7

Guide 7. Data and data governance

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 (AI Act) 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. This guide has been updated in accordance with the Digital Omnibus amendments to the AI Act.

Among the relevant technical references currently under development and applicable are **ISO/IEC 5259-1 to 5259-5, "Artificial intelligence – Data quality for analytics and machine learning (ML)," prEN 18229, "AI Trustworthiness Framework," prEN XXX, "Quality and governance of datasets in AI," and prEN XXX, "Concepts, measures and requirements for managing bias in AI systems,"** which will serve as the basis for establishing a comprehensive framework for governance, data quality, life-cycle management, and trust in artificial intelligence systems.

NIPO: 23026015X

Revision date: 17, July 2026

- Document version control:

Version	Revision date	Major changes
1.0.	10/12/2025	First official release
2.0.	17/07/2026	Revised to reflect the changes introduced by the Digital Omnibus and minor editorial changes

General content

1. Preamble 6
2. Introduction 8
3. European Regulation on Artificial Intelligence 9
4. Data management 12
5. Other elements to consider 32
6. Technical documentation 37
7. Self-assessment questionnaire 38
8. Annexes 39
9. References, Standards and Norms 67

Detailed Index

1. Preamble 6
1.1 Purpose of the document 6
1.2 How to read this guide? 6
1.3 Who is it for? 6
1.4 Use cases and examples throughout the guide 7
2. Introduction 8
2.1 What is data governance? 8
2.2 What elements could I implement and how could I do it to develop adequate data governance? 8
3. European Regulation on Artificial Intelligence 9
3.1 Preliminary analysis and relationship of the articles 9
3.2 Content of the articles in the AI Act 10
3.3 Correspondence of the articles with the sections of the guide 11
4. Data management 12
4.1 Information requirements 12
4.2 Data collection 12
4.3 Data preparation 13
4.3.1 Measuring and improving data quality 14
4.3.2 Data transformation 20
4.3.3 Aggregation of data 21
4.3.4 Sampling of the data 22
4.3.5 Creating and Selecting Features 24
4.3.6 Data Enrichment 25
4.3.7 Data labelling 26
4.3.8 Analysis of bias in data 27
4.4 Data Disposition 29
4.5 Deletion of data 30
5. Other elements to consider 32
5.1 Processing of special categories of personal data 32
5.1.1 The special categories of personal data in the European Regulation on Artificial Intelligence and their processing 32
5.1.2 Detection and/or bias correction of special categories of data in the sandbox framework 36
6. Technical documentation 37

7. Self-assessment questionnaire 38
8. Annexes 39
8.1 ANNEX A – Methods of Data Collection 39
8.2 ANNEX B – Data quality 39
8.2.1 ANNEX B.1 – Data Quality Dimensions 39
8.2.2 ANNEX B.2 – Data quality controls 44
8.3 ANNEX C – Biases 50
8.3.1 ANNEX C.1 – Sources of bias 50
8.3.2 ANNEX C.2 – Bias assessment techniques 53
8.3.3 ANNEX C.3 – Bias treatment measures 55
9. References, Standards and Norms 67

1. Preamble

1.1 Purpose of the document

This guide presents the measures that will be used by providers and deployers to comply with the requirements of Article 10 "Data and data governance" of Regulation (EU) 2024/1689 of the European Parliament and of the Council (European Regulation on Artificial Intelligence). This article sets out the data governance requirements that must be incorporated by any high-risk AI system (HRAIS) and certain general-purpose AI systems (Chapter V, General-purpose AI models). In this regard, throughout the guide we will generally refer to these systems as "AI systems" with the aim of simplifying the discourse.

1.2 How to read this guide?

This section has been prepared to try to **accompany the reader** in reading the guide and help them to achieve a more **agile and effective** understanding **of the contents**.

If the reader does not have experience in data governance, a **first full reading of the guide is recommended at a minimum**, paying particular attention to the key chapters inherent in the article of the European Regulation on Artificial Intelligence (chapters 4 and 5.1). If the reader has extensive experience in data governance, it is recommended to he or she to focus on the **essential chapters inherent to** the article of the European Regulation on Artificial Intelligence (chapters 4 and 5.1), without prejudice to the fact that a complete reading of the guide is also recommended.

In addition, a series of Annexes are provided, the content of which is introduced and referenced in the corresponding sections of the guide. The relevance and importance of these Annexes should be highlighted, as they are catalogues of essential elements for completing and shaping the content of the guide (data collection methods, data dimensions and quality controls, common sources of bias, assessment techniques and measures for the treatment of bias).

1.3 Who is it for?

The requirements described in Article 10 "*Data and data governance*" of the IA Act are primarily geared towards the development of the system, carried out by the provider. In this article, no requirements are specified for the person who makes use of the system, that is, the deployer. In the event that the deployer, in a given situation, participates in the development of the system, or is responsible for the training data or test, they should apply the measures developed for the provider. In spite of everything, the deployer is required to make responsible and ethical use of the system at all times.

However, in accordance with Article *26* "Obligations of deployers of high-risk AI systems"*, in point 4,* to the extent that the deployer exercises control over the input data, said the deployer shall ensure that the input data is relevant to the intended purpose of the high-risk AI system. Therefore, in this case, the deployer must comply with the same measures as the provider.

In this context, the measures detailed throughout this guide, both organizational and technical, are intended to serve as a guide for the provider, except in the cases mentioned above where they also affect the deployer.

1.4 Use cases and examples throughout the guide

In order **to facilitate** the **understanding of the guide,** different examples **are incorporated in it** that are intended to serve as **a reference** for the adequacy of HRAIS in accordance with the data and data governance requirements of the Regulation.

These examples are developed based on the **use cases** described in the Practical Guide with examples and practices to understand AI Act. Specifically, the use cases used in the preparation of this guide are the AI system for employee promotion, the smart insulin pump and the biometric recognition AI system for recording time and attendance at work.

Finally, it should be noted that whenever an example is given, it will be done in an illustrative way. It is recommended for both the provider and the deployer to consider implementing all of the 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 typology. 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 data governance?

In the context of AI, data governance is the set of elements (policies, procedures, processes, standards, etc.) that are implemented to ensure that the data used in the training, validation, and testing of AI systems is adequate, relevant, sufficiently representative, and meets the established quality and completeness requirements.

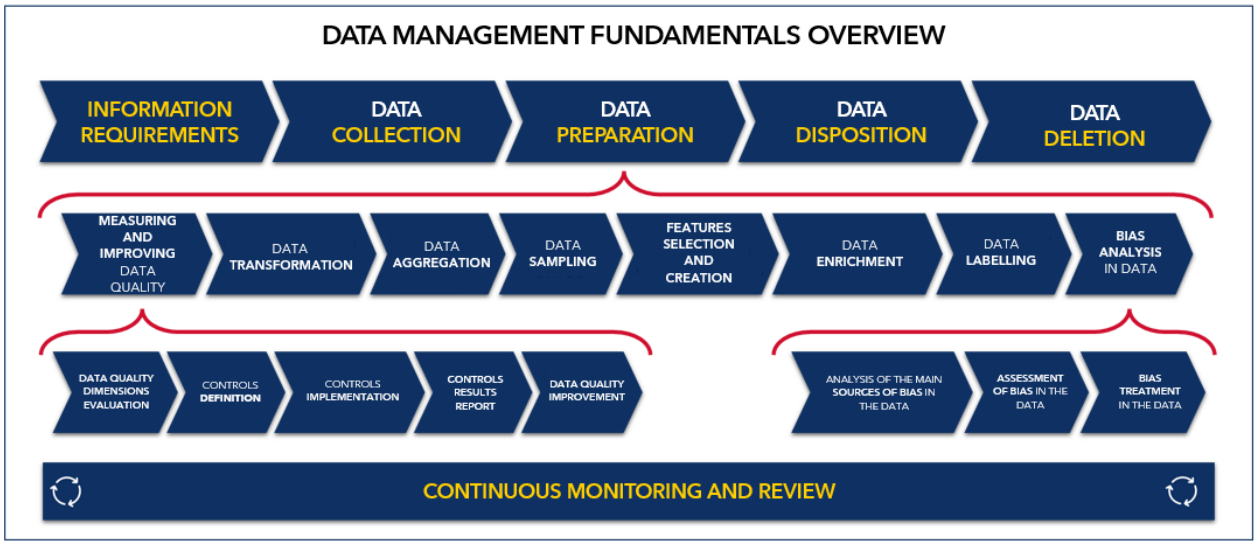
Proper data governance is essential for AI systems to function properly and in accordance with their intended purpose. The lack of data governance can lead to biased or inaccurate results, which could materialize as a risk to the health, safety, or fundamental rights of users.

2.2 What elements could I implement and how could I do it to develop adequate data governance?

This section presents the phases of the proposed model, as an example of how to implement adequate data governance. This model aims to cover all the elements required in Article 10 of the IA Act.

However, some additional elements have been incorporated to what is explicitly provided for in the article due to the completeness and coherence of the essential elements of an adequate data governance (for example, in the article, within the processing processes, the transformation or sampling of the data is not mentioned).

In a broad way, the process can be seen as:



Organized into phases that will be developed in detail in section 4 of this guide.

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 data and data governance. These 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 or individual rights recognized and protected by Union law.

This section includes the articles on data governance of Regulation 2024/1689 of the European Parliament and of the Council of June 13th 2024 (European Regulation on Artificial Intelligence) and details the sections of this guide that address the different elements of these articles.

3.1 Preliminary analysis and relationship of the articles

In order to facilitate the implementation of the measures proposed to comply with the requirements established in the article, the content of the article has been restructured so that some points that seem more generic (such as 2.h and 3) are intertwined with other more specific points (for example, point 2.e), which is where they would have a more specific application:

1. The **training data, validation, and testing sets** will be subject to the following data governance and management **practices**:
    a) Choosing a **suitable design**.
    b) **Data collection processes**.
    c) **Processing operations** for data **preparation** (*annotation, labelling, cleansing, enrichment, and aggregation*).
    d) **Formulation of the relevant assumptions** with respect to **the information the data measure and represent.**
    e) **Pre-assessment of the availability, quantity and adequacy** of the required datasets:
        i. They will be relevant and representative.
        ii. They will be complete.
        iii. They will have the appropriate statistical properties.
        iv. They shall take into account, depending on their intended purpose, the particular characteristics or elements of the geographical, behavioural or functional context in which the HRAIS is intended to be used.
    f) **Analysis of biases** (*taking into account those that may affect the health and safety of people or give rise to some type of discrimination prohibited by Union law).*
    g) **Detection and remediation** of **gaps or deficiencies** in data:
        i. They will be error-free.

2. If the output data is to be used as input data for another model, it should be subject to the aforementioned data governance and management practices.

3. To the extent strictly **necessary** to ensure the **monitoring, detection and correction of bias,** special categories of personal data may be processed, always providing appropriate safeguards **for the fundamental rights and freedoms** of individuals.

Based on this analysis, in section 4 we will propose an orientative approach to be able to address each of these tasks appropriately.

3.2 Content of the articles in the AI Act

AI Act

Art.10 – Data and data governance

1. **High-risk AI systems** which make use of techniques involving the **training of AI models with data** shall be developed on the basis of **training, validation and testing data** sets that meet the quality criteria referred to in **paragraphs 2, 3 and 4** of this Article and in **Article 4a(1)** whenever such data sets are used.

2. **Training, validation and testing data sets** shall be subject to **data governance and management practices** appropriate for the intended purpose of the high-risk AI system. Those practices shall concern in particular:

(a) the relevant **design** choices;
(b) **data collection processes and the origin of data**, and in the case of personal data, the original **purpose** of the data collection;
(c) relevant **data-preparation** processing **operations**, such as **annotation, labelling, cleaning, updating**, enrichment and aggregation;
(d) **the formulation of assumptions**, in particular with respect to the **information** that the data are supposed to measure and represent;
(e) an assessment of the availability, quantity and suitability of the data sets that are needed;
(f) **examination** in view of **possible biases** that are likely to affect the **health and safety of persons**, have a negative impact on **fundamental rights** or lead to **discrimination prohibited** under Union law, especially where data outputs influence inputs for future operations;
(g) appropriate **measures to detect, prevent and mitigate** possible biases identified according to point (f);
(h) the identification of relevant data **gaps or shortcomings** that prevent compliance with this Regulation, and how those gaps and shortcomings **can be addressed.**

3. **Training, validation and testing data sets** shall be **relevant**, sufficiently representative, and to the best extent possible, **free of errors** and **complete** in view of the intended purpose. They shall have the **appropriate statistical properties**, including, where applicable, as regards the persons or groups of

persons in relation to whom the high-risk AI system is intended to be used. Those characteristics of the data sets may be met at the level of individual data sets or at the level of a combination thereof.

4. **Data sets** shall take into account, to the extent required by the **intended purpose**, the **characteristics** or elements that are particular to the specific **geographical, contextual, behavioural or functional** setting within which the high-risk AI system is intended to be used.

5. *(Deleted according to Digital Omnibus)*

6. For the development of high-risk AI systems not using techniques involving the training of AI models, paragraphs 2, 3 and 4 of this Article and Article 4a(1) shall apply only to the testing data sets.

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	Regulation requirement	Section
10.2.a	What elements could I implement and how could I do it to develop adequate data governance?	Section 2.2
10.2.b	Data collection	Section 4.2
10.2.c	Data preparation	Section 4.3
10.2.d	Information requirements	Section 4.1
10.2.e	Measuring and improving data quality	Section 4.3.1
10.2.f	Data bias analysis	Section 4.3.8
10.2.g	Measuring and improving data quality	Section 4.3.1
10.2.h	Measuring and improving data quality	Section 4.3.1
10.4		
4a	Processing of special categories of personal data	Section 5.1

4. Data management

In this section of the guide, the stages of the data management model presented in the introduction will be developed in detail.

4.1 Information requirements

What is it?

It is the process of **establishing** what **information** we need to feed our AI system to achieve the objective for which it has been designed.

How should I approach it?

Every AI system is created with an objective or purpose, for example, to solve an existing problem or cover an identified need. This will be the first phase of an AI system's lifecycle. When it comes to data, the first thing we need to do is identify what information we need to fill that need or solve that problem.

Example

Taking the smart insulin pump AI system **as an example**, the first step will be to analyse the solution we want to implement and the goal we seek to achieve, which in this case is to determine the amount of insulin that a diabetic patient needs at a given time. Thus, the data we will need to collect, for example, will be the blood sugar level, heart rate or volume of oxygen in the blood.

4.2 Data collection

What is it?

It is the process of **obtaining** the **data** that contains the information necessary for the development of our AI system, which will be the data that will feed our AI system to achieve the objective for which it has been designed.

How could I approach it?

The data collection process is generally carried out by extracting data from different sources and this is a particularly relevant element. Why? Because the accuracy and quality of the solution we develop using our AI system will depend, among other elements, on how representative and sufficient this data is (the quantity, sufficiency, completeness and adequacy of the data, along with other related elements, we will discuss in more detail in section 4.3).

