currently, one of the possible solutions to facilitate its transparency (not 100%) involves algorithms (e.g. *profweight*) that probe the system by analysing the corresponding inputs and outputs they

produce. Once this data has been analysed, an equivalent directly interpretable system is generated that serves to detail the operation of the black box.

Who does it apply to?

- This design decision applies to the AI system provider.
- The deployer should identify whether the complexity of the system that will support their use case may become blocker to the needs of transparency about the system.

To which sections does this measure apply?

- Section 1. Design and development
- Section 3b.ii. Level of accuracy
- Section 3d. Human

Example – Aid granting automatic system and insulin pump

Decision trees and linear regressions will be used. The use of Bayesian probabilistic models or neural networks is ruled out, since the improvement of accuracy with the latter might not be relevant, for example, in a use case with a high probabilistic factor such as the granting of aid, while on the other hand they would complicate the objective of transparency of the system.

5.8. Use built-in metrics in the system lifecycle

The metrics allow monitoring the level of accuracy, robustness and cybersecurity (referred to in the guides detailing Article 15 of the regulation). Such metrics can be applied to data and system. From a formal point of view, the metrics used in data and model are developed in the standard [ISO/IEC 23053]. Its use is necessary because AI systems degrade their performance over time due, for example, to new input data received by the system (*data drift*), or even due to changes in the system (*model drift*).

It is therefore necessary to:

- Define the specific metrics to use on the data, the model, the quality of the results and the performance.
- Validate these metrics in the final tests before putting the system into operation.
- Monitor the value of these metrics during operation, specifying the minimum acceptable values from which it is necessary, for example, to retrain the system (when the degradation is due to data) or even stop its execution (for the above reason and/or when the degradation is due to a system error).

These metrics must be associated with a continuous integration system on the system and its data (*MLOps*), allowing the analysis of these metrics every time there is a change in the system and/or its data and thus identifying the level of the metrics defined for each version. To do this, it is necessary to adhere to the life cycle of the system defined in the standards [ISO/IEC 22989] and [ISO/IEC 5338] (see in Glossary the definition of Life Cycle of an AI system).

Who does it apply to?

- The provider shall provide the necessary tools to define and monitor such metrics, as detailed in the guidance covered by Article 15 of the AI Act.

- The provider should include a continuous integration system (MLOps) to identify how changes affect different versions of the system.
- The user, as the final responsible for the system in its operation, must monitor these metrics following the recommendations provided in the guides detailed in Article 15 of the AI Act.

To which sections does this measure apply?

- Section 1. Design and development
- Section 3b.ii. Level of accuracy
- Section 3c. Changes
- Section 3d. Human
- Section 3e. HW/SW Resources & Maintenance

Example – Insulin Pump

Accuracy metrics will be used to compare, for example, the reactions of patients to certain doses.

In relation to continuous integration, these metrics must be specially monitored, for example, when deploying a new *version* of the system that includes, for example, a new drug or a new composition of it, or that includes the medication of groups of people with certain additional pathologies not previously contemplated. In this way, based on the aforementioned metrics, it will be transparently guaranteed that the system works according to the expected specifications.

Example – Aid granting automatic system

Accuracy metrics (specified in the accuracy guide, Article 15) will be used to make it possible, for example, to compare the predictions made by the system on families at risk of social exclusion that have finally led to this situation.

In relation to continuous integration, these metrics should be specially monitored, for example, when deploying a new version of the system that includes a new geographical region (data *release*), or that includes the weighting of dependent or disabled people within the family when predicting social exclusion or granting support (*release* model). In this way, based on the aforementioned metrics, it will be guaranteed in a transparent manner that the system does not discriminate positively or negatively against these new concepts included in these releases.

5.9. Apply prudence

The information provided must be relevant. This implies that such information has to be prudent, identifying scenarios in which the information provided may not be appropriate, at least publicly if, for example:

- It leads to greater confusion for users.
- They can be exploited by external agents to violate the security of the system and/or degenerate its operation, learning and reasoning processes. This is something

particular in AI systems compared to traditional systems, and is known as "poisoning attacks", a concept described in detail in the corresponding cybersecurity guide of Article 15 of the European AI Act.
- They reveal private and confidential information, or information subject to GDPR, that may be misappropriated by third parties.

