13

Companies developing compliance with requirements

Guide 13. Post-market monitoring

European
Artificial Intelligence Act

Financiado por
la Unión Europea
NextGenerationEU

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

Plan de
Recuperación,
Transformación
y Resiliencia

España | digital 2026

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

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

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

Among the currently applicable relevant technical references, the following standards stand out. On the one hand, **ISO/IEC WD 25059:2021** ***"Software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality model for AI systems"*** for defining the indicators to be monitored in our intelligent systems. On the other hand, **ISO/IEC 25000:2021** ***"Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE)"*** for designing the monitoring system and defining how such activity is to be carried out.

Revision date: 10 December 2025

General content

Detailed Index

1. Preamble

1.1 Purpose of this document

The European Regulation on Artificial Intelligence (*AI Act*) indicates the need to carry out a **Post-market monitoring plan** for high-risk artificial intelligence systems. The objective of this report is to document the processes that must be carried out in this plan and establish recommendations to achieve it.

A post-market monitoring plan is a **set of activities** conducted by **providers/users**, to collect and evaluate experience obtained from artificial intelligence systems, considered high risk, which have been **put on the market**, and thus identify the need to take any action. This is an important tool to ensure that AI systems remain secure and function properly. In addition, in this way, the **development of actions** is contemplated in the event that the continued risk of the high-risk system begins to outweigh the benefit. The evaluation carried out with this post-market monitoring can also contribute to a continuous improvement of the system in question.

The objective of this Guide is to present the processes that must be carried out in this plan and establish recommendations to achieve it.

1.2 How to read this guide?

As mentioned above, **this document provides** implementation measures for providers and users of AI systems **to facilitate compliance with** the obligations expressed in Article 72 of the AI Act, dedicated to post-market Monitoring.

To this end, the document **goes through** all the sections of said article in order, answering the fundamental questions necessary to **facilitate the fulfilment** of the obligations expressed in these sections.

In addition, we must take the following issues into account for an efficient reading of this guide:

1. **Connection with other guides**: in the event that you do not have knowledge or context about the rest of the guides, it is recommended to start by reading the introduction of this guide to understand the relationship with the rest of the guides and understand the previous steps to be taken.
2. **Development of the post-market monitoring system**: section 4 of the Guide explaining how to develop and implement the monitoring system as well as the measures that should be contemplated in the monitoring plan. Before reading it, it is recommended to review Annex A to deepen the concept of indicator and the lists of minimum indicators.
3. **Technical documentation**: finally, a reading is recommended to obtain an intuition of what the objectives to be covered are.

1.3 Who is it for?

The requirements described in Article 72 "*Post-market monitoring by providers and post-market monitoring plan for high-risk AI systems*" are focused on the measures that **the service provider** must take once the intelligent system is in production. Therefore, it is the responsibility of the provider to assess that the requirements set out in this article are met throughout the life cycle.

The deployer's duties **focus on reporting incidents and anomalous behaviour to the provider**. In particular, when an abnormal change in the system's behaviour with respect to its instructions for use is detected, the provider must be notified. In addition, in the event that the user is unable to contact the provider, it will be responsible for applying the necessary changes and suspending the use of the system as specified in Article 26-Section 5 of the European Regulation on Artificial Intelligence.

1.4 Use cases and examples throughout the guide

To contextualize, where applicable, the measures exposed that allow the requirements of the AI Act to be met, examples will be used on two use cases:

- Employee promotion
- Chronic Disease Management - Smart Insulin Pump

These use cases are developed in detail in the Concepts and Cross-Cutting Information Guide.

The examples of these use cases are presented at a high level, without going into detail or being exhaustive, in order to try to cover as many cases as possible. In addition, they do not respond to real experiences (but with the intention of being realistic from a didactic point of view), with the aim only of clarifying the measures a little more, therefore they cannot be taken as specifications in a real implementation.