Therefore, if we collect data from a single source of information, we run the risk that the accuracy of our AI system generates certain dependencies on that data source. That is, we get good results when we use data from that source to test our system, but we find ourselves with a totally different accuracy scenario if we test our system with data from a new source. In other words, if we train our system with biased data, we run the risk of getting

a system that makes biased decisions (for more details on biases in the data, see section 4.3.8).

Example

Let us now take as an example the AI system for recording **attendance at work** using biometric recognition. If, for example, we train this system with gender- and race-biased imagery, there is a high risk that the system will also show bias and discrimination in its behaviour.

For example, if we primarily use images of white men to train the facial recognition system, the system will likely struggle to accurately recognize and classify people of other genders and races. This could lead to the system making mistakes in identifying people of certain races or genders, and therefore discrimination by the system.

In this context, there are different methods and avenues by which we can approach the data collection process. A detailed list of some of these methods and ways is shown in Appendix A, in order to provide the reader of this guide with some representative examples.

Importantly, the data collection process is closely related to the processes of assessing the availability, quantity and adequacy of the data detailed in section 4.3. Once we address these processes, it is advisable to analyse whether it is necessary to reevaluate or modify the data collection process (for example, incorporating data from new sources). This process of incorporating new data from additional sources has a specific name and is the data enrichment process that we will detail in section 4.3.6.

Additionally, it should be noted that the data collection process and the sources that are finally selected will depend on the need to be covered and the context of the implementation of each AI system. It is appropriate to considerate special consideration to the applicable regulations on the protection of personal data[1].

4.3 Data preparation

What is it?

It is the phase of processing and conditioning the data collected in the previous phase for its disposition and use in the AI system.

How could I approach it?

This processing and conditioning of the data is generally made up of a set of operations that we can carry out depending on the nature of the data that we have collected in the previous phase. In other words, we have collected some data and we need to condition it to be able to use it and for this it is recommendable to address a sequence of treatments depending on how that data is.

[1]Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation); Organic Law 3/2018, of 5 December, on the Protection of Personal Data and guarantee of digital rights; Organic Law 7/2021, of 26 May, on the protection of personal data processed for the purposes of prevention, detection, investigation and prosecution of criminal offences and the execution of criminal sanctions.

In the following subsections, we will describe the main operations or treatments that complete this data preparation phase.

In each section we will address **what each element is** and **how we could approach it**. Additionally, we will include the reflection of **when we could address it**, this is because we will not always have to face each of the processes indicated in this section, it will depend on the circumstances and that is what we will try to solve in that section.

4.3.1 Measuring and improving data quality

What is it?

Data quality is the degree to which the data is fit for the purpose for which it was defined or collected. Data quality measurement is about analysing how well the data is fit for purpose. Improvement is the process that allows us to adjust the degree of adequacy of the data to its purpose, that is, it allows us to increase the quality of the data.

Maximizing data quality is undoubtedly one of the most important elements in the data preparation process.

When could I address it?

Measuring data quality and improving data quality is an essential process in proper data governance. As we will see in the following phases and as we have previously mentioned, it is appropriate to approach each of them according to a series of circumstances that we will analyse in each case. However, measuring and improving data quality is not an optional or circumstance-dependent process, but one that must always be present and properly implemented in our data governance system.

How could I approach it?

The best way to approach any process of measuring and improving data quality is to establish an appropriate data quality methodology. In this section we intend to accompany the reader of the guide in the process of developing their data quality methodology. To do this, we will detail below the phases that are recommended to face in this process:

A. Determine the quality dimensions to be evaluated

The first step is to determine which dimensions of quality we are going to evaluate for our data. To do this, we must know and understand what these dimensions are and then determine which of them we need so that our data has the right quality to fulfil its function. It is important to note that it is appropriate to address this exercise for each of the data whose quality we want to guarantee, that is, the quality dimensions must be analysed individually for each data.

To facilitate this task, Annex B.1 provides and details some examples of the most common and relevant data quality dimensions (such as completeness, representativeness, auditability, consistency, relevance, diversity or credibility).
In this context, the first thing we must do is select which of these dimensions (or others that our organization considers and are not in the inventory provided in Annex B.1) we need so that each instance of our data has the appropriate quality to fulfil its function.

For example, it's quite likely that the completeness dimension is necessary for all of our data. However, the dimension of consistency between data will only be necessary in certain cases, for example, if we have two pieces of data that represent related dates (the date of entry into an infrastructure must be less than the date of departure).

Example

Let's select for this example the AI system related to the use case of **employee promotion**, let's assume that we are an organization that decides to develop this **AI system** that analyses the profiles and performance of workers and helps us determine who deserves a promotion to a greater extent.

Suppose we are also about to determine the quality dimensions of the data that our AI system uses, including the "date of incorporation to the company", "employee status" and "date of departure from the company" data of the employees, and we decide to start defining the quality dimensions for them.

Now, what is appropriate to be asked is this: What dimensions of quality do I need my data to have in order for it to meet its purpose and for my AI system to function properly? Let's start by selecting two of the dimensions set out in Annex B.1, "Completeness" and "Consistency", to develop this example. We, who know our purpose and our data, determine that the data "employee status" and "date of joining the company" must always be informed since every employee must have a date of incorporation and a situation (e.g. active, leave, former employee). On the other hand, the data "date of departure from the company" will only be reported for those former employees who are no longer in our organization. As for "Consistency", we have several relationships that must be fulfilled between these data. For example, if the data "date of departure from the company" is reported, then the data "employee's situation" can only take the value "former employee". In addition, the data "date of departure from the company" must be greater than the data "date of incorporation into the company".

B. Define the quality controls to implement for each dimension

Once we have identified the dimensions of our data quality, what we could do is define the data quality controls for each requirement. Defining a quality control consists of establishing a metric that provides us with the information we need to in relation to the state of the quality of our data.

The selection of the quality dimensions could be a relatively simple task since it will generally depend on the nature of the data itself, as we have seen in the previous phase. However, it is likely that at this point we will ask ourselves, what controls do I have to define, and how many will I have to define? Again, the answer is linked to the objective of this exercise, which is to ensure that each of our data has the appropriate quality to fulfil its function. Therefore, it is useful to analyse and to determine which controls are appropriate in each case in order to achieve this objective.

To facilitate this task, Annex B.2 provides and details some examples of quality controls of the most common and relevant data for each of the dimensions detailed in Annex B.1. In addition, the detail of how to define them in each case is also incorporated.

Example

Continuing with the example described in the previous phase, we are now looking at what controls we need for the data completeness requirement "employee status", "date of joining the company" and "date of departure from the company" and for the consistency requirement between the last two. Based on the inventory of controls provided in Annex B.2, we decided to define the following controls to ensure the quality of our data:

ID	Data	Type of Control	Description	Definition of control
0001	Employee status (A)	Completeness	Ratio of reported data (not null) in the dataset	Non-null records of data A / Total records of data A
0002	Incorporation date (B)	Completeness	Ratio of reported data (not null) in the dataset	Non-null records of data B / Total records of data B
0003	Departure date (C)	Completeness	Ratio of reported data (not null) in the dataset	Non-null records of data C / Total records of data C
0004	Joining date (B) and departure date (C)	Consistency	Ratio where the consistency rule "B<C" is correctly fulfilled	Records where B<C is met / records where B and C are fulfilled

Additionally, it is important to note that the definition of data quality controls must be addressed at each of the points of the data life cycle where we consider appropriate to ensure that they arrive with the appropriate quality when they are used by our AI system.

Example

Let's say I download data from an open data source and dump it into an Excel sheet. This Excel sheet is loaded directly into a SQL database and from there I make an extraction on my Python platform where I have my AI system implemented. In this life cycle we have our data in three different repositories and for this, three different processes take place, let's assume the following:

a) Download from the internet to the Excel sheet.
b) Loading the Excel sheet to SQL database.
c) Extraction of a selection of the data from the SQL database to the Python platform.

The definition of quality controls may be different depending on the repository. Going back for a moment to the previous example, if when going from "b)" to "c)" we no longer select the "date of incorporation into the company", in point "b)" we can define the controls mentioned in the previous example, but in "c)" we can no longer define the controls that implied the data that we have not selected.

In this context, it is appropriate to decide at which points in this life cycle we should define and subsequently implement quality controls. In order to determine this, it is recommended to analyse our life cycle and the needs that we consider appropriate to guarantee the quality of our data. With regard to data quality measurement, it should ideally be carried out at the source repositories. From there, controls at the different layers or stages of the lifecycle should be more focused on process quality (i.e.,

ensuring that data have been correctly copied/ingested/transferred), and therefore be more technical in nature rather than functional. This approach helps to avoid duplicating the same controls across multiple layers and reduces the complexity associate with their management and remediation. In the example above, it's a very short lifecycle, and we may have the ability to define and implement quality controls at all points. But what if we have a much longer life cycle? Then it is recommended to determine which points are the most critical and which are the ones that allow us to best guarantee the quality of our data.

Finally, it is appropriate to determine the necessary quality requirements for each control, this will allow us to evaluate whether the quality of our data is sufficient or not. For example, if the completeness of one of the characteristics of the dataset that feeds our AI system is essential, we must establish a quality requirement for the completeness control of that data of 100%.

C. Implement defined controls

The next phase is to implement the quality controls defined in the previous phase. It consists of technically developing the controls we have defined, which means translating the defined rules into the technical language of each point in the data life cycle. Therefore, this exercise will depend on the type of platform and language on the points at which we have defined the controls. Here are some examples to help the reader better understand this process:

Example

Let's assume that we have defined the quality controls listed in the previous example and that we have defined them for the three points in the data lifecycle specified in that example. In this context, we can now implement these controls and, as we have explained, how we do it will depend on each type of platform.

In this context, we will be able to implement the rules that allow us to count the number of records reported (not null) and the rules that allow us to compare the records of the incorporation and exit dates in the language of each of these points, as follows:

a) Implementation of controls in Excel à we will use Excel Macros and the VBA language.
b) Implementation of SQL controls à We will use the SQL language itself.
c) Implementation of controls in Python à we will use the Python language itself.

D. Report the results of the controls

In the previous phases, we have been in charge of defining and implementing data quality controls, i.e. we have provided the appropriate means for measuring the quality of our data.

At this point, the next task will be to report the results of the controls and arrange these results appropriately to carry out the analysis of the degree of adequacy of the data for the purpose for which they have been defined or collected, that is, their quality.

To do this, what we will have to do is determine how we want to report these results and what information we want to incorporate into these reports. To facilitate this task, we detail below some of the most relevant elements that an adequate data quality report can contain:

- **Context of the report**: in this section we will explain the scope and context of the report and the objective it aims to cover.
- **Information that is reported**: what data is incorporated, the definition of the quality controls reported, the quality requirements of each control, the assessment of the quality of the data, as well as any additional information or detail that is considered necessary.
- **Information on execution times**: the date and time at which each control was executed, the date and time at which the report was prepared, the frequency of execution of the controls, the frequency of execution of the report, as well as any information related to periodicities and times of execution that is considered necessary.
- **Responsible for the report:** among them are: responsible for the execution of each control, responsible for the evaluation of the results of the controls, responsible for the preparation of the report, responsible for the issuance of the report, responsible for the reception and review of the report.
- **Conditions of storage of the report**: the storage location of the report must be informed, for any subsequent consultation.

Example

In this example we will try to arrange, without going into detail, some examples of each of the elements to be incorporated. It is not intended to be an exhaustive reflection of what would be a complete report of the results of the data quality assessment.

Let's assume that we continue with the example of the AI system related to the use case of **employee promotion:**

- **Report context:** This report represents a report that collects the results of the quality checks of the data that feeds the AI system for employee promotion. The main objective of this report is to provide a transparent and accurate assessment of the reliability of the data used by the system.
- **Reporting Information:**

ID	Data	Type of Control	Description	Definition of control	Minimum quality requirement	Control evaluation
0001	Employee status (A)	Completeness	Ratio of reported data (not null) in the dataset	Non-null records of data A / Total records of data A	0.9	0.98
0002	Incorporation date (B)	Completeness	Ratio of reported data (not null) in the dataset	Non-null records of data B / Total records of data B	0.9	0.92
0003	Departure date (C)	Completeness	Ratio of reported data (not null) in the dataset	Non-null records of data C / Total records of data C	0.9	0.94
0004	Joining date (B) and departure date (C)	Consistency	Ratio where the consistency rule "B<C" is correctly fulfilled	Records where B<C is met / records where B and C are fulfilled	0.9	0.97

- **Runtime Information**:
  - Date and time of the last execution of each control:
    - 0001 : 28/09/2023 at 16:45
    - 0002 : 28/09/2023 at 16:50
    - 0003 : 28/09/2023 at 16:55
    - 0004 : 28/09/2023 at 17:00
  - The date and time the report was prepared: 09/28/2023
  - The frequency of execution of the controls: daily.
  - The periodicity of execution of the report: weekly.
- **Responsible for the report:**
  - Responsible for the execution of each control: IT department.
  - Responsible for the evaluation of the results of the controls: the person responsible for the AI system, i.e. human resources.
  - Responsible for the preparation of the report: analyst of the human resources area.
  - Responsible for the issuance of the report: automatic issuance after generation via email.
  - Responsible for the receipt and review of the report: director of human resources.

E. Develop data quality improvement measures

At this point, we will have defined and implemented the data quality controls and prepared the corresponding report that allows us to analyse whether the quality of our data is sufficient or not. What is appropiate to do now is to determine the appropriate measures to remediate or correct the data quality incidents identified.

At this point, we may have doubts such as, Who should remedy the incidents? If we have identified a large number of incidents, how will we remedy all of them?, Which one should we remedy first?, Should we take into account all those affected by these data?

To try to answer these questions and facilitate the reader's understanding of the data quality remediation process, we detail some of the most relevant elements that could be considered below:

- First, it is appropriate to inventory all the data quality incidents identified.
- Then, it is recommended to determine which data and incidents are the most critical for the proper functioning of our AI system.
- Next, we can consult all the interested parties (*stakeholders*) that could be affected by the impact on the quality of this data.
- Then, after analysing the criticality of the incidents and after agreeing with the interested parties, it is advisable to establish an order of priority for the remediation of these incidents.
- Finally, it is appropriate to define a remediation plan for each incident identified, according to the established order of priority.

Example

In this example, we will try to show, without going into detail, some examples of each of the elements that can be incorporated. It is not intended to be an exhaustive reflection of what a complete data quality improvement process would look like.

Let's assume that we continue with the example of the AI system related to the use case of employee promotion**:**

- First, we will document all identified data quality issues (e.g., inventorying them in an Excel sheet, where we will assign an ID to each of them, identify the control that has detected it, the person responsible for that control, which processes impact the quality of that data, and all the documentation we consider appropriate to help us identify the incident, categorize and prioritize it).
- Then, with the information provided in the incident inventory, what we will do is, after agreeing with the different actors involved (for example, the owners of the data and the developers of the AI systems) determine an order of priority of these incidents according to which we will try to remedy each of them.
- *Then, and also jointly with the actors involved and stakeholders who could be affected by the impact on the quality of this data, we will agree on a remedy plan.*

4.3.2 Data transformation

What is it?