Who does it apply to?

- The deployer, as the owner of the information, will identify sensitive information that should not be provided either because, for example, it is confidential information of its customers, or because it even reveals sensitive information of the supported business process.
- The provider must implement the relevant filters so that the information is not revealed.

Example – Aid granting automatic system

The system provides families with a response of acceptance/denial of aid through the online web channel, with arguments that support it. Such arguments should not include, for example, detailed financial information of the entire family, since some of this information may be private individually.

Example – Insulin Pump

The system provides patients with a track of past and immediate doses through a mobile app. In case of variation in the dose, it will be explained to the patient in a simple way to ensure their understanding, and this information will be secured to avoid access to confidential information by third parties.

To which sections does this measure apply?

- Section 1. Design and development
- Section 3d. Human

5.10. Use causation, minimize correlations

Correlation does not imply causation, as two variables may have a high correlation, but may not be because of the other. A simple example of these concepts is directly using height, not age, to predict whether it is physically safe for a child to ride a roller coaster. With height we would be using a direct causality, while with age we would be using a theoretical correlation that, if not met, could cause accidents.

Correlation **can complicate the transparency of the AI system and its level of accuracy**, since it is necessary to take into account the large number of variables that these systems handle, and the many explicit and implicit relationships that can exist between them. Analysis of the data can help identify such correlations, as detailed in the measure for data transparency reflected herein.

In addition, correlation may depend on the subjectivity of the person who designs and implements the correlation rules they consider, in addition to not contemplating possible unusual but relevant scenarios in the AI system. It is therefore a potential source of bias, and as a consequence a risk in the correct application of the principle of equity.

Who does it apply to?

- When defining the requirements, the deployer should identify valid correlations as well as explicitly unwanted ones, as they are the connoisseur of the information managed by the AI system from a business perspective.
- The provider shall have a detailed identification of how the correlations have been implemented and their impact on the accuracy of the system.

Example – Aid granting automatic system

The data from the available sources will be analysed as indicated in the transparency measure on the data used in this document, since the assignment considers many variables and there are relationships between them.

For example, for the allocation of aid, one of the variables used is the **number of members** of the family unit. A direct correlation can be made mistakenly between the number of family members and the need and/or amount of the aid. It should be kept in mind that this variable must be weighted with other variables of these members, such as, for example, their age (for example, minors do not have the capacity to contribute income to the family), or the possible belonging of any of them to groups of dependent people (for example, a family with only one child being dependent may have greater need than another family with three children without such dependency).

Example – Insulin Pump

In the design of the system, it will be necessary to analyse the data from the available sources as indicated in the transparency measure on the data used in this document, given that the decision about the dose to be applied may have many variables to consider and there are relationships between them.

For example, one of the variables used is the patient's age. A direct correlation between blood sugar level and dose volume can be mistakenly established. In the design of the solution and its associated documentation, it must be taken into account that this variable must be weighted with other variables such as age, sex, physical activity, etc.

To which sections does this measure apply?

- Section 1. Design and development
- Section 3b.ii. Level of accuracy
- Section 3d. Human

5.11. Use counter factuality

In the transparency process, it is necessary to use comparisons. That is, it is not enough to detail the why of an action (prediction or decision), but also why, said action was one instead of another (concept of counter factuality). Or also in what circumstances this action could have been different if some of the information taken into account were modified. This is especially relevant for the people who are the subject of the decisions of the AI system, who have the right to know on the basis of which criteria and information a certain prediction or decision has been made, and in what particular circumstances such an action would be different.

Who does it apply to?

- The deployer, as a connoisseur of the information managed by the AI system, must identify the counterfactual it is necessary for the use case supported by said system and take them into account in the functional definition of the AI system.
- The provider must design and implement the AI system, providing tools that facilitate the implementation of this measure.

Example – Aid granting automatic system

As a result of this measure, the applicant family will be able to find out through the web channel the reasons why the aid has not been granted, and under what particular conditions if it was granted to them (for example, if they include the data of a dependent person that the system has detected in the family, but which was not provided in the application).

To which sections does this measure apply?