2. Introduction

2.1 What do we mean by 'post-market monitoring system' and why is it necessary?

A post-market monitoring system for high-risk AI systems is conceived as a **set of processes** and **tools** aimed at collecting data from a system to transform it into a series of indicators about its activity with the aim of monitoring artificial intelligence (AI) systems **after its market launch**. The objective is for the provider to be able to assess whether the AI systems meet the requirements set out in Chapter III, Section 2 (Section referring to the 'Requirements of high-risk systems'), throughout the entire life cycle of the intelligent system.

The post-market monitoring system operates through the following subsystems:

- **Indicators capture system.** Different processes that collect data on the performance of the intelligent system, its infrastructure, user interactions and different data on security (See Annex A of this guide: "Monitoring indicators" for the minimum list of indicators).
- **Systems for recording these indicators**. Storage services for such records in accordance with the measures described the Record-keeping guide: "What elements should I implement and how should I do so in order to develop an adequate records management system?".
- **Automated alert system**. Monitoring processes for changes in indicators based on their pre-established scales to alert on possible risk scenarios (See Annex A of this guide: "Monitoring indicators" for a list of minimum and maximum thresholds).
- **Different analysis interfaces for those in charge of Monitoring**. Access point and analysis by the system's guards to be able to analyse the extracted indicators. It can be in the form of a web application, a list of raw data or through any other tool that allows searches and operations on groups of records.

Example – Evaluation system for staff promotion

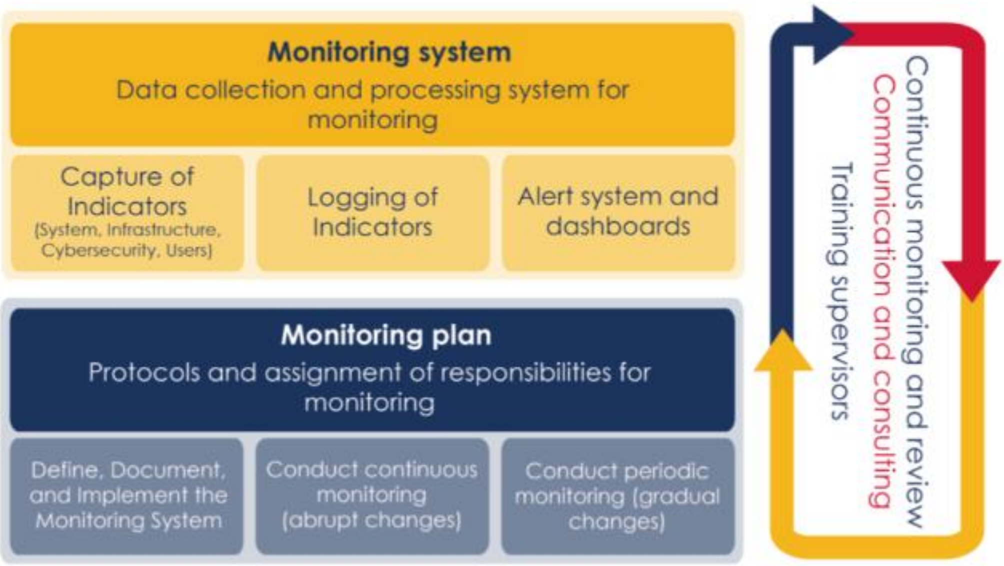
Let's take the case of a company that has implemented an AI system to evaluate staff promotion, which is based on a variety of parameters and data collected from employees. The system has been validated and approved by the relevant regulators and has been released to the market.

Considering that the established AI system is considered high risk, the company has implemented a post-market monitoring system to continuously collect and analyse data on the system's real-world effectiveness. Such monitoring system collects employee promotion data, as well as other relevant data, such as turnover rate and employee satisfaction, and analyses it to detect any unexpected patterns or trends.