It is the process of converting data into a homogenized format.

When could I address it?

It is recommended to address this process when we have non-homogeneous data for its use in our AI system. The data may not be homogeneous because it has different formats,

is expressed in different units of measurement or has indices on different scales, among other reasons. This may happen, for example, if we are collecting data from a variety of different sources.

How should I approach it?

By homogenizing, normalizing or scaling the data collected.

Example

Let's select for this example the AI system related to the use case of employee promotion.

Let's also suppose that our organization has just absorbed another organization that carried out an activity similar to ours (this implies a change in the internal context of the organization as we can see in section 2.2 of the risks guide). In this context, we will have to incorporate new workers into our promotion and promotion AI system so that it also analyses their profiles and performance.

Finally, let's assume that our organization, located in Europe, incorporates the information regarding the salaries of employees in Euros, that the absorbed organization is Japanese and incorporates this information in Yen. If this information is used by our AI system for its employee promotion analysis, it is appropriate to homogenize this data, for example, by transferring the information of new workers from Yen to Euros.

4.3.3 Aggregation of data

What is it?

It is the process of grouping data in order to be able to analyse them and provide observations to draw conclusions about the information they represent.

When could I address it?

It is appropriate to approach this process when we need to know characteristics of data groupings.

How could I approach it?

By transforming the original datasets into datasets grouped according to the characteristics that we consider appropriate for the analysis we will carry out.

Example

Let's select for this example the same AI system from the previous section, i.e. the one related to the use case of employee promotion**.** Let's now assume that we are in the most incipient phase of the AI system development process and that we are selecting the

characteristics (a process explained in section 4.3.5) of the data with which we will train our AI system.

Let's also assume that in this specific case we are developing this AI system in the context of a company manufacturing a certain type of industrial parts.

In this feature selection phase, we want to incorporate the average number of parts processed by each employee per day. We only have a database that stores a record for each employee and working day with the total number of parts processed in each day.

In this context, if we finally consider incorporating this feature as training data of our system, which we want to train with data at the employee level, we must first have the

Employee ID	Day	Processed parts
E00001	1	234
E00001	2	259
E00001	3	289
E00002	1	301
E00002	2	297
E00002	3	278
E00003	1	223
E00003	2	245
E00003	3	237

Employee ID	Average number of pieces processed per day
E00001	260.7
E00002	292.0
E00003	235.0

aggregated data per employee. To do this, we will transform the previously mentioned table and add it per employee by calculating the average number of pieces manufactured per day. A simple example of this process is provided below:

4.3.4 Sampling of the data

What is it?

It is the process by which a subset of data is extracted from a larger dataset.

When should I address it?

We will generally make use of data sampling when we need to generate and run our AI system agilely, for example, to carry out a test that allows us to observe its operation without having to make use of the entire dataset.

We can also make use of data sampling when selecting training data, validation and testing datasets.

How should I approach it?

There are two techniques mainly known and used to carry out an adequate data sampling:

- Random sampling: Each sample in the dataset has an equal chance of being selected.

- Stratified sampling: the data are divided into subgroups according to certain characteristics. This type of sampling is generally used to ensure that each subset is adequately represented.

Example

Let's select for this example the same AI system from the previous section, i.e. the one related to the use case of **employee promotion.** Again, we are developing the AI system in the same context as a company manufacturing a certain type of industrial parts.

Now suppose that when analysing the data to train our AI system and selecting the characteristics we will use, we realize that the characteristics we need to collect to train our AI system are not the same for all groups of employees.

For example, for operators who manufacture industrial parts, we need the characteristic related to the number of parts that each operator processes. On the other hand, for employees who work in the offices we do not have this data, as it does not exist. For this other subset, we will use other characteristics to train our system, for example, the turnover that each worker manages.

In this context, one solution will be to use stratified sampling, so that we divide our dataset into two subsets, one with the data of the factory operators and the other with the data of the office operators. In this way, we will be able to train two different AI systems and we will use each of them in the processes of promoting employees depending on whether they are factory or office.

4.3.5 Creating and Selecting Features

What is it?

They are the processes that allow us to increase or decrease the dimensionality of our dataset. We can increase dimensionality by creating new features and reduce it by selecting a subset of the available features.

When could I address it?

It is recommended to create new features when we require them and do not have them in our feature set. On the other hand, we can select a subset of the features we have available when we determine that we only need some of them and not all of them.

How could I approach it?

The process of creating features will usually be approached by combining some of the available features. The selection of characteristics will be carried out by determining which characteristics we are interested in using to train our system.

We may also want to get rid of features that may be redundant or don't add differential value to our AI system. For this specific scenario, there are techniques such as Principal *Component Analysis (PCA*). This technique allows us to identify those characteristics that are highly correlated and that might not provide a differential value if, for example, we are training a classifier based on neural networks.

Example (Feature Creation)

Let's select the **smart** insulin pump **AI system** for this example and suppose that we have determined that the characteristics we need to train our system are blood sugar level, heart rate, and blood oxygen volume.
Let's also suppose that due to a hardware limitation it is very difficult to obtain the volume of oxygen in the blood, which usually results in inaccurate or incomplete measurements. Therefore, measuring the volume of oxygen in the blood is not a viable solution, but we need to obtain this data to be able to determine the insulin that the patient needs and, therefore, to be able to develop our AI system. What can we do? One option is to investigate whether there is any way to estimate this data.
Suppose, in this context, that there was a very precise way of estimating the volume of oxygen in the blood by using other parameters that we are able to obtain, for example, the heart rate and the volume of oxygen in the lungs. If we make the decision to incorporate this estimated characteristic of blood oxygen volume, we will be creating a new characteristic in our dataset.

Example (feature selection)

Let's continue with the example developed for the creation of features. Now suppose that in our initially designed dataset we also had the characteristic represented by the cholesterol level. Suppose we also develop a Principal Component Analysis on our dataset and observe that cholesterol level is a characteristic that is fully correlated with blood sugar level. In this scenario, generally, we will decide to do without one of the two characteristics since it would not provide us with a differential value in the training of our AI system.

4.3.6 Data Enrichment

What is it?

This process consists of extending the features and data available in our dataset. The objective is to try to provide information of additional value to the available information.

When could I address it?

Data enrichment involves the incorporation of data from new sources, and is a process closely related to the data collection process detailed in section 4.2 (as we already anticipated when we introduced the concept of data enrichment in that section). However, when we talk about data enrichment, we should interpret it as a process after the initial data collection, which we will address if we identify additional data needs or data characteristics.

How could I approach it?

We can approach it in the same way that we approach the data collection process (section 4.2) for the new data sources we consider

4.3.7 Data labelling

What is it?

Data labelling or data annotation is the process by which we assign identification labels to the data we have collected. For example, if we have collected data to develop an animal image classifier, it is the process by which we assign the label that identifies the type of animal contained in each of the images.

When could I address it?

We will address this in the event that, to develop our AI system, we need labelled data and we do not have such labels in the sources that feed our dataset.

If we look at the most common types of machine learning, labelled data is necessary in the development of AI systems based on supervised learning (e.g., image classifiers). In contrast, AI systems based on unsupervised learning do not require labelled data (for example, segmentations of related data groups).

How could I approach it?

Data labelling tasks are generally costly in terms of time and resources. This is because we typically need large amounts of data to develop our AI systems. It is true that there are also tools for the automatic labelling of data, however, it is appropriate to assess whether we should incorporate some type of human surveillance into this process (HITL – *Human in the Loop*) to guarantee that the labelling is appropriate to our needs (for more details on human oversight processes, see the guide in the article Human oversight).

Here are some of the most common data labelling approaches:

- **Internal labelling**: in this case, the labelling process will be addressed by the experts in charge of the AI systems within the organization. It is a methodology that will provide us with guarantees in accuracy and quality since we control the entire process internally, but it is also the type of labelling that will consume the most resources directly.
- **External labelling**: in this case we hire experts from outside our organization to carry out the tasks of labelling the data. This option can be particularly interesting for high-volume labelling processes at short intervals of time. However, it will involve a greater direct economic cost than the internal labelling option.
- **Crowdsourced tagging**: This approach is based on a distributed data labelling process across the web. There are platforms that are responsible for designing and offering these services. This solution is perhaps the most efficient and profitable of all those described; however, it does not provide us with the same guarantees of quality and accuracy as internal labelling nor do we have a contract in between with an organization that responds in the event of any type of incident. Additionally, it should be considered that the poisoning of training data is a possible attack vector on our system (See Cybersecurity Guide)

- **Automatic labelling:** This approach is based on an automated process that labels data according to a predefined set of rules or a stored value associated with the model's objective. Human oversight is carried out on a random and representative sample of the labelled records: if there is a significant number of discrepancies between what a human would have labelled and what the automated process produced, the rule system or suitability of the automatic labelling is reviewed. Such labelling should be accompanied by sufficient validation and supervision to mitigate the risk and drift.

  In this regard, the integration of AI agents into data governance frameworks such as UNE 0085 and Data Governance Act (DGA) can foster innovation by automating management and compliance tasks. However, it requires specific governance measures to ensure transparency, traceability, and human oversight. The literature proposes hybrid human-machine models and alignment architectures to audit automated decision-making processes.

Example

Let's say we go back to the AI system example related to the use case of **employee promotion.** Let's assume that we have already selected the necessary characteristics and collected the historical employee dataset with which we intend to train our AI system. This AI system will consist of a classifier that will help us determine whether an employee should be promoted or not.

In this scenario, we need to incorporate an additional feature into our dataset that will be the label used by the supervised learning system that we will train. This label will consist of a binary mark that will identify whether the employee should be promoted or not.

From what information will we build that label? Well, based on the characteristics that we have selected to train our system because they are those that determine whether an employee should be promoted. We assume that the process is complex enough to require a specific evaluation by an expert for each specific case (otherwise we would not need to train this type of system and we could define a set of simple rules).

Who will address this evaluation of each case and incorporate the promotion label? Here we could choose between any of the alternatives explained above. For example, we can suppose a scenario where our organization has a large team of data scientists and in this case they will be the ones who take charge of this task after aligning a series of criteria with the human resources department who have the specific knowledge of the data used for the evaluation of employees.

4.3.8 Analysis of bias in data

What is bias in data?

Bias is the tendency or potential systematic error that we could incur if we had an imbalance in the data. This occurs when certain elements of a dataset are overweighted or overrepresented.

Example

For example, if we were to train our **AI system**, previously described, for **promoting employees** with historical data from management positions between the years 1900 and 2000, the vast majority of cases of successful promotion from which the system would learn would be of male people. If we try to use this system (trained with the aforementioned data) in the current context, we could obtain that all the cases of promotion of proposed employees are of male people, since our dataset was biased by this characteristic and the decisions of our system could be conditioned in this direction.

When could I address it?

The analysis of biases in data, like the measurement and improvement of data quality, is an essential process in proper data governance. As we have seen in the previous phases, we must look at specific circumstances to determine when we will address each of them. However, data bias analysis is not an optional or circumstance-dependent process, it must always be present and properly implemented in our data governance system.

How could I approach bias analysis?

The goal of data bias analysis is to have the ability to identify and evaluate the potential bias that our data may present and how it could impact the development of our AI system. Also, to be able to determine the need to incorporate measures that mitigate this impact.

We place special emphasis on defining this objective since it is important to emphasize that bias in data is not a purely negative or harmful element that we should always try to eliminate or mitigate. What is important is to have the ability to identify it and evaluate its impact.

Example

Following the previous example, the impact on the selection process would be negative for people of a gender other than male. Most importantly, we must understand that this is an inherent bias of the historical data we have collected, i.e., the data is not incorrect or contains errors, it simply represents an imbalance with respect to a characteristic (in this case gender).

Should we correct this imbalance in our data? This is a question that is appropriate to address in each specific context and scenario, there is no one-size-fits-all answer or solution.

What we can do is try to design a process that describes the steps we should take when dealing with a bias analysis and that is what we will try to address in this section of the guide. This process is divided into three phases that we detail below as a guideline:

1. **Analysis of the main sources of bias in the data:** The first thing we can do is know and understand what the main sources of bias in the data are, that is, the main elements that can cause my data to be biased. To facilitate this task, some examples of the most relevant sources of bias are set out and detailed in Annex C.1.
2. **Assessment of bias in the data:** Secondly, we can make use of the main techniques for assessing bias in the data. These techniques will help us identify if there is any type

of bias in our data and measure and assess its impact. As in the previous phase, a detailed list of the most relevant techniques for assessing bias in the data is provided in Annex C.2.

3. **Treatment of bias in the data:** Finally, it is recommended to determine the need to implement measures to address the biases identified and evaluated. These measures will help us mitigate the impact of biases present in our data. Likewise, the most relevant bias treatment measures are detailed in Annex C.3.

4.4 Data Disposition

What is it?

It is the process of providing the data, after going through all the preparation processes detailed in the previous sections, for use in the AI system.

How could I approach it?

In the first phase of the process, we determine the information requirements we need to develop our AI system, then we collect the data that contains and represents that information, then we process it and prepare the data and now it is advisable to make it available for proper use in the development of the AI system.

The availability of the data consists of providing them once they have been prepared, using the appropriate technical tools. In the same way as when we implemented quality controls (where we saw how it depended on the type of platform and language on the points in which we had defined the controls) providing the data will depend on where we are going to implement our AI system.

Example

For example, let's suppose that in the data collection phase we started with a download in an Excel sheet from an open data source, let's also suppose that we later loaded that Excel sheet into our Python platform and there we developed all the processes of preparing the data using Python code.

Now suppose that we have designed our AI system in such a way that it can only be fed by a relational database in SQL in order to ensure the appropriate security and information integrity measures. In this context, in order to properly have the data available and that it can be used by the person in charge of developing the AI system (which can be ourselves) we must load the data into this relational database in SQL.

In addition, there are also other elements that are appropriate to be considered for the proper development of the data disposition process:

- Specify the data disposition process itself, where the data is stored, and how we should access it appropriately.
- Implement authentication and data access processes.
- Document and maintain a version control process of data disposition processes.

- Keep a record of the different updates of the data that are available.
- Inform about the processes developed until the availability of the data:
  - Evaluation of information requirements.
  - Data collection.
  - Data preparation (data quality measurement and improvement processes, data transformation, aggregation and sampling, feature creation and selection, data enrichment and labelling, and bias analysis).
- Inform about the possible presence of personal data or special categories of personal data (see section 5.1).
- Inform about the security and privacy protection measures (such as pseudonymization or anonymisation) that have been implemented (see section 5.1).
- To report on possible improper and reasonably foreseeable uses of the data provided.
- Document any additional details of the data and metadata necessary to facilitate the understanding of the data and its appropriate use for the development of the AI system (e.g. the distribution and statistical properties of the data, the types of data, the format of the data, or the dates of acquisition and updating of the data).