- Section 1. Design and development
- Section 3b.vii. Output Information
- Section 3d. Human
- Section 3b.v. Impact on people

5.12. Enable a channel with system usage information

In addition to the measures to be taken into account in the design and development of the system so that it can provide information about its operation in a transparent manner, it is important that there is a means external to the system itself (*web/wiki/doc page*) that collects the information of said system that is aimed at facilitating transparency about its operation.

In view of the specific information reflected in the third section of this article, the structure of this channel could be as follows:

- Contact information for the provider.
- Description of the purpose of the system.
- Use cases for which it is intended.
- Potential use cases not recommended, and the associated risk if used for such purposes.
- Information on its operation:

- System performance and accuracy.
- Potential biases.
- Possible attacks.
- Description of the data used.
- Recording files, to be able to analyse the actions performed.
- System lifecycle and upgrade procedures, including Change version log

Who does it apply to?

- The provider, as the creator of the system.
- The deployer, as they will be the one who consumes the information provided by the provider and must deploy it in their organization among the people who have some level of responsibility for the system.

Example – Grant and Insulin Pump

The content of this channel (*web/wiki/doc page*) may be equivalent to that developed in each of the examples of the measures in this document.

To which sections does this measure apply?

- Section 2. Instructions for use

5.13. Executive summary. List Section-applicable measures

	Measures															
Sections	MT1	MT2	MT3	MT4	MT5	MT6	MT7	MT8	MT9	MT10	MT11	MT12	MT13	MT14	MT15	MT16
1. Design and development		X	X	X	X	X	X	X	X	X	X					
2. Use Instructions												X				
3. Specific informatior																
3a. Contact	X															
3b. Features, Capabilities, and Limitations																
3b.i. Purpose		X														
3b.ii. Accuracy level							X	X		X					X	
3b.iii. Risks from unintended uses			X										X			
3b.iv. Explanation of its output results				X	X	X			X		X					
3b.v. Impact on individuals and groups				X	X											
3b.vi. Input data				X												
3b.vii.Output Informatior				X	X				X		X					
3c. Changes	X							X								
3d. Human oversight		X	X	X	X	X	X	X	X	X	X			X		
3e. HW/SW Resources& Maintenance								X								
3f. Log files																X

MT1	Provide contact with provider
MT2	Catering to the business domain
MT3	Ensure the functional objective of the system
MT4	Transparency about the data used
MT5	Detailing from the most global to the most particular
MT6	Adapting the language
MT7	Manage complexity
MT8	Use built-in metrics in the system lifecycle
MT9	Apply prudence
MT10	Use causation, minimize correlations
MT11	Use counter factuality
MT12	Enable a channel with system usage information
MT13	Article 09. Risk Management System
MT14	Article 14. Human oversight
MT15	Article 15. Accuracy, robustness and cybersecurity

6. Technical documentation

Article 11 (Technical Documentation) states that the system must be documented in such a way as to demonstrate that it meets the requirements set out in Section 2 (to which this article on transparency corresponds), providing the competent national authorities and notified bodies with the information necessary to assess the conformity of the AI system with those requirements in a clear and comprehensive manner.

The aforementioned article states that such documentation shall contain, **as a minimum**, the elements set out in **Annex IV.** [1]

Furthermore, this transparency guide sets out measures to meet the requirements set out in the European AI Act in the article dedicated to transparency in AI systems. **As a result of these measures, aspects of the system set out below can be documented**, which may help to generate the minimum documentation required.

About contacting the provider

1. Channel user manual that allows the management of requests and incidents between user and provider.
2. A way of accessing this channel from the user and the provider.

On the mastery of functionality

3. Deployer. Requirements document where the user details the transparency needs of the AI system.
4. Deployer. Document that identifies the transparency requirements in accordance with the legislation applicable by the Public Administrations.
5. Provider. Document with the functional scope of the AI system that it makes available to the user, also specifying the possible risks due to unforeseen uses of the same.

On the assurance of the functional objective

6. Deployer and provider. Document with the foreseeable circumstances where the AI system could be used for purposes other than the one for which it was conceived.
7. Deployer and Provider. Document with the risks plan for such uses.
8. Deployer and Provider. Document with the description of the procedure that proactively supervises that such uses do not occur during the operation of the system.