Let's imagine that, after a few months of post-marketing monitoring, the indicator of one of the inputs (ethnic group of employees) undergoes a significant variation with respect to the average of said input in the historical record. In this case, the post-market monitoring system sends an alert via emails from the monitoring system to the team responsible for the AI system, which investigates the problem in the analysis interface.

The team finds that the system is using a historical data set that results in employees of a certain ethnicity or gender being less likely to be promoted compared to others, despite having similar qualifications. The incident is reported through the incident report to the designated managers in the monitoring plan. Subsequently, the company's team works quickly to develop a solution and releases an update to the AI system that eliminates bias. Once this is done, the post-market monitoring system continues to collect and analyse data to make sure that the problem has been resolved and that the AI system is working properly. Finally, those responsible for the system will document what happened in the means established in the monitoring plan, attaching the incident, the solution applied, and the results obtained.

As a visual introductory summary, an infographic is provided, that tries to give an overview of the post-market design of high-risk AI systems:


Monitoring system
Data collection and processing system for monitoring
Capture of Indicators
(System, Infrastructure, Cybersecurity, Users)
Logging of Indicators
Alert system and dashboards
Monitoring plan
Protocols and assignment of responsibilities for monitoring
Define, Document, and Implement the Monitoring System
Conduct continuous monitoring (abrupt changes)
Conduct periodic monitoring (gradual changes)
Continuous monitoring and review
Communication and consulting
Training supervisors

2.2 Why is a post-market monitoring system needed?

Post-market monitoring of high-risk AI systems includes a variety of tasks, considering the intended purpose of the system. It also considers the collection of data on the performance and security of the system, the evaluation of the possible causes of detected problems, the implementation of solutions to correct problems, and the communication of the results and recommendations to stakeholders.

It may also include implementing additional security controls to protect the system from malicious attacks, and conducting periodic assessments to ensure that the system continues to comply with the various requirements set out in the AI Act for these systems in terms of 'Data and Data Governance', 'Technical Documentation', 'Records', 'Human oversight, 'Transparency and communication of information' and 'Accuracy, robustness and cybersecurity'. These processes are important to ensure the safety and reliability of high-risk AI systems once deployed.

2.3 Entities subject to sectoral legislation

Those providers subjected by the legislative acts specified in Section A of Annex I, who have implemented a post-market monitoring system and plan in accordance with these provisions, have the power to use such mechanisms in accordance with the established regulations. This prerogative also extends to those providers that develop artificial intelligence (AI) systems considered to be high risk. That is, systems placed on the market or put into service by financial institutions subject to internal governance requirements, mechanisms or processes established in accordance with Union financial services law.

3. European Regulation on Artificial Intelligence

3.1 Preliminary analysis and relationship of the articles

In this section, the content of the article is simplified and structured in order to facilitate its understanding. Likewise, the structure followed in this Guide to cover the content of this article is presented.

What we understand from the article

The article discusses the factors required for the **implementation and documentation** of post-market monitoring systems for high-risk AI systems. In doing so, it aims to ensure that such systems continue to **meet the necessary requirements** set out in the AI Act once placed on the market, and throughout **the entire life cycle of the** system.

What key measures we understand the article establishes

- Establishment of monitoring systems proportionate to the risks.
- Collection, documentation, and analysis of relevant data on the performance of AI systems.
- Creation of a post-market monitoring plan as part of the technical documentation.
- Acceptance of existing monitoring documentation if it complies with certain legislative acts and financial regulations.

3.2 Content of the articles in the AI Act

AI Act

Art.72 Post-market monitoring by providers and post-market monitoring plan for high-risk AI systems

1. Providers shall establish and document a post-market monitoring system in a manner that is proportionate to the nature of the AI technologies and the risks of the high-risk AI system.

2. The post-market monitoring system shall actively and systematically collect, document and analyse relevant data which may be provided by deployers or which may be collected through other sources on the performance of high-risk AI systems throughout their lifetime, and which allow the provider to evaluate the continuous compliance of AI systems with the requirements set out in Chapter III, Section 2. Where relevant, post-market monitoring shall include an analysis of the interaction with other AI systems. This obligation shall not cover sensitive operational data of deployers which are law-enforcement authorities.