4.5 Deletion of data

What is it?

It is the last phase of the data life cycle, consisting of the removal of the data that has been arranged and used to develop the AI system.

How could I approach it?

The process of deleting the data that we have made available and used for the development of our AI system will generally be carried out, either by deleting it, or by transferring it, as detailed below.

For the **transfer of data**, we could:

- Identify a recipient who presents the appropriate guarantees to assume the responsibility of safeguarding the data with the respective associated legal and contractual obligations.
- Document a clear agreement with the recipient according to which they agree to take charge of the data and the associated obligations.
- Ensure that the transfer of data does not violate any legal, contractual or data retention obligations.
- Generate a report with all the details of the data transfer process addressed.

For the **deletion of the data**, we could:

- Make sure that there is no user or actor involved in the development of the AI system that still requires access to the data for a specific reason or purpose.
- Ensure that the data deletion process we are about to carry out does not violate any legal, contractual or data retention obligations.
- Ensure that they are removed from all locations where they may be stored.

- Evaluate whether there is any possibility of restoring the data, partial or total, from the AI system trained with this data. In this case, we will need to document it properly and consider it in the data deletion process.
- Review whether the data has any kind of cultural, social, or historical significance. In such a case, we must document and detail why we do not opt for the transfer of the data to a recipient who can maintain all data obligations and their importance.
- Review whether it is possible and adds value to donating data to the public domain. This can benefit different fields of research.
- Generate a report with all the details of the data deletion process addressed.

In addition, we must have the appropriate means to manage possible specific requests for partial deletion of data. As is known, some regulations such as GDPR, can force us to delete specific sets of data in different scenarios and contexts. With respect to this **partial data deletion process**, we must:

- Assess the impact of this partial data deletion on the remaining dataset, its specifications, and its quality.
- If the remaining data no longer meets its specifications and properties, we will need to document and inform all stakeholders of this event and seek to provide mitigating measures.
- Ensure that data that has been deleted is effectively deleted from all locations where it may be stored.
- Generate a report with all the details of the partial data deletion process addressed.

Example

Continuing with the example of the AI system related to the use case of **employee promotion,** let's now assume that we receive a request for deletion of data from a former employee and, let's also assume that data protection law supports this particular request in the context and circumstances in which it happens.

In this scenario we should:

- **Impact assessment on the total set of data and the operation of the system**: in this case it is a specific deletion of data from a single former employee and has been considered not to have a material impact on the total set or on the operation of the system.
- **Documentation and information to interested parties**: an email is sent to the areas that had this data stored in their systems. These are human resources and the *Advanced Analytics team*.
- **Deletion of the data in all locations**: the human resources area (*owner* of the data) is contacted to consult the functional trace and identify all the locations in which the data was stored. Then, in collaboration with the IT team, these databases are accessed and the data is deleted.
- **Generation of the report**: a report is generated with all the details of the partial data deletion process addressed (description of the request, impact assessment, locations where the data was located, dates and times of the actions carried out).

5. Other elements to consider

5.1 Processing of special categories of personal data

As a starting point, it is necessary to remember that providers and deployers are obliged to comply with personal data protection regulations when they carry out personal data processing during the different stages of the life cycle of AI systems.

5.1.1 The special categories of personal data in the European Regulation on Artificial Intelligence and their processing

European and national regulations on the protection of personal data cover special categories of data whose processing may pose a particular risk to the rights and freedoms of individuals. The special categories of personal data are as follows:

- Personal data revealing racial or ethnic origin;
- Personal data that reveals political opinions, religious or philosophical beliefs;
- Personal data revealing trade union membership;
- Genetic data;
- Biometric data processed for the purpose of uniquely identifying a natural person;
- Personal data relating to the health, sex life or sexual orientation of a natural person.

Article 4a of the European Regulation on Artificial Intelligence aims to allow providers of high-risk AI systems to process special categories of personal data with the **aim of detecting and correcting biases** that may be present in these AI systems. This is allowed to avoid discrimination caused by biases present in AI systems.

Thus, Article 4a. "**Processing of special categories of personal data for bias detection and correction**" of the European Regulation on Artificial Intelligence states that:

AI Act

> Art.4a – Processing of special categories of personal data for bias detection and correction
>
> 1. To the extent strictly necessary to ensure bias detection and correction in relation to high-risk AI systems in accordance with Article 10(2), points (f) and (g), of this Regulation, providers of such systems may exceptionally process special categories of personal data, subject to appropriate safeguards for the fundamental rights and freedoms of natural persons. In addition to the provisions set out in Regulations (EU) 2016/679 and (EU) 2018/1725 and Directive (EU) 2016/680, as applicable, all the following conditions shall be met in order for such processing to occur:
>
> (a) the bias detection and correction cannot be effectively fulfilled by processing other data, including synthetic or anonymised data;

(b) the special categories of personal data are subject to technical limitations on the re-use of personal data, and state-of-the-art security and privacy-preserving measures, including pseudonymisation;

(c) the special categories of personal data are subject to measures to ensure that the personal data processed are secured and protected, subject to suitable safeguards, including strict controls and documentation of the access, to avoid misuse and to ensure that only authorised persons have access to those personal data with appropriate confidentiality obligations;

(d) the special categories of personal data are not transmitted, transferred or otherwise accessed by other parties;

(e) the special categories of personal data are deleted once the bias has been corrected or the personal data has reached the end of its retention period, whichever comes first; and

(f) the records of processing activities pursuant to Regulations (EU) 2016/679 and (EU) 2018/1725 and Directive (EU) 2016/680 include the reasons why the processing of special categories of personal data was strictly necessary to detect and correct biases, and why that objective could not be achieved by processing other data.

2. Providers and deployers of other AI systems and models and deployers of high-risk AI systems may exceptionally process special categories of personal data to the extent that:

(a) such processing is strictly necessary to ensure bias detection and correction in view of possible biases that are likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to discrimination prohibited pursuant to Union law, especially where data outputs influence inputs for future operations; and

(b) all of the conditions and safeguards set out in paragraph 1 are applied. This paragraph does not create any obligation to conduct such bias detection and correction.';

This article serves as a **basis of legitimacy to be able to process special categories of data**. In this regard, Article 9(1) of Regulation (EU) 2016/679 (GDPR) allows the processing of special categories of data when an EU or EU Member State regulation has provided for this for reasons of an essential public interest. Thus, the European Regulation on Artificial Intelligence, through Article 4a. " **Processing of special categories of personal data for bias detection and correction**" is the **specific European standard that expressly allows these special categories of data to be processed based on an essential public interest**, in this case, **the detection and correction of biases to mitigate the discrimination made by AI systems with respect to individuals or groups**.

Data protection regulations do not only require that a standard allows special categories of data to be processed. In addition, this regulation in turn must establish different measures and guarantees to protect the data subjects in the framework of such processing.

Below, in order to help the reader properly understand, we detail some of the most relevant security and privacy protection measures:

a) Data anonymization

The European Regulation on Artificial Intelligence establishes that whenever the main purpose of the processing, i.e. the detection and correction of bias, is not significantly affected, this data must be anonymized.

Therefore, the initial premise is that **whenever possible**, when it is intended to use special categories of data to detect bias, they **should be anonymized**.

Where the provider does not anonymize these personal data because it considers that the anonymization of these affects or will significantly affect the bias detection process itself, **the latter must document and justify** how the anonymization process affects that bias detection process.

It is important to note that the European Regulation on Artificial Intelligence establishes that the anonymization process must have a "significant" effect, which means that, in principle, if the anonymization does not affect the bias detection process itself in a very relevant way, the data must be anonymized despite the fact that the detection of bias is less accurate.

Anonymized data fall outside the scope of data protection legislation to the extent that it is possible to objectively demonstrate that there is no material capacity to associate anonymized data with a specific natural person, directly or indirectly, either through the use of other data sets, information or technical and material measures that may exist at the disposal of third parties[2].

On its website, the AEPD makes available to citizens, companies and Public Administrations a whole set of tools and guides on the different processes of anonymization of personal data[3].

b) Pseudonymization of data

Where the provider considers that anonymization of data will significantly affect the detection and correction of bias, this data should be pseudonymized.

The purpose of pseudonymization is to protect personal data by concealing the identity of the persons who are the owners of such data, replacing one or more personal data identifiers with so-called pseudonyms and adequately protecting the link between the pseudonyms and the additional linked information.

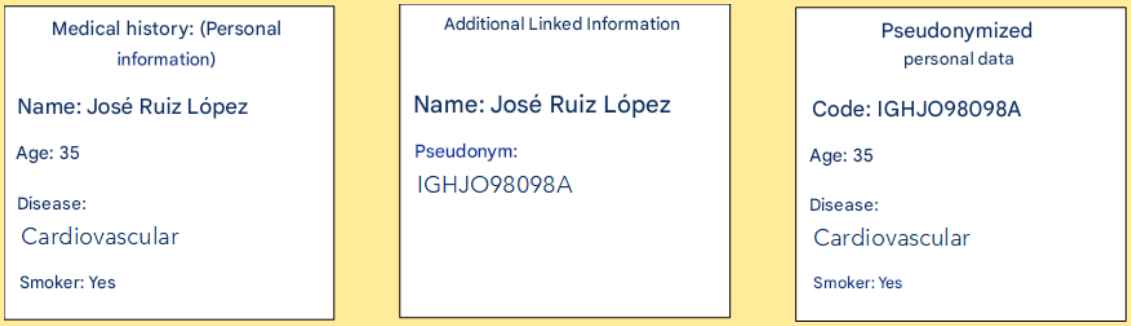
[2] Fountain. Spanish Agency for the Protection of Personal Data. It can be consulted at: https://www.aepd.es/es/prensa-y-comunicacion/blog/anonimizacion-y-seudonimizacion

[3] See the content space dedicated to innovation and technologies and specifically to anonymization processes. It can be consulted at: https://www.aepd.es/es/areas-de-actuacion/innovacion-y-tecnologia

The pseudonymized dataset and the additional information linked to this dataset fall within the scope of the GDPR and the rest of the regulations on the protection of personal data, since, although the direct identification of individuals is reduced, this possibility is not completely eliminated.

A simple example of pseudonymization would be the replacement of a person's identification data, such as name and surname, with a code. In this way, on the one hand, we would have a set of pseudonymized data to which we assign different codes and on the other hand the additional information linked to this set of data.

Example



The safeguards applicable to the pseudonymized dataset and the information linked to that dataset shall include[4]:

- The pseudonymization process itself, which must prevent re-identification without having the additional information.
- Application of data protection principles to pseudonymized data. For example: data retention period, data communication, etc.
- Deployment of security measures to avoid personal data breaches both on the pseudonymized set and on additional information.

c) Other guarantee measures derived from the regulations on the protection of personal data

In addition to the previously indicated measures that are expressly indicated in the European Regulation on Artificial Intelligence, another series of guarantees may be implemented to ensure adequate compliance with data protection regulations with respect to the processing defined in Article 10. "***Data and data governance***" of the European Regulation on Artificial Intelligence Among others, we can highlight the following measures that the Spanish Data Protection Agency has indicated when detecting bias:

- Define procedures to identify and eliminate, or at least limit, biases in the data used to train the model.

[4] For more information on pseudonymization, please consult the content space dedicated to innovation and technologies and specifically to pseudonymization processes. It can be consulted at: https://www.aepd.es/es/areas-de-actuacion/innovacion-y-tecnologia

- Security and access: Identification measures can be established for access to these specific data, and it is advisable to establish a process for managing access rights to these data. Therefore, only certain people in the organization should have access to this data.
- Transmission of data to third parties: It is advisable to encrypt information in transit.
- Verify that there are no previous historical biases in the training data used as input to the model and that, if not, either another source of training data that does not contain them has been chosen, or an appropriate cleaning and purification has been carried out for its normalization.
- Take steps to assess the need for additional data to improve accuracy or eliminate potential biases.
- Implement human oversight mechanisms to control and ensure the absence of bias in the results[5].

In addition to being able to incorporate the techniques detailed in this section, providers must implement all those general guaranteed measures that derive from general compliance with the regulations on the protection of personal data[6].

5.1.2 Detection and/or bias correction of special categories of data in the sandbox framework

Currently, the European Regulation on Artificial Intelligence is a standard in the process of being applied, therefore, while this regulation is not applicable in all its terms until August 2, 2026, Article 10 "***Data and data governance***" may not be used by providers to process special categories of data based on this precept.

That is why, **within the framework of the sandbox**, this precept cannot be put into practice, until the European Regulation on Artificial Intelligence is a current regulation, so that the processing of special categories of data to detect biases based on this precept is not allowed.

That said, it is possible to resort to other ways to try to legitimize the processing of special categories of data in order to detect and/or correct biases, although such processing could not be carried out using the article "***Data and data governance***" of European Regulation on Artificial Intelligence as a reference.

It is important to note that the indications made in this Guide to this precept must be followed with caution. Above all, it is recommended for providers to review the possible modifications that this precept has undergone during the legislative development, as well as the indications and general interpretations that may have been indicated by the national and European authorities in the field of data protection on this article.

[5] Spanish Data Protection Agency. Guide Requirements for Audits of Treatments that include AI. 2021. p. 25.
[6] The Spanish Data Protection Agency has prepared a series of Guides focused on the use of Artificial Intelligence systems and compliance with data protection regulations. These documents are: Guides Adaptation to the GDPR of processing that incorporates Artificial Intelligence. An Introduction (AEPD, 2020) or Requirements for Processing Audits that include AI (AEPD, 2021). Likewise 10 Misunderstandings About Machine Learning (AEPD, 2022).

6. Technical documentation

Article 11 (Technical Documentation) states that the system must be documented in such a way as to demonstrate that it complies with the requirements set out in the regulation, providing the national competent authorities and the notified bodies with the necessary information, clearly and comprehensively, to assess the AI system's conformity with those requirements.

The article further specifies that this documentation shall include, at a minimum, the elements listed in Annex IV. The following outlines the aspects to be documented regarding data governance in accordance with that annex:

> 2. (d) where relevant, the data requirements in terms of datasheets describing the training methodologies and techniques and the training data sets used, including a general description of these data sets, information about their provenance, scope and main characteristics; how the data was obtained and selected; labelling procedures (e.g. for supervised learning), data cleaning methodologies (e.g. outliers detection);
>
> 3. Detailed information about the monitoring, functioning and control of the AI system, in particular with regard to: its capabilities and limitations in performance, including the degrees of accuracy for specific persons or groups of persons on which the system is intended to be used and the overall expected level of accuracy in relation to its intended purpose; the foreseeable unintended outcomes and sources of risks to health and safety, fundamental rights and discrimination in view of the intended purpose of the AI system; the human oversight measures needed in accordance with Article 14, including the technical measures put in place to facilitate the interpretation of the outputs of AI systems by the deployers; specifications on input data, as appropriate;

These two points alone could be considered sufficient to cover the essential documentary requirements of the regulation regarding data and data governance. However, it is considered good practice to expand the documentation by also incorporating and justifying the steps indicated in section 4, along with the additional considerations mentioned in section 5.