About the data used

9. Deployer and provider. Document with the description of the data sources used by the system both for its learning and in its use.
10. Deployer. Document with the description of the usefulness and the implication in the use of said data in the specific use case.
11. Provider. Technical and user manuals of the tools that allow a detailed exploratory data analysis (EDA) of these sources.

[1] SMEs, including start-ups, may provide the technical documentation specified in Annex IV in a simplified manner. To this end, the Commission shall establish a simplified technical documentation form tailored to the needs of small and micro-enterprises. Where an SME, including start-ups, chooses to provide the information required in Annex IV in a simplified manner, it shall use the form referred to in this paragraph. Notified bodies shall accept that form for the purposes of conformity assessment.

12. Deployer. Documents demonstrating that the deployer has carried out a detailed analysis of such data using at least those tools.

About the operation

13. Provider. Technical and user manuals that allow the deployer to use the functionality of the AI system that allows the understanding of the global reasoning mechanism of the AI system.
14. Provider. Technical and user manuals that allow the deployer to use the functionality of the AI system that allows the understanding of the predictions and decisions made by the AI system on a subset of information with similar characteristics.
15. Provider. Technical and user manuals that allow the deployer to use the functionality of the AI system that allows each one of the individual predictions and decisions of said system to be analysed.

On complexity

16. Deployer. A document that analyses whether the technical complexity of the AI system that will support your use case can become a blocker for the needs of transparency about the system.
17. Deployer. Document that describes the possible alternatives (if any) that are less technically complex but that guarantees transparency about the system, as well as the possible loss of other capabilities (e.g., accuracy) with these alternatives.
18. Provider. In case the AI system uses black box models, technical and user manuals on the tools that allow you to have transparency about the system are especially necessary.

About the metrics integrated into the system lifecycle

19. Provider. Document with the metrics to be used on data, the model, the quality of the results and the performance of the AI system, specifying the minimum acceptable values for these metrics from which it is necessary, for example, to retrain the system (when the degradation is due to data) or even stop its execution (for the above reason and/or when the degradation is due to a system error).
20. Provider. User manual so that the deployer can monitor the value of these metrics through the tools provided by the provider.
21. Deployer. Recording reports with the value of these metrics in the final tests before putting the system into operation.
22. Deployer. Periodic recording reports with the value of these metrics when the system is already in operation.
23. Deployer. Reports that allow you to identify how changes in these metrics affect the different versions/releases of the system when it is in operation.

On prudence

24. Deployer. Requirements document that identifies sensitive information that should not be provided by the AI system.
25. Provider. Document with the description of the mechanisms used so that such information is not revealed.

About correlations

26. Deployer. Document that identifies valid correlations, as well as explicitly unwanted correlations that could lead to difficulty for the transparency of the system.

27. Provider. Document describing how correlations have been implemented in the AI system, as well as, for example, their impact on the accuracy of the system.
28. Provider. User manual that allows the user to analyse such correlations.

On counter factuality

29. Deployer. Requirements document identifying the counterfactual it is necessary for the AI system.
30. Provider. User manual that allows the user to obtain such counterfactual it is.

Over the channel with system usage information

31. Provider. Access and user manual of the external media to the AI system (web/wiki/doc page) that collects information from said system aimed at facilitating transparency on its operation.
32. Deployer. How the user has given access to that external medium in your organization.

7. Self-assessment questionnaire

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

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

8. Annexes

8.1. Glossary

The content of this document aims to be didactic using understandable language and minimizing technicalities, but at the same time being precise from a technical and formal point of view. When technicalities are used, they are explained in the same text where they are exposed, but in others it is not done so since their explanation could divert the thread of argument of the document. These concepts are detailed in this section.

Life cycle

The life cycle of an AI system is the phases that the system goes through from its conception until it is retired.

The [ISO/IEC 22989] and [ISO/IEC 5338] standards define in depth what the phases of the life cycle of an AI-based system are, from a regulatory point of view. For example, [ISO/IEC 22989:2022, clause 6.1] essentially defines the following states in the life of an AI-based system:

- Conception.
- Design and development.
- Verification and validation of the product or service.
- Deployment.
- Operation and supervision.
- Re-evaluation.
- Removal or dismantling.