3. The post-market monitoring system shall be based on a post-market monitoring plan. The post-market monitoring plan shall be part of the technical documentation referred to in Annex IV. The Commission shall adopt an implementing act laying down detailed provisions establishing a template for the post-market monitoring plan and the list of elements to be included in the plan by 2 February 2026. That implementing act shall be adopted in accordance with the examination procedure referred to in Article 98(2).

4. For high-risk AI systems covered by the Union harmonisation legislation listed in Section A of Annex I, where a post-market monitoring system and plan are already established under that legislation, in order to ensure consistency, avoid duplications and minimise additional burdens, providers shall have a choice of integrating, as appropriate, the necessary elements described in paragraphs 1, 2 and 3 using the template referred in paragraph 3 into systems and plans already existing under that legislation, provided that it achieves an equivalent level of protection.

The first subparagraph of this paragraph shall also apply to high-risk AI systems referred to in point 5 of Annex III placed on the market or put into service by financial institutions that are subject to requirements under Union financial services law regarding their internal governance, arrangements or processes.

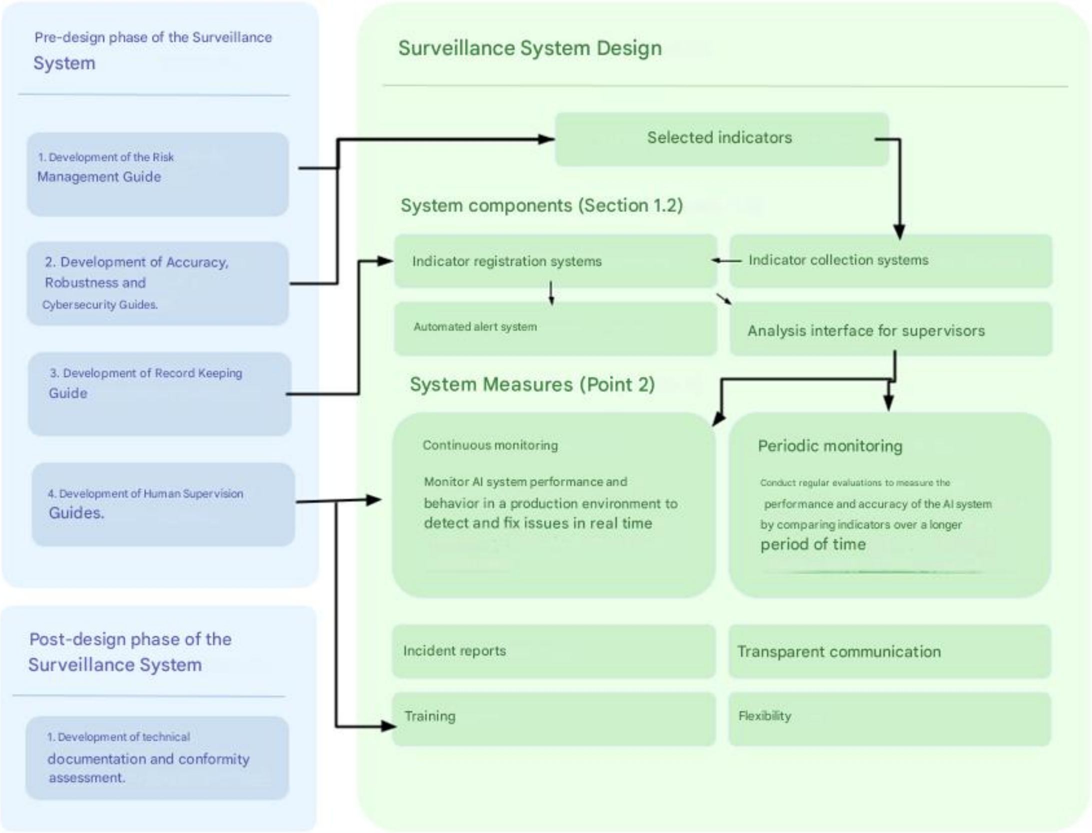
3.3 Correspondence of the articles with the sections of the guide