These sections detail the elements that must be implemented to meet the requirements of the article, as well as specify how each of them should be implemented. For the documentation of these elements, a document accompanying the system should describe all the data governance measures outlined in this guide that have been implemented. Each implemented measure should be specified and detailed in terms of how it was implemented, and the person responsible for its implementation should be identified.

7. Self-assessment questionnaire

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

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

8. Annexes

8.1 ANNEX A – Methods of Data Collection

Data exists in many forms, and we can create or collect it in different ways. In this section we detail some of the most common data collection methods that we can address [1][3].

- **Automated data collection tools:** In this case, data collection is accomplished using applications within the organization (e.g., email) or externally through a website, mobile application, or similar application. It involves using computer programs to automatically collect data from online data sources. Some methods to automate data collection are: Web-scraping, web crawling, use of APIs, etc.
- **Transactions from other systems:** Data collection or updating performed on other systems may flow into the organization's system through electronic data interchange or other interconnection processes.
- **Sensors:** More and more data is being introduced into the organization through mechanical systems such as sensors. Sensors encompass a wide range of data acquisition devices, such as website records, social media feeds, and Internet of Things devices, including everyday devices, from simple temperature sensors to TVs, cars, traffic lights, and buildings. Sensor data can also include potentially urgent signals, such as alerts and alarms.
- **New contexts:** Data from different reports can be combined with other data to provide additional information, which in turn feeds back into the organization's data. In many cases, this additional data provides new context to the original data and may need to be treated differently. New contextual data can come from decisions that give relevance or value to existing data.
- **Subscription**: Data can be made available to your organization through a subscription to a data source or virtual data warehouses.
- **Public crowdsourcing:** consists of collecting data through the help or contribution of independent people. For example, if we need to collect images of the streets to train a traffic sign recognition system, we could obtain them through public crowdsourcing by creating a platform for it and providing certain instructions to the participants. This method is commonly used for data labelling.
- **Private crowdsourcing:** it is a method similar to public crowdsourcing, the difference is that the contributions in this case will be made by groups of private specialists who will be hired to carry out this task.

8.2 ANNEX B – Data quality

8.2.1 ANNEX B.1 – Data Quality Dimensions

Data dimensions and quality checks are used to specify and evaluate data quality. Generally, more than one data quality control is available to assess the quality associated with each dimension. In this Annex we detail the main dimensions of quality around data

[6]. Some examples of data quality checks for each of these dimensions are set out in Annex B.2.

1. Accessibility: refers to the degree to which data can be accessed in a specific context of use, particularly by people who need special assistive technology or settings due to a disability.

2. Auditability: refers to the degree to which all or part of the dataset has undergone an audit review or, failing that, the availability of all or part of the dataset to undergo an audit review.

For example, suppose a dataset used to train an image recognition AI system has been tagged by a hired third-party specialist. In this situation, to ensure correct tagging, we might hire an independent third party to audit a subset of the tagged images.

3. Identifiability: refers to the ability to identify a principal of personally identifiable information (PII) directly or indirectly on the basis of a given set of PII. It is important to understand whether any PII in a dataset can be used to identify a PII principal, as legal requirements in some jurisdictions may restrict such activity. De-identification or anonymization processes can be applied to training data, validation, and testing to reduce the possibility of identification.

For example, suppose we have an AI system developed to show specific advertising to each user, trained by search engine queries. The data set includes the user's IP address, which is considered PII in some jurisdictions, including the Spanish one by the doctrine of the Supreme Court and the CJEU. Anonymization is applied to the dataset to remove the IP address before the dataset is split into training datasets, validation, and testing, thereby trying to reduce the possibility of identification.

4. Portability: refers to the ability to move data from one system to another, within a specific context, preserving its quality.

For example, let's say that for the development of our AI system we need to process the data on multiple platforms or systems. We need to maintain data quality when we transfer it from one platform or system to another.

5. Understandability: refers to the ease with which data can be interpreted by the organization. AI systems are based on mathematical models and constructions whose main language is based on numbers and algorithms. If these could not be interpreted by users, they could lose their function and therefore it is a relevant dimension of quality.

6. Temporal coherence: refers to the alignment of the data with the reality they currently represent in terms of their age.

For example, data on people may be incomplete for underrepresented populations prior to changes in regulations and social norms. On the other hand, AI systems based on economic data collected over several decades can be incorrect if the data is not corrected for inflation, exchange rates, and other factors that vary over time.

7. Effectiveness: refers to the extent to which the dataset meets a series of requirements for its use in the development of the specific AI system:

- For a computer vision system, the effectiveness of the dataset may be the lowest acceptable ratio at which the number of images with brightness or resolution is below a required threshold over all samples in the dataset.
- For an image classification system, the effectiveness of the dataset may refer to the lowest acceptable proportion of the number of images belonging to a class over the total number of samples in the dataset.
- For an object detection system, the effectiveness of the dataset can refer to the lowest acceptable proportion at which the number of images with bounding box numbers or areas is below a required threshold across all samples in the dataset.

8. Efficiency: refers to the degree to which data has attributes that can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use.

9. Accuracy: refers to the degree to which each piece of data in the dataset has the correct value. In other words, it is the degree to which the data has attributes that correctly represent the true value of the predicted attributes. We can differentiate, mainly, two types of accuracy:

- Syntactic accuracy that considers the proximity of data to a syntactically correct dataset in a relevant domain
- Semantic accuracy that considers the proximity of data to a semantically correct dataset in a relevant domain

A data element is syntactically correct if its data value has the same type as its explicit data type, and semantically correct if its data value has an expected value corresponding to the ML (*Machine Learning*) task. ML models are mathematical constructs, which means that the low syntactic or semantic accuracy of the data in training data, validation, or test can cause the model itself to be incorrect or the inferences made by the model to be incorrect.

For a supervised learning classification system, correcting labels can affect the inference accuracy of a model trained from data. Factors to consider in measuring labelling accuracy include:

- Correcting Category Tags
- Fix labelled bounding boxes

10. Completeness: refers to the degree to which the data presents values for all characteristics. Incompleteness of datasets can have a serious impact on the development of the AI system.

The completeness of labelled data in a dataset is relative. In different scenarios, the meaning of completeness may be different and should be considered with specific usage contexts. Factors to consider include:

- The completeness of a dataset that is used for ML-based image classification should check for unlabelled samples in a dataset, which cannot be used directly in supervised ML.
- The completeness of a dataset that is used for ML-based object detection should check for the incompleteness of the bounding boxes labelled on objects.

In particular, it is common in real life for a sample to have multiple objects in several categories, as it is difficult to capture a scene with a single isolated object taking up the entire viewing space. In this case, to measure the completeness of the dataset for an ML-based image recognition, you should consider whether:

- Any target object exists in a sample
- All target objects are categorized
- All detected target objects are tagged with bounding boxes or other methods

11. Regulatory compliance: refers to the degree of compliance with regulations, standards, regulations, or other compliance rules around data. For example, personal data used for the development of AI systems may be subjects to legal and regulatory requirements.

12. Credibility: refers to the degree to which data has attributes that users consider true and credible in a specific context of use. Credibility is applicable for individual data, for related data in a data record, and for the entire data set. The context in which data is used can affect its perceived veracity and credibility. Data may be disturbed during processing by authorized and unauthorized parties.

An emerging concern for ML is that unauthorized parties disrupt training data, validation, testing, and production to deliberately render trained models unusable or to manipulate inferences made by a trained model.

Processes used in data preparation can change data without changing its meaning (e.g., normalization, imputation, segmentation, or combination of characteristics). In these cases, the data maintains its credibility.

13. Balance: refers to the distribution of samples for all aspects of the dataset. For example, if the dataset represents X number of item categories, the number of samples per category must be evenly distributed for this dataset to balance. For an image dataset, these aspects can include category labels that are significant for business logic, figure resolution, figure brightness, the width/height ratio of labelled bounding boxes, the size of labelled bounding boxes, and any other aspects that may influence the performance of the machine learning model.

The balance of a dataset can affect a portion of the overall performance of a machine learning model. For an ML-based machine vision system, the balance of the data set must be considered.

- When there are considerable differences in brightness or resolution between samples in a training data set and real-world data, ML models can fail due to noisy data introduced due to weakness or vagueness.
- In an ML-based classification system, the presence of an unbalanced category of sample population can result in the failure of rare instance discovery and classification. Such cases may even be misclassified or identified as noisy data.
- In an ML-based object detection system, significant differences in the width/height ratio or size of bounding boxes can result in size inconsistency over the objects to be detected, given a fixed receptive field size. Consequently, this can also cause a

loss of generalization, if additional multi-size object verification or adjustment approaches are not applied.

14. Consistency: refers to the degree to which the same data represented in different sources and different data that are meaningfully related do not present contradictions or inconsistencies. Consistency is a key aspect of the data used for the development of AI systems, as the features used in the training data must together provide a model that allows correct inferences in the production data. In addition, ML can be literal in its interpretation of the data. Duplicate records can cause an overweight of certain characteristics. Contradictions between the characteristics of the training data can cause a trained model to perform below requirements.

The distribution of data for each characteristic can be considered a measure of consistency. In some cases, ML models require data with a normal distribution to meet performance requirements.

15. Diversity: refers to the difference between samples in terms of the target data. In a dataset used for an AI system, a proper difference between samples is important. If all or most of the records in a dataset are the same, a machine learning model trained from that dataset may have the risk of overfitting and consequently less generalizable. The diversity of a dataset represents a degree of how the dataset contains multiple value domains, labels, clusters, and distributions among individual data. Improving data using generative machine learning models can improve data diversity, but these approaches can fail if the diversity of the original dataset is limited. Diversity is closely related to representativeness and balance. It is a data quality feature that can be used to evaluate the fidelity of a dataset.

Diversity measurement can be performed in the context of target-specific data, as determined by the ML task requirements.

16. Relevance: refers to the degree to which a dataset is appropriate for a given context.

For an AI system, relevance can mean that the characteristics selected in the training data and their values are good predictors for the target variable. For example, suppose an AI system is used to determine the creditworthiness of people. The training data are representative of the sample of the population that is expected to appear in the production data. Training data includes relevant characteristics such as previous credit history, income, job tenure, and net worth, which are good predictors of creditworthiness. The training data also includes each person's height and weight. Statistical tests show no correlation of height and weight with past credit history and are considered poor predictors of future credit performance. To improve the overall relevance of the dataset, height and weight characteristics are removed.

17. Representativeness: refers to the degree to which a sample reflects the target population being studied. For supervised ML, the training data set can be thought of as the sample and the production data as the study population. When the training data does not sufficiently represent the production data, the trained AI system may not function as required.

Data representativeness is related to relevance in the sense that a dataset that does not represent the population under study is unlikely to provide good predictors for the target

variable. For example, for a facial recognition application, a training data set that does not contain images of people with dark skin is unlikely to result in a trained AI system that works properly on production data that includes images of people with dark skin.

18. Similarity: refers to the degree of similarity between samples in terms of the characteristics concerned. This is relevant to classification tasks that are typically implemented using supervised learning. This is also relevant for grouping tasks that are typically implemented using unsupervised learning. Both sorting and clustering tasks require an appropriate level of difference between samples to function successfully.

An AI system trained on a dataset containing fairly similar images (e.g., which are generated by a slight pixel shift based on a few seed images) may have the risk of overfitting and consequently less generalization. In this case, it is possible to consider applying data enhancement approaches, such as rotation and scrolling. which can improve the generalizability of the AI system. These approaches cannot work if the number of seed images is limited. In this case, the proportion of similar samples should be checked. Another approach is to consider clustering algorithms with topic drift mitigation methods.

Other measurements identify data similarity through the geometric approach: that is, a dataset of N records and characteristic M, can be represented as N vectors in an M-dimensional space, so that it can be analysed and compared using the tools of geometry. In particular, similarity can be associated with the mutual position of vectors in space.

8.2.2 ANNEX B.2 – Data quality controls

This Annex provides some examples of data quality checks for each of the quality dimensions described [6].

1. Accessibility

- **User accessibility:** The degree to which users find the values in the data accessible.
  - **X = A/B**, where:
  - **A=**number of data relevant to the user's task that is easily accessible.
  - **B=**number of data relevant to the user's task whose access is required.
- **Data format accessibility:** The degree to which a specific device allows accessibility.
  - **X = 1- A/B**, where:
  - **A=**number of data that are not accessible due to their format.
  - **B=**number of data for which an accessible format can be defined.

2. Auditability:

- **Records audited:** The proportion of records in the dataset that have been audited.
  - **X = A/B**, where:
  - **A=**number of audited records.
  - **B=**total number of records in the dataset.
- **Auditable Records:** The proportion of the records in the dataset that are available to be audited.
  - **X = A/B,** where:
  - **A=**number of records available to be audited.

- **B=**total number of records in the dataset.

3. Identifiability:

- **Identifiability Index:** The proportion of records in the dataset that can be used for identification.
  - **X = A/B**, where:
  - **A=**number of records containing data that can be used for identification.
  - **B =** total number of records in the dataset.

4. Portability:

- **Data portability ratio:** The proportion of data that preserves its quality after portability.
  - **X = A/B**, where:
  - **A=**number of data that preserves its quality after portability.
  - **B=**number of data that have been ported.

5. Comprehensibility:

- **Understandability of symbols:** The degree to which comprehensible symbols are used.
  - **X = A/B**, where:
  - **A =** number of data represented by known symbols.
  - **B =** number of data for which it is necessary to understand the symbols.
- **Semantic comprehensibility:** Proportion of terms defined in the data dictionary.
  - **X = A/B**, where:
  - **A=**number of data defined in the data dictionary using a language understood by users.
  - **B=**number of data defined in the data dictionary.
- **Understandability of data values:** The degree to which data values are understood by users.
  - **X = A/B,** where:
  - **A =** number of data easily understandable by users.
  - **B=**number of data that users need to understand in a particular context.
- **Comprehensibility of data representation:** The degree to which data is represented in a way that is understandable to users in a specific system.
  - **X = A/B**, where:
  - **A=**number of data considered understandable in a data model.
  - **B=**number of data in the data model.

6. Temporal coherence:

- **Temporal consistency of features:** The proportion of data for a feature in the dataset that is within the required time range.
  - **X = A/B**, where:
  - **A=**number of data of the characteristic that are within the required time range.
  - **B=**number of data in the characteristic.

- **Temporal consistency of records:** The proportion of records in the dataset where all the data in the record is within the required time range.
  - **X = A/B**, where:
  - **A=**number of records that are within the required time range.
  - **B=**total number of records.