Source**:** ISO.org

Decision tree

Model for which inference is encoded as paths from the root to a leaf node in a tree structure.

Source: ISO.org

Neural network

A network of two or more layers of neurons connected by weighted links with adjustable weights, which takes input data and produces an output.

Note: While some neural networks aim to simulate the functioning of biological neurons in the nervous system, most neural networks are used in artificial intelligence as implementations of the connectionist model.

Source: ISO.org

Bias

Bias is a systematic deviation from a true state. From a statistical point of view, an estimator is biased when there is a systematic error that causes it not to converge to the real value it is trying to estimate.

In humans, bias can manifest itself in deviations from perception, thinking, memory, or judgment, which can lead to personal decisions and outcomes based on their membership in a protected group. There are different forms of bias, such as subjective bias of individuals, data and algorithm bias, developer bias, and institutionalized biases rooted in the underlying social context of the decision.

Source: JRC. Glossary of human-centric artificial intelligence

Equity

Fairness refers to a variety of ideas known as fairness, fairness, egalitarianism, non-discrimination, and justice. Equity embodies an ideal of equal treatment between individuals or between groups of individuals. This is what is generally sought from a procedural perspective, i.e., the ability to seek and obtain redress when individual rights and freedoms are violated.

Source: JRC. Glossary of human-centric artificial intelligence

GDPR

The General Data Protection Regulation (GDPR) is the European Directive on data protection in criminal matters and other rules relating to the protection of personal data.

Source: European Commission

MLOps

They support the task of deploying AI models into production from upstream environments (development, integration, pre-production, etc.) using automated processes and workflows.

Source: Arxiv.org (Cornell University)

9. References, Standards & Norms

9.1. Standards

This document compiles some of the recommendations of a set of international standards in the field of artificial intelligence, following a standards-based approach. Some of these standards have been published, while others are in the process of being developed, as documented in the publication of the Joint Research Centre (JRC), the science and knowledge service of the European Commission, entitled "*AI Watch: AI Standardisation Landscape. State of play and link to the EC proposal for an AI regulatory framework*".

The Normative Standards, which include the contents of this document are:

[1] ISO/IEC 22989:2022 Information technology – Artificial intelligence– Artificial intelligence concepts and terminology.
[2] ISO/IEC AWI 12792 Information technology – Artificial intelligence – Transparency taxonomy of AI systems.
[3] ISO/IEC TR 24368:2022 Information technology – Artificial intelligence – Overview of ethical and societal concerns.
[4] ISO/IEC DIS 5338 Information technology – Artificial intelligence – AI system life cycle processes.
[5] ISO/IEC TR 24030:2021 Information technology – Artificial intelligence (AI) – Use cases
[6] IEEE 7000-2021 Standard Model Process for Addressing Ethical Concerns during System Design

Additionally, the following standards can also be consulted, where the concept of Transparency is mentioned:

[7] ISO/IEC TS 4213:2022. Information technology – Artificial intelligence – Assessment of machine learning classification performance
[8] ISO/IEC AWI 5259-2. Data quality for analytics and ML - Part 2: Part 2: Data quality measures
[9] ISO/IEC AWI 5259-3. Data quality for analytics and ML - Part 3: Data quality management requirements and guidelines
[10] ISO/IEC CD TR 5469. Artificial intelligence – Functional safety and AI systems
[11] ISO/IEC 23894-2. Information technology – Artificial intelligence – Guidance on risk management
[12] ISO/IEC TR 24027:2021. Information technology – Artificial intelligence (AI) – Bias in AI systems and AI aided decision making
[13] ISO/IEC TR 24028. Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence
[14] ISO IEC TR 24029-1. Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 1: Overview
[15] ISO/IEC CD 24668. Information technology – Artificial intelligence – Process management framework for big data analytics
[16] ISO/IEC 38507. Information technology – Governance of IT – Governance implications of the use of artificial intelligence by organizations
[17] ISO/IEC 42001. Information technology – Artificial intelligence – Management system.
[18] prEN 18229-1 AI Trustworthiness Framework – Part 1: Logging, Transparency and Human Oversight (In progress).

Financiado por
la Unión Europea
NextGenerationEU