This table details the correspondence of the sections of this guide that address the elements of that article:

Article	AI Act requirement	Section
72.1	Proportionality of the post-market monitoring system based on the nature of the AI technologies applied.	Section 4
72.2	Deployers or that can be collected through other sources on the operation of high-risk AI systems throughout their lifetime.	Section 4.1 and Section 4.2
72.3	Post-market Mmnitoring plan shall form part of the technical documentation referred to in Annex IV.	Section 6
72.4	Specific documentation of high-risk AI systems regulated by the Union harmonisation legislation listed in Section A of Annex I.	Section 6

4. What elements should be implemented and how should I do it to develop an adequate post-market monitoring system?

Based on what was discussed in point 1 about what a post-market monitoring system is and what components it should have, we can follow the following flow chart in its design:


Pre-design phase of the Surveillance System
1. Development of the Risk Management Guide
2. Development of Accuracy, Robustness and Cybersecurity Guides.
3. Development of Record Keeping Guide
4. Development of Human Supervision Guides.
Post-design phase of the Surveillance System
1. Development of technical documentation and conformity assessment.
Surveillance System Design
Selected indicators
System components (Section 1.2)
Indicator registration systems
Indicator collection systems
Automated alert system
Analysis interface for supervisors
System Measures (Point 2)
Continuous monitoring
Monitor AI system performance and behavior in a production environment to detect and fix issues in real time
Periodic monitoring
Conduct regular evaluations to measure the performance and accuracy of the AI system by comparing indicators over a longer period of time
Incident reports
Transparent communication
Training
Flexibility

How to implement the post-market monitoring system?

In order to be able to carry out the main Monitoring activities, the provider must design, develop, deploy and validate the components that are part of the system. Specifically:

1. **Selection of indicators**: the most important indicators for the correct monitoring of the system must be established. (See Annex A to this guide: "Monitoring indicators.") Such selection must be based on the previously developed risks management system. Once selected, the intelligent system must generate records with the information of these indicators based on the technical recommendations of

the Records Guide: "What elements should I implement and how should I do it to develop an adequate records management system?". After specifying the design of the records, the intelligent system must deposit them in a temporary log file or similar so that the capture system can begin its processing. It may be the case that the intelligent system itself sends records to the indicator recording system without the need to deposit the records in a temporary file. Both options are valid and will depend on the architecture of the intelligent system itself.

2. **Development of indicator capture and sending systems:** The collection systems will collect data from the log files where they are deposited by the intelligent system and will be sent to the indicator recording systems. These systems should pay particular attention to the security of sending records.
3. **Indicator recording system**: A records management system must be implemented where the indicator capture systems will send the data generated by the system. This system may be any type of structured or unstructured database for storing records. In addition, the policies for controlling access, retention and deletion of these records will have to be established as indicated in the Guide for Records.
4. **Joint development of the alert system and the analysis interface for supervisors**: From the records, there will be a system for monitoring indicators for early warning of anomalous values. You can develop your own solution or use one of the solutions for viewing and monitoring records. You must have both:
   a. An alert system assigned to system supervisors. It is recommended that the system sends alerts through different means of communication and also periodically notifies about the appropriate behaviour of the system.
   b. An interface for real-time monitoring of system indicators. Such an interface shall clearly show when one of the system indicators is outside the expected scale. It should also allow manual scanning of the records for deeper inspection of the collected data.

All components of the system must have functionality tests at the unit level and integration with the rest to ensure the correct operation of the Monitoring system.