7. Effectiveness:

- **Brightness Efficiency:** The ratio of samples to unacceptable brightness in an image dataset.
  - **X = A/B**, where:
  - **A=**number of samples with unacceptable brightness.
  - **B=**total number of samples.
- **Resolution Efficiency:** The proportion of samples with unacceptable resolution in an image dataset.
  - **X = A/B**, where:
  - **A=**number of samples with unacceptable resolution**.**
  - **B=**total number of samples**.**
- **Category size effectiveness:** The proportion of categories where the number of categorized samples are below a threshold.
  - **X = A/B**, where:
  - **A=**number of categories where the categorized samples are below a threshold.
  - **B=**total number of categories.

8. Efficiency:

- **Data format efficiency:** Proportion of space unnecessarily occupied due to the definition of the data format.
  - **X = 1-A/B**, where:
  - **A=** Size (in bytes) of the record of a data file that is unnecessarily occupied due to the definition of the data format.
  - **B=** total size (in bytes) of the record in a data file due to the definition of the data format.
- **Data processing efficiency:** Lost labour time due to data representation (data formatting).
  - **X = 1-A/B**, where:
  - **A=**time lost due to data representation.
  - **B=**total processing time.

9. Accuracy:

- **Data syntactic accuracy:** The relationship of closeness of data values to a set of values defined in a domain.
  - **X = A/B**, where:
  - **A =** number of data that present syntactically accurate values.
  - **B=**number of data for which syntactic accuracy is required.

- **Semantic Data Accuracy:** The relationship between the semantic accuracy of data values in a specific context.
  - **X = A/B**, where:
  - **A =** number of semantically precise data.
  - **B=**number of data for which semantic accuracy is required.
- **Risk of dataset inaccuracy:** The number of *outliers* indicates a risk of data inaccuracy.
  - **X = A/B**, where:
  - **A=**number of *outliers*.
  - **B=**total number of data in the dataset.
- **Data model accuracy:** The data model describes the system with the required accuracy.
  - **X = A/B**, where:
  - **A=**number of data in the model that are accurately described.
  - **B=**total number of data used in the development of the model.
- **Class labelling accuracy:** The proportion of samples within incorrect classes in a dataset.
  - **X = A/B**, where:
  - **A=**number of samples, in which each one is labelled with an incorrect class.
  - **B=**total number of samples.

10. Completeness:

- **Completeness of value:** The proportion of data with no nulls in a dataset.
  - **X = A/B**, where:
  - **A =** number of data whose value is not empty.
  - **B=**total number of data.
- **Completeness of value occurrence**: The ratio of the number of occurrences of a given data value to the expected number of occurrences of the value in data with the same domain in a dataset.
  - **X = A/B**, where:
  - **A=**number of occurrences of a given data value
  - **B=**expected number of occurrences of the value in data with the same domain in a dataset.
- **Feature completeness:** The proportion of data with no nulls for a given feature in a dataset.
  - **X = A/B**, where:
  - **A =** number of data with no nulls for a given feature in a dataset.
  - **B=**total number of data associated with that characteristic.
- **Completeness of the record:** The ratio of records with no presence of empty data in a dataset.
  - **X = A/B**, where:
  - **A=**number of records that have all the data reported.
  - **B=**total number of data .
- **Category Label Completeness:** The proportion of samples with unlabelled or incompletely labelled categories in a dataset.
  - **X = A/B**, where:

- **A=**number of samples with unlabelled or incompletely labelled categories.
- **B=**number of samples.

11. Regulatory compliance:

- **Data Element Compliance:** The degree to which data meets regulatory compliance requirements.
  - **X = A/B**, where:
  - **A=**number of data that meet compliance requirements.
  - **B=**total number of data

12. Credibility:

- **Credibility of values:** The degree to which the elements of information are considered true, real and credible.
  - **X = A/B,** where:
  - **A =** number of data whose values have been satisfactorily validated by a specific process.
  - **B=**number of data that have been subjected to this process.

- **Source credibility:** The degree to which values are provided by a qualified organization.
  - **X = A/B,** where:
  - **A=**number of data validated by a qualified organization.
  - **B =** number of data for which the credibility of the source can be defined.
- **Data Dictionary Credibility:** The degree to which the data dictionary provides credible information.
  - **X = A/B,** where:
  - **A =** number of data in the dictionary whose values have been satisfactorily validated by a specific process.
  - **B=**total number of data in the dictionary.
- **Data model credibility:** The degree to which the data model provides credible information.
  - **X = A/B,** where:
  - **A=**number of data in the model whose values have been satisfactorily validated by a specific process.
  - **B=**total number of data in the model.

13. Balance:

- **Brightness Equilibrium:** The maximum ratio of the difference in brightness of an image sample to the average luminosity of the samples in a dataset.
  - **X = A/B**, where:
  - **A =** average brightness value of the samples.
  - **B=** maximum value of the absolute differences between the brightness value of each image in the sample and A.
- **Resolution balance:** The maximum ratio between the difference in resolution of an image sample and the average resolution of the samples in a dataset.

  - **X = A/B**, where:
  - **A =** average resolution value of the samples.
  - **B=** maximum value of the absolute differences between the resolution value of each image in the sample and A.
- **Image Balance Between Categories:** The maximum ratio between the difference in category size (number of samples contained) and the average category size of a dataset.
  - **X = A/B**, where:
  - **A =** average size of the dataset category.
  - **B=**maximum value of the absolute differences between the size of each category in the dataset and A.

14. Consistency:

- **Data recording consistency:** The proportion of duplicate records in the dataset.
  - **X = A/B**, where:
  - **A=**number of duplicate records in the dataset.
  - **B=**total number of records.
- **Data format consistency:** Consistency of data format for the same data element.
  - **X = A/B,** where:
  - **A=**number of data where the format of all features is consistent across the different locations where they are stored.
  - **B=**number of data for which a given format has been defined.
- **Semantic consistency:** The degree to which semantic rules are respected.
  - **X = A/B**, where:
  - **A =** number of semantically correct data.
  - **B=**number of data for which semantic rules have been defined.

15. Diversity:

- **Tag Richness:** The number of different tags in a dataset.
  - **X = A**, where:
  - **A=**number of different labels in the dataset.
- **Relative abundance of labels:** The proportion of the number of data that the same label has in a dataset.
  - **X = A/B**, where:
  - **A=**number of data sharing the same label in a dataset.
  - **B=**number of total data.

16. Relevance:

- **Feature relevance:** Proportion of features in the dataset that are relevant to the given context.
  - **X = A/B**, where:
  - **A=**number of data considered relevant in a given context.
  - **B=**number of total data.
- **Relevance of the record:** Proportion of records in the dataset that are relevant to the given context.

- **X = A/B**, where:
- **A=**number of records considered relevant in a given context.
- **B=**number of total records.

17. Representativeness:

- **Representativeness ratio:** Proportion of relevant attributes found in a sample with respect to the attributes found in the sample.
  - **X = A/B**, where:
  - **A=**number of relevant attributes found in a sample.
  - **B=**number of attributes in the sample.

18. Similarity:

- **Sample Similarity:** Proportion of similar samples in a dataset (the lower, the better).
  - **X = A-B/A,** where:
  - **A=**number of samples in the dataset.
  - **B=**number of sample groups processed by a pooling algorithm.
- **Sample independence:** Ratio between Principal Component Analysis (PCA) and the size of the dataset.
  - **X = 1-A/B**, where:
  - **A = number of** main components.
  - **B=total number** of components (dimension).

8.3 ANNEX C – Biases

8.3.1 ANNEX C.1 – Sources of bias

In this Annex we detail some of the most common sources of bias that can affect the desired functioning of our AI system [4]:

1. Inherent biases in the data: one of the most common sources of bias is that introduced by the data used to develop AI systems, at some stage of their life cycle. Below, we will analyse the different types of biases that we can differentiate in this category.

1.1 Bias in data selection: this type of bias is associated with the mismatch between the selected data and its real-world distribution. This may be closely related to the human cognitive bias introduced into the data selection process.

1.1.1 Data sampling bias: This type of bias is related to the non-random collection of data within a population. For example, if we train an image recognition system for the development of autonomous cars only with images within a city, it may not be able to react appropriately in other environments (a highway, a national road, a road, etc.).

1.1.2 Coverage bias: This type of bias is related to a mismatch between the population represented in the dataset against which the AI system has been trained and the population represented in the dataset on which the predictions are being made. For example, if we train a music recommendation system by interviewing only people in the age range between 65 and 80, we will be introducing a coverage bias.

1.1.3 Participation bias: this type of bias is generated when a given population group has a very different participation rate in a survey than another group. For example, suppose we want to train an AI system to predict the outcome of an election. To do this, we collected data by interviewing the population and it turns out that we only managed to collect information from voters of one ideology since those of the opposite ideology refuse to participate in the survey.

1.2 Bias in data labelling: the data labelling process itself can introduce biases when trying to discretize certain variables in the dataset. For example, if we try to label a dataset based on the variable "age" using the categories "young", "adult" and "elderly" we will be discretizing the population of this dataset into categories that will not always be a faithful reflection of the reality they represent. On the other hand, the labelling process itself can be affected by cognitive biases of the person who performs this process.

1.3 Data incompleteness bias: in many cases the data we collect may be incomplete in some of its characteristics. This fact is very common when we use data collected from the real world. If there is a considerable imbalance of completeness between different populations, our data could be skewed. For example, the medical data available on patients often varies greatly between population groups of different purchasing power, especially in societies where healthcare is private.

1.4 Bias in data preparation: Certain actions or decisions in the data preparation stages may introduce bias. For example, in the stage of measuring and improving quality we may find many unreported values of a characteristic that we decide to impute. Approaching this imputation process and how we choose to approach it can involve introducing bias into our dataset.

1.5 Simpson's paradox: This paradox poses a scenario in which a trend present in several (separately) data sets disappears and is reversed when these groups are combined. An example that is usually presented to illustrate this situation is the comparison of the mortality rates of two hospitals, which can globally favour hospital A over B, and yet when analysed by procedures it is discovered that the sign of difference changes, because patients with a worse prognosis and more serious pathologies are admitted to hospital B more frequently.

1.6 Confounding variables: A confounding variable is a variable that influences both the dependent variable and the independent variable causing a spurious association. Because of this, a perceived relationship between two variables could be shown to be partially or totally false.

1.7 Non-normality: Most statistical methods assume that data are subjects to a normal distribution. However, if the data is subjects to a different distribution (e.g., Chi-Square, Beta, Lorentz, Cauchy, Weibull, or Pareto), the results can be biased and misleading.

1.8 Aggregation of the data: Aggregating data covering different groups of objects that may have different statistical distributions can introduce biases into the data used to train AI systems. This could be caused by human cognitive biases, such as the homogeneity bias outside the group.

1.9 Other sources of bias in the data: Data can also be biased by external disruptive influences. This bias would be considered by an AI algorithm as part of the model to be generalized and, therefore, would lead to undesired results. An example is the "*outliers*" or extreme values that represent very low probability events. Another example is the noise components, which are caused by stochastic processes and cannot be described in a deterministic way. Noise can have a negative influence on the model if overfitting occurs. In addition, artificially generated noise can be used to create contradictory examples that will cause undesired results.

2. Human cognitive biases: Thinking is often based on opaque and subjective processes that lead us to make decisions without always knowing what leads to them. These human cognitive biases can affect decisions about data collection and preparation and the development and use of the AI system.

2.1 Automation bias: This type of bias is generated when the decision-maker favours recommendations made by an automated decision-making system over information made without automation, even when automation makes mistakes.

2.2 Group attribution bias: This type of bias is generated when we assume that what is true for one sample is also true for all samples in that group. The effects of group attribution bias can be exacerbated if a convenience sample is used for data collection. In a non-representative sample, attributions may be made that do not reflect reality.

2.3 Implicit bias: This type of bias is generated when we make associations or assumptions based on our mental models and memories. For example, for a wedding photo identification image classifier to look for the presence of white dresses, because not in all times and cultures the wedding dress is white.

2.4 Bias within the group: this type of bias is generated when we show bias towards the group to which we belong. It is quite common to find tendencies to make more internal (dispositional) attributions for events that reflect positively on the groups to which we belong and more external (situational) for events that reflect negatively on these groups.

2.5 Homogeneity bias outside the group: this type of bias is generated when we have the perception that the members of the external group have closer similarities than those of our own group. For example, Europeans might be perceived as a more homogeneous group by Americans and vice versa.

2.6 Self-confirmation bias: this type of bias is generated when we tend to gather information with the sole intention of confirming our beliefs or hypotheses by ignoring opposing arguments that may contradict them.

2.7 Social bias: This type of bias is generated when one or more similar cognitive biases (conscious or unconscious) are being held by many individuals in society. It manifests itself in AI systems when they learn or amplify pre-existing historical patterns of bias in datasets. Social bias also manifests itself when cultural assumptions are applied to data without taking into account cross-cultural variation. For example, when historical data is inappropriate for the inferences being made, possibly reinforcing common but inaccurate social views.

8.3.2 ANNEX C.2 – Bias assessment techniques

In the development of an AI system, it is very important to consider potential biases that could lead to unwanted and inequitable system behaviours. One way to identify potential biases is by evaluating the results of the AI system using metrics developed for this purpose. These metrics try to evaluate the differences between the observed average values and the true values [4].

Most of the work developed around the aforementioned metrics has focused on AI systems based on classification or regression with respect to groups defined in terms of one or more biographical, demographic or behavioural attributes. This section presents the most relevant approaches to assessing the bias of classification-based AI systems.

1. Rating systems: Bias in rating systems can be detected by measuring different types of errors with respect to various groups. The approach of dividing data into training, validation, and testing sets is complemented by subdividing each of those sets based on the characteristics for which the system is expected to be fair. If there are multiple features relevant to detecting potential biases in a particular system, then those features could be considered independent or intersectional. For example, a system that is impartial with respect to gender and race independently might be biased toward a specific combination of the two.

Once the data has been properly divided, bias metrics are calculated in each group and comparisons are made between related groups. A classification system could be "unbiased" with respect to relevant characteristics if measurements based on intergroup metrics are within a sufficiently small delta.

2. Reinforcement learning systems: The bias in reinforcement learning systems differs from that of classification systems. In this case, the bias focuses on the series of actions that the system is taking. While testing the bias of reinforcement learning systems is a less developed practice, some methods are emerging. Reinforcement learning systems are usually trained based on a measure of usefulness by following their actions. Bias metrics can be applied to this usefulness measurement across all groups.

In addition to using utility as an analogue to accuracy, bias in continuous learning systems that use reinforcement learning can manifest as different "*path attractors*" for different groups. This could be particularly the case in reinforcement learning systems that make content recommendations. In such systems, *"ground truth"* is a less defined concept. If, for example, an AI assistant to guide college students throughout their course of study consistently led women toward care professions and men toward engineering professions, then that system is potentially biased regardless of utility metrics.

3. Confusion matrix: A confusion matrix (Figure 1) is a method used to uncover classifier bias. It reports the number of false positives, false negatives, true positives and true negatives and includes other performance criteria derived from these values. Since a confounding matrix contains and compares multiple metrics, it allows for detailed analysis of a classifier's performance and is useful for circumventing or uncovering weaknesses in individual metrics. For example, using individual metrics to measure a classifier's

performance would result in misleading results if the underlying datasets were unbalanced and therefore unable to provide reliable results.

		true condition				
	total population	condition positive	condition negative	Prevalence = $\frac{\sum condition\ positive}{\sum total\ population}$	Accuracy (ACC) = $\frac{\sum TP + \sum TN}{\sum total\ population}$	
predicted condition	prediction positive	True Positive (TP) Power	False Positive (FP) Type I error	Positive Predictive Value (PPV), Precision, Relevance = $\frac{\sum TP}{\sum prediction\ positive}$	False Discovery Rate (FDR) = $\frac{\sum FP}{\sum prediction\ positive}$	
	prediction negative	False Negative (FN) Type II error	True Negative (TN)	False Omission Rate (FOR) = $\frac{\sum FN}{\sum prediction\ negative}$	Negative Predictive Value (NPV) Separation Ability = $\frac{\sum TN}{\sum prediction\ negative}$	
		True Positive Rate (TPR) Sensitivity, Recall, Probability of Detection = $\frac{\sum TP}{\sum condition\ positive}$	False Positive Rate (FPR) Fall-out, Probability False Alarm = $\frac{\sum FP}{\sum condition\ negative}$	Positive Likelihood Ration (LR+) = $\frac{TPR}{FPR}$	Diagnostic Odds Rate (DOR) = $\frac{LR+}{LR-}$	F_1 score = $\left(\frac{TPR^{-1} + PPV^{-1}}{2}\right)^{-1}$
		False Negative Rate (FNR) Miss Rate = $\frac{\sum FN}{\sum condition\ positive}$	True Negative Rate (TNR) Specificity, Selectivity = $\frac{\sum TN}{\sum condition\ negative}$	Negative Likelihood Ration (LR-) = $\frac{FNR}{TNR}$		

Figure 1 – Confusion matrix and derived ranking performance metrics [17]

4. Equality of probabilities: also known by its term *"equalized odds"* means that the decisions of an algorithm are independent of a category *A* given an input *Y*. A predictor $\hat{Y}$ satisfies the equal probabilities with respect to category *A* and output *Y*, if $\hat{Y}$ and *A* are conditional independent of *Y*:

$$P(\hat{Y} = \hat{y} | Y = y, A = m) = P(\hat{Y} = \hat{y} | Y = y, A = n), \forall Y, m, n \in A$$

This implies that true positive rates (TPRs) are the same across all demographic categories and false positive rates (FPRs) are the same across all demographic categories.

We must consider that this definition allows models to take into account demographic information. TPR equals 1 – False Negative Rate (FNR), so this also encourages equal false negative rates across all demographic categories. To see the trade-offs between false negatives and false positives, comparing the false negative rate and the false positive rate might help.

5. Equality of opportunity: also known by its term " *equality of opportunity*" means that the decisions of an algorithm where $\hat{Y}=1$ are independent of a category *A* given the input *Y=1*.

A binary predictor $\hat{Y}$ satisfies equality of opportunity with respect to *A* and *Y* if $\hat{Y}=1$ and *A* are conditionally independent of *Y=1*. Formally:

$$P(\hat{Y}=1|Y=1, A=m) = P(\hat{Y}=1|Y=1, A=n), \forall m, n \in A$$

This implies equal true positive rates (TPRs) across all demographic categories.

6. Parity: Statistical parity means that there are equal prediction rates between categories. Demographic parity says that there are equal prediction rates among demographic categories, such as race. Demographic parity, which is a case of statistical parity, means that a decision, such as accepting or rejecting a loan application, must be independent of a demographic attribute. Formally, given demographic variable A:

$$P(\hat{Y} = \hat{y}|A = m) = P(\hat{Y} = \hat{y}|A = n), \forall\, m, n \in A$$

Parity does not capture cases where the exit decision is correlated with one of the groups or attributes being evaluated, and there is no guarantee that the predictions made will be equally good for each category.

7. Predictive Equality: Predictive equality implies equal false positive (FPR) rates across all demographic categories. Formally:

$$P(\hat{Y} = 1|Y = 0, A = m) = P(\hat{Y} = 1|Y = 0, A = n), \forall\, m, n \in A$$

8.3.3 ANNEX C.3 – Bias treatment measures

In this Annex we detail some of the most common bias treatment measures that can be implemented to mitigate the impact of the possible biases we have identified and affect the desired functioning of our AI system. The measures detailed below go beyond the context of the data, it is an approach that tries to encompass the treatment of biases at different stages of the life cycle of an AI system [4]:

1. Analysis of the context and system requirements

Analysing system requirements is an important activity to mitigate bias. It is the stage where internal and external requirements are analysed, system stakeholders are determined, and system objectives are evaluated. For this milestone, the risks posed by the system have been identified, the impact for the identified stakeholders has been assessed and the levels of stakeholder involvement have been defined.

1.1 External requirements

Identifying external requirements as part of the system analysis activity is a normal part of the system development and procurement lifecycle. During this process, particular attention could be paid to the following regulatory frameworks:

- International human rights and equality instruments.
- Specific laws and guidance related to the provision of technical solutions.
- Data protection and privacy legislation could include provisions relating to automated decision-making.
- Competition and business law.

Below are examples of possible types of obligations for the responsible entity:

- The need for a risks assessment, which could include social concerns from the perspective of affected stakeholders.

- Notification to users that they are subjects to an automated decision, the requirement to obtain explicit consent, and to provide a non-automated alternative when consent is not given.
- Ensure a certain level of auditability or explainability in the solution, to support the analysis of a particular decision or event.
- Activities to quantify or mitigate risks, such as collecting metadata on data sources to understand provenance and quality.
- Provision for the meaningful participation of a human being in the decision-making process.

The equivalent provision and price of services for groups of people with certain characteristics. This might require the ability to demonstrate that equality is achieved in practice.

1.2 Internal requirements

In addition to regulatory requirements, many other factors could contribute to a stakeholder's desire to mitigate bias, such as:

- Goals, strategies, and internal policies of an organization.
- Moral or cultural values.
- Avoid social concerns or reputational damage.

The analysis process may pay particular attention to five specific areas: inclusion of transdisciplinary experts; identification of stakeholders; the selection of data sources; external change; and specification of acceptance criteria, including acceptable levels of bias.

1.3 Transdisciplinary experts

While unwanted bias is a relatively new topic in the context of technology, it is a well-understood topic in the social sciences. As part of the requirements analysis process (and, indeed, of the entire system lifecycle), it is relevant to consider the expertise that might be required to fully mitigate societal concerns about bias and to take into account diverse perspectives. This could include:

- Social scientists and ethicists;
- Data scientists and quality specialists;
- Legal and data privacy experts;
- User representatives or external stakeholder groups.

For example, designers of a facial recognition system might give importance to the facial contour feature in their design and overlook the fact that the contour might be (partially/completely) covered for people with particular cultural/religious backgrounds. A sufficiently diverse team is more likely to identify such limitations with designs, assumptions, and data sets.

1.4 Identification of stakeholders

Traditional requirements analysis includes identifying stakeholders. However, in order to comply with the aforementioned aspects of the regulatory frameworks and adequately

mitigate societal concerns, it might be necessary to broaden this traditional definition of stakeholder to include those directly and indirectly affected by the system in place.

Depending on the types of data being used to make automated decisions, it might be necessary to further decompose stakeholder lists into groups of people who might be affected differently by bias in the system, have different skills in using the system, or have different levels of knowledge and access. It is important to consider what biases, negative experiences, or discriminatory outcomes might occur.

These can be considered through a variety of participatory design or ethnographic methods, which involve active outreach and discussion with affected groups. It is usually insufficient to indicate who could theoretically be affected without the direct input of these groups. Often, it's also not enough for a member of the affected group (who may also work as part of the design team) to assess how that group is affected. Groups are not monoliths, and a person cannot always adequately represent the range of possible perspectives.

Stakeholder identification and engagement could be included in a formal description and documentation of anticipated areas of concern and potential consequences for affected groups, positive and negative. From these, more qualified and quantitative requirements can be derived. A human impact assessment conducted in later phases can review these areas of concern and assess whether the concern was successfully mitigated.

1.5 Selecting and Documenting Data Sources

The selection of data used to form explicit rules within an expert rules-based system, or data that is used to train ML models, is an essential activity that has a significant influence on bias.

In the case of explicit knowledge, it is important to consider the human cognitive biases already present in those who specify knowledge. Human cognitive bias could be present in a human judgment, which is then encoded in a rules-based system, and then propagated throughout the system's lifetime on a larger scale. While it might be considered acceptable for such a cognitive bias to be present in a single human decision, propagating that bias through automated decision-making could have a much greater impact.

AI statistical systems that learn from data without explicit knowledge being specified suffer from many risks. The individual data source could be considered to determine:

- **Integrity**. A data source that excludes certain records because they do not contain the same characteristics for all records might not provide a complete picture and cause defects in the training process. Publicly available data (e.g., from the Internet) are unlikely to have an equivalent distribution among groups of people.
- **Accuracy**. A data source that contains inaccurate data will propagate those inaccuracies in a machine learning model. This could lead to general accuracy issues, but these issues could also be skewed towards certain groups of people, for example, those with less credit history.
- **Collection procedures**. It's important to understand the lineage of data, how it's collected, how it's entered, and whether these processes affect completeness and accuracy. The location and environment from which the data is obtained can be considered.

- **Temporal coherence.** The frequency with which data is collected or updated may be relevant to ensure its accuracy. Conversely, the upgrade may require reassessment. A system that passes an audit can be out of compliance, especially if the updates come from a different process than the initial data.
- **Consistency**. The consistency with which input data (or labels) are determined can be important. For example, if a human is categorizing items that don't have a clear boundary between categories, different categories or labels could result from the same data.

1.6 External change

Attention may need to be paid to how changes might occur in the use of the developed or acquired system, as this may require a complete re-evaluation of a system.

Some examples are:

- Deploying an existing system in a different environment, including different users, target markets, and data sources, can change risks and require a system to be re-evaluated.
- Over time, the relationship between the inputs and outputs of the system can change. For example, a system that uses an ML model to make decisions based on correlations established during initial system training might suffer if those correlations change over time.

Use cases for the system can be developed, either deliberately or organically, requiring a reassessment of the risks.

- Social norms change over time (e.g., attitudes toward gender behavioural norms, ideal body shape, or smoking). Bias in AI systems may need to be reassessed to consider the resulting changes (such as metrics, risks, stakeholders, or requirements) and addressed accordingly.

1.7 Acceptance criteria

The effective requirements are verifiable, since when they are evaluated, it is possible to determine whether a system meets them. Often, the performance of the AI system is compared to that of humans. Still, it's beneficial to be able to specify performance statistically. For example, stakeholders can specify a threshold for false positive or false negative decisions, in addition to an overall accuracy metric.

Establishing acceptance in terms of specific characteristics of the system and the degree of its compliance in advance allows for effective evaluation and decision-making.

Failure criteria can be the lower bound of acceptance, setting clear limits for the acceptable performance of a model. Without these criteria being established and monitored, an AI system could be diverted in such a way that unwanted biases arise without being noticed or remedied. The way a system fails might also need to be carefully considered as part of the design process to prevent extreme cases of bias from occurring.

2. Design and development

The models themselves can be biased if care is not taken to avoid this. Human biases can be encoded into ML systems through implicit assumptions that make their way into design. Therefore, it helps to identify and make implicit assumptions explicit.

2.1 Data Representation and Labelling

A key step in developing a machine learning system is deciding how to best represent the training data into features that the model can interpret. This is also called feature engineering, and Annex C.1 describes some types of data bias that could affect this process. There are several often implicit criteria that go into this process, including the criteria by which data is considered "good" or "bad" (e.g., whether an overexposed photo could be kept in a dataset). These criteria for what data is included in the training data and which characteristics are selected can be made explicit. It is important to consider how the data relates to the purpose of building the system, the process by which the characteristics are chosen, and the people who choose the characteristics and their justification.

It is important to evaluate the characteristics chosen for any data and human cognitive bias, such as missing entity values, unexpected feature values, or data bias. Any of these could indicate that certain groups or characteristics are not accurately represented in the data.

The lack of entity values could be the result of implicit biases in the data collection process, which might need to be modified.

In deep learning algorithms where features are created during training, the right tags are critical. Annotations made by humans to create labels for data can be prone to bias due to human cognitive biases or human error due to difficulties in the labelling task itself. It is important to ensure that labels are correct by evaluating both the labellers and the final labelled data.

Data annotators often annotate and label the data that will be used for supervised ML training. The consistent errors or human cognitive biases that manifest during that process are propagated in a trained model.

When using data annotators for data labelling, it might be helpful to understand the diversity and goals of the people annotating the data. For example, you could consider what success looks like for different workers, and the trade-offs between time spent on task and enjoyment of the task.

Deployers can develop tasks that take into account human differences (and cognitive biases) in annotation, for example, by using gold test questions with known answers or other forms of participant pre-screening.

Clarity of instructions, as well as getting feedback from crowd workers on potentially confusing tasks, could be important in reducing unwanted bias. Human variability, including accessibility, muscle memory, and human cognitive biases in annotation, could be explained by using a standard set of questions with known answers.

2.2 Transparency tools

To clarify the intended use cases for ML models and minimize their use in contexts for which they are not suitable, published models may be accompanied by documentation detailing their performance characteristics. The model's transparency tools could provide a framework for transparent reporting on ML model provenance, use, and fairness-based evaluation. Model documentation may include:

- Qualitative information, such as ethical considerations, target users, and use cases;
- Quantitative information, which consists of a disaggregated evaluation of the model (divided into the different target subgroups) and intersectional (including the evaluation of multiple subgroups in combination, e.g., race + gender).
- Data information, if possible, that could be formalized as a data transparency tool.

The usefulness and accuracy of a transparency tool depends on the integrity of the creator(s) of the tool itself and can be stored as documentation or metadata associated with each model.

2.3 Training

2.3.1 Training data

A seemingly straightforward approach to bias mitigation is to remove relevant characteristics that could be directly responsible for bias. To illustrate that, in a use case that automatically shortlists candidates based on resume information, examples of such characteristics are race, gender, and age. At the same time, the characteristics that are relevant to the use case include experience, skills, qualification, certification, and professional membership. While the elimination of race, gender, and age could address the problem, other characteristics and indirect variables could indirectly reflect biases. For example, characteristics such as greeting/prefix (Mr./Mrs.) or occupation can reveal gender. Therefore, removing only the characteristics associated with bias may not always work.

Data-driven methods can be used to mitigate bias in training data. Data reweighting, for example, could increase the weight of samples that align with a goal. Such techniques include:

- Sampling techniques to measure the representativeness of samples from different sources to assess selection bias.
- Stratification sampling to overcome a rarity phenomenon. Stratified sampling could be used by increasing the relative frequency of positive cases compared to negative cases.
- Careful selection of characteristics in cases where the characteristics of the sample have a strong correlation with the bias to be excluded (e.g., gender or colour).

Another approach is to find out the amount of bias present in the data and compensate for the bias of the result. Using a series of steps, it is possible to find out the contribution of the feature and the relative importance of each feature in the model's prediction. It is possible to offset all the influence by a characteristic that causes the bias. The process of determining the materiality of characteristics may include the following:

- Iterative orthogonal projection of features. Given the input and output of an ML model, the method seeks to produce an input classification that corresponds to the machine learning system's dependence on each input in its decision-making process, and thus could detect biases involving certain characteristics.
- Minimal redundancy, maximum relevance. Feature selection identifies subsets of data that are relevant to the parameters used. One scheme is to select the features that correlate most strongly with the classification variable and at the same time are mutually distant from each other. This scheme, called minimum redundancy and maximum relevance selection, has been found to be more powerful than other modes of feature selection.
- Ridge Regression/LASSO Regression. LASSO and Ridge are linear regression methods with regularization to avoid overfitting to the training data. These techniques are also used to help with feature selection.
- Random Forest. Random Forest is an approach that combines several random decision trees and aggregates their predictions averaging. This technique has shown excellent performance in environments where the number of variables is much greater than the number of observations.

2.3.2 Testing

Bias assessment is a difficult task. Methods of grouping and visualizing training data, derived features, or resulting predictions can help detect imbalances or potential biases in the data or training system. In many cases, preparing or curating a balanced dataset can take up most of the development time in deep learning-based systems. Having separate teams working on training and evaluation could also prevent the influence of individual cognitive biases.

Investigating apparent flaws in the model could reveal why you're not maximizing overall accuracy. Solving these defects could improve overall accuracy. Sub-representative datasets from certain groups might need additional training data to improve decision-making accuracy and reduce biased results.

2.3.3 Adjustment

Bias mitigation algorithms have been created to achieve several goals. Bias mitigation algorithms (also sometimes referred to as fair algorithms) can be classified as follows:

- Data-driven methods, such as bottom-up sampling of underrepresented populations or the use of synthetic data.
- Model-based methods, such as adding regularization terms or constraints that impose a goal during optimization, or learning representation to hide or reduce the effect of a specific variable.
- Post-hoc methods, such as identifying group-specific decision thresholds based on predicted outcomes to match false positive rates or other relevant metrics.

Examples of bias mitigation algorithms that are applied are:

- Disparate Impact Remover: A pre-processing technique that edits values to be used as characteristics in such a way as to reduce different treatments between groups;

- Individual Bias Detector and Remover: A technique that creates a new ML model for individuals in the disadvantaged group who receive a different decision compared to similar individuals in the favoured group. Sometimes you can apply different thresholds for positive classification between groups;
- Decoupled Classifiers: A technique for training a separate classifier in each group. The separate classifiers could be thought of equivalently as a single classifier branching into the group function.
- Joint Loss Function: A technique for capturing group parity by using a joint loss function that penalizes differences in classification statistics between groups.
- Transfer Learning: A technique for mitigating low-volume data issues for groups where there is a smaller data population.

2.4 Bias assessment

Bias in AI systems is measured comparably to how other properties, such as aggregate performance, are measured. However, aggregated performance metrics across the validation set may not indicate whether there is bias in the model. The overall metrics in the confusion matrix may appear to work well across the whole. However, calculating accuracy and recall in subsets of demographically important or certain categories can often reveal biases such as lower accuracy for women than for men, or lower accuracy for a specific demographic. These differences in performance likely indicate that undetected biases are present early in the development process. For example, a certain group might be underrepresented in training data.

2.5 Adversarial methods to mitigate bias

One method to mitigate bias is to incorporate an adversarial unit into the model architecture. In these methods, an "adversary" is predicting some property or characteristic that defines groups toward which equity is desired. The output of the model for which bias is being mitigated is the input to the adversarial model. The weight update for that model is modified so that, in addition to being optimized for the task it is performing, it also reduces the amount of information it makes available to the adversary useful for its prediction. The net effect of this system is that the system learns to perform its task in an orthogonal manner to the characteristics for which bias is unwanted.

2.6 Bias in rule-based systems

Diversity in designers' backgrounds and experiences, along with leveraging transdisciplinary experts, as explained above, could help reduce the chances of introducing bias into a system's design. For example, consider a system for the automatic identification of potential smugglers with coded rules based on the knowledge of a few very experienced subject matter experts. If subject matter experts have expertise primarily in a particular area of smuggling, using the system in a different area could result in an unintended profiling of specific classes of people. This, in turn, could lead to a systematic difference in how these classes will be treated in relation to other classes in the new context. A common consequence in such a scenario would be the unbalanced chances of pointing someone out wrongly among different cohorts.

3. Verification and validation

Verifying and validating a newly developed ML model could identify and mitigate potential biases prior to deployment. A retention dataset obtained from a data source separate from the training data is typically used in verification and validation. This safeguard for model generalization is also important to guard against any bias implicit in the training data set. In general, any steps taken during dataset processing and model training would be beneficial to apply to data and validation procedures where appropriate.

While verification of ML systems is carried out intensively using training data and test sets, it is limited to verification of results based on the selection and variation of available data. An AI system may need to be evaluated in a specific context.

Software testing traditionally relies on a "body of knowledge used as the basis for test and test case design." The success of any empirical testing activity is often limited by the degree to which surrounding requirements or risks management processes have explicitly identified potential biases or sources of bias.

The techniques described in this section are intended to be carried out on a statistically significant scale. The techniques generally measure the sensitivity of the result to a subgroup not explicitly included in the input data.

This section is intended to apply to the development of new systems, the deployment of existing systems, and the evaluation of whether systems maintain quality over time. A change in the relationship between expected and actual input data may be subject to evaluation. The assessment can also include the results of the deployment in system deployers and bystanders (such as people or objects that are incidentally present but are not the target or subject of an AI system deployment). For example, a system that is unfair to gender and race independently might be fair to a specific combination of the two.

3.1 Static analysis of training data and data preparation

Analysis of training and input data could provide useful information and detect the types of data bias described in Annex C.1.

Evaluators can identify the profile of training and input data and validate whether the propagation of a given variable represents the expected actual dataset. An example of this is identifying that records from a certain age group have been used for training, when a different distribution of ages is expected in actual datasets. This activity could aim to validate the potential for selection bias, sampling bias, and coverage bias, but it cannot do so exhaustively, as it is limited by the evaluator's knowledge.

Evaluators could identify stages in the data preparation process that could introduce bias through "missing data." For example, if specific data is not consistently available in an input data set, engineers can impute that information for the remaining records or they can delete it. If the absence of that data element correlates with specific groups of records, this could result in bias that would not normally be detected in model testing.

3.2 Label Controls Samples

The risk of mislabelling described in Annex C.1, i.e. human labellers incorrectly specifying labels for an input data set that are then used to train the model, could be assessed by sample controls of the submitted labels.

Labelling based on expert opinion could be more complicated to evaluate. Double-blind reviews, or evaluation by multiple experts, may be necessary to assess the quality of the initial label.

3.3 Internal Validity Tests

Internal validity tests evaluate the correlation between individual input data and system outputs. Internal validity tests then review whether these correlations are adverse in the context of specific requirements or acceptance criteria.

This process relies on data that causes a bias to be included in the input data domain. It could detect biases in the models and their interaction.

This could include evaluation in a fully integrated environment, in order to detect any bias in the data collection or preparation activities used during the development of the AI system. The integration can also detect non-representative sampling. For example, data collected in an integrated environment can have variable characteristics, such as lighting levels or sensor refresh rate. These variations can influence the input data to the AI system.

3.4 External validity tests

External validity tests may involve re-evaluating previous observations using external data sources. This is a useful technique because it can detect many types of bias that have been described in this document, including indirect bias. The aspect of the input data to which indirect bias refers is not explicitly contained within the input data, but is a second-order derivation.

For example, some media reports on AI bias have focused on research that correlates model results with census data or zip code with results to illustrate the disparity of results.

External validity tests can also include integrating new input data and validating that the results are consistent with internal validity tests.

External validity tests are particularly important for indirect biases introduced by proxy variables. If a model designer attempts to mitigate bias by simply removing demographic information from input, bias is likely to continue to exist through proxy variables. For example, the model could perpetuate "colourblind" racism, a sociological concept that describes how claims of not "seeing" a person's skin colour prevent understanding and addressing the persistence of racial inequality in society. To avoid this outcome, external validity tests might need to include originally excluded demographic data, or an exploration of proxy variable effects. Further investigation might be necessary to understand why such proxies exist and whether the purpose can be fulfilled without them. External validity tests might also need to include qualitative data demonstrating disparate impacts of the same classification. For example, if a certain model is used to identify people

before boarding an airplane, the emotional damage of a false negative could be greater for those groups stereotyped as likely "terrorists" than for other groups.

In this context, it might be necessary to integrate input data sets with additional data points to properly assess system bias.

3.5 User Testing

Testing with different types of end users can be useful when a user's interaction with the system influences results and predictions in a way that correlates with the user's membership in a group.

Assessing user experience in real-world scenarios across a broad spectrum of users, use cases, and usage contexts is a useful technique for detecting bias in model interactions, data processing issues, and issues with data labels.

3.6 Exploratory tests

System developers could organize a group of reliable and diverse testers who could adversely test the system and incorporate a variety of potentially harmful inputs into unit or functional tests. This could help uncover unforeseen ways in which a system might be biased.

4. Deployment

Once implemented, proper training and support for the AI system is important for the deployers to enable effective use of the product. This includes guidance for system developers on what constitutes an appropriate and inappropriate implementation of a model within a software system. For example, an attention tracking system might be perceived as unfair if it is used in an educational system to monitor student behaviour, but that might not be the case if the same system is used as a research tool in a psychology experiment.

Deployed systems may also include instructions for system end users. For example, it is desirable that recruiters who use a hiring recommendation system understand the capabilities and limitations of the system. Both system developers and software end users may need to be aware of known areas of bias. This can be achieved through a transparency tool, which contains information about the data on which the model was trained, distributions for populations of its false positive and false negative errors, and other associated information.

The data subjects, the people to whom the training data refer, are not necessarily responsible for the deployment of the system. They do not need to be trained, but they may need to be informed of any bias within the system that may affect them, in language appropriate to the context. Failures in training or support can lead to additional biases that can be difficult to detect sooner.

4.1 Continuous validation

Models can degrade and lose performance over time. Yield degradation can be attributed to changes in the environment, new emerging behaviours, changes in the composition of

the input population, and changes in requirements. In addition, a system could be biased towards a historical position.

Continuous performance monitoring may be required when the system is deployed. This includes checking system performance, especially outliers, both manually and using different metrics and techniques to assess bias and fairness. If there are indications of bias, the system may need to be retrained or redesigned.

Monitoring is a well-known process in many industries that leverage automated decision-making in their processes. For example, in banking, dashboard models are being developed and introduced along with their approved monitoring processes. Monitoring processes could not only be applied to the accuracy and performance of models/systems, but could also be used for the identification and monitoring of biases in systems or models.

9. References, Standards and Norms

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

[1] ISO-IEC 38505-1 - Information technology - Governance of IT - Governance of data. Part 1: Application of ISO/IEC 38500 to the governance of data

[2] ISO-IEC 38507 - Information technology - Governance of IT - Governance implications of the use of artificial intelligence by organizations

[3] ISO-IEC 5338 - Information technology - Artificial intelligence - AI system life cycle processes

[4] ISO-IEC 24027 - AI Algorithmic bias - Information technology - Artificial Intelligence (AI) - Bias in AI systems and AI-aided decision making

[5] ISO-IEC 5259-1 - Artificial intelligence - Data quality for analytics and machine learning (ML) - Part 1: Overview, terminology, and examples

[6] ISO-IEC 5259-2 - Artificial intelligence - Data quality for analytics and machine learning (ML) - Part 2: Data quality measures

[7] ISO-IEC 5259-3 - Artificial intelligence - Data quality for analytics and machine learning (ML) - Part 3: Data quality management requirements and guidelines

[8] ISO-IEC 5259-4 - Artificial intelligence - Data quality for analytics and machine learning (ML) - Part 4: Data quality process framework

[9] ISO-IEC 5259-5 - Artificial intelligence - Data quality for analytics and machine learning (ML) - Part 5: Data quality governance

[10] ISO-IEC 8183 - Information technology — Artificial intelligence — Data life cycle framework

[11] ISO-IEC 42001 - Information technology - Artificial intelligence - Management system

[12] ISO-IEC 22989 - Information technology - Artificial intelligence - Artificial intelligence concepts and terminology

[13] prEN 18229 AI Trustworthiness Framework

[14] prEN XXX Quality and governance of datasets in AI

[15] prEN XXX Concepts, measures and requirements for managing bias in AI Systems

[16] Especificación UNE 0085:2024

[17] VERMA, Sahil and RUBIN, Julia, 2018. Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness - FairWare '18

[online]. Gothenburg, Sweden: ACM Press. 2018. p. 1–7. [Accessed 6 May 2019]. Fairness definitions explained | Proceedings of the International Workshop on Software Fairness (acm.org)

[18] MITCHELL, Shira, POTASH, Eric, BAROCAS, Solon, D'AMOUR, Alexander and LUM, Kristian, 2020. Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. arXiv:1811.07867 [stat] [online]. 24 April 2020. [1811.07867] Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions (arxiv.org)

[19] GAJANE, Pratik and PECHENIZKIY, Mykola, 2018. On Formalizing Fairness in Prediction with Machine Learning. arXiv:1710.03184 [cs, stat] [online]. 28 May 2018. [Accessed 30 October 2020]. [1710.03184] On Formalizing Fairness in Prediction with Machine Learning (arxiv.org)

[20] AGARWAL, Alekh, DUDÍK, Miroslav and WU, Zhiwei Steven, 2019. Fair Regression: Quantitative Definitions and Reduction-based Algorithms. arXiv:1905.12843 [cs, stat] [online]. 29 May 2019. [Accessed 30 October 2020]. [1905.12843] Fair Regression: Quantitative Definitions and Reduction-based Algorithms (arxiv.org)

[21] Confusion matrix, 2020. Wikipedia [online]. [Accessed 30 October 2020]. Confusion matrix - Wikipedia

[22] Spanish Data Protection Agency. "Adaptation to the GDPR of processing that incorporates Artificial Intelligence. An Introduction". 2020.

[23] Spanish Data Protection Agency. "Requirements for Audits of Treatments that include AI". 2021.

[24] ENISA. "The Adoption of Pseudonymization Techniques: The Case of the Health Sector. 2022.

[25] Spanish Association for Standardization. Guide to the evaluation of Data Governance, Management and Quality Management (UNE 0080).

[26] Spanish Association for Standardization. Data Quality Management (UNE 0079).

[27] Spanish Association for Standardization. Data Management (UNE 0078).

[28] Spanish Association for Standardization. Data Government (UNE 0077).

The version of the European Regulation on Artificial Intelligence that has been taken as a reference has been the one published by the Council of the European Commission on June 13, 2024.

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