Before carrying out the design and implementation of such a Monitoring system, a post-market monitoring plan must be established that includes the following tasks and characteristics: continuous Monitoring, periodic Monitoring, incident reporting, transparent communication, training, flexibility and independent evaluation (if possible). The items listed are described below:

- **Continuous monitoring:** Monitor the performance and behaviour of the AI system in a production environment to detect and correct problems based on the indicators selected in the risks assessment.
- **Periodic monitoring:** Conduct regular manual assessments to measure the performance and accuracy of the AI system and detect any issues.
- **Incident Reporting:** Establishing a system to collect and analyse incident reports related to the AI system, including bugs, privacy and security issues.
- **Transparent communication:** Provide clear and transparent information about the performance and security of the AI system to users, regulators, and other stakeholders through reporting.

- **Training:** Train users and operators of the AI system to detect and handle issues associated with anomalous system operation.
- **Flexibility:** Having a flexible and scalable plan to adapt to changes in AI system performance and security and to comply with current and future regulations.

If possible, an **independent assessment** would be valued, which could provide an objective view of its performance and safety. For example, the achievement of a seal or certification that guarantees this by an accredited entity.

The following sections of point two present each of these requirements in greater depth. Subsequently, section 4 of "Technical documentation" will indicate how to document both the monitoring system and the proposed monitoring plan.

4.1 Continuous monitoring

Continuous monitoring of high-risk AI systems is necessary to ensure that the system continues to operate safely and effectively and to avoid performance, security, and legal liability issues once in the market, in the event of **abrupt changes** in intelligent system performance.

Keep in mind that AI systems operate in changing environments, and can be affected by changes in input data, environmental conditions, and regulations. Continuous monitoring of the indicators selected after the risks assessment allows problems related to these changes to be detected and corrected.

Additionally, AI systems can experience performance issues due to a variety of factors, such as aging training data, using insufficient or inaccurate training data, or lack of proper training. Continuous monitoring allows you to detect and correct performance issues before they significantly impact the system.

It should not be forgotten that these systems can be subject to malicious attacks, such as adversarial learning, phishing, and data theft. Continuous **monitoring allows you to detect and correct security issues before they cause significant damage**.

Measures to carry it out

Existing techniques for monitoring the performance and behaviour of an AI system in a production environment include:

- **Monitor intelligent system indicators:** Collecting and analysing data on the performance and behaviour of the AI system to detect patterns and trends. Data analysis tools can be used to detect problems, such as errors and deviations.
- **Monitor indicators about user actions:** Collect information about how users interact with the AI system, including commands, queries, and responses. This can help detect usability issues and provide feedback to improve the system.
- **Monitor infrastructure indicators:** Monitor the health of AI system components, such as CPU, memory, and storage, to detect performance and capacity issues.
- **Monitor security indicators:** The goal is to detect and prevent security breaches, including phishing, brute force, and malware attacks.

- **Monitor indicator changes using alerts:** Configure alerts to detect anomalous events, such as bugs, security breaches, and privacy issues. This can include configuring rules to detect specific patterns or configuring thresholds to detect deviations.

Example – Smart Insulin Pump

Using the example of the smart insulin pump, continuous monitoring would focus on monitoring selected indicators to **avoid abrupt performance changes.**

Let's assume that the indicator of the number of predictions, which continuously had values close to 45 predictions per minute, has become 3640 predictions per minute.

First, the records with the indicators will be transferred by the indicator capture and sending system to the records management system through a secure communication protocol. Subsequently, the alert system will analyse the data obtained and detect the anomaly with respect to the indicator's normality scale based on the rules pre-established in its design. Finally, the alert system will send notifications to supervisors through the means defined in its design: emails, SMS or any other communication system of the organization.

After notification, the supervisory team will evaluate the indicators captured through the monitoring interface and review the system.

4.2 Periodic monitoring

Regular evaluations to measure the performance and accuracy of the AI system is essential to ensure that it remains accurate, scalable and useful in a real-world environment, a **longer review over time,** depending on the characteristics of the system itself, allowing problems to be detected early and measures to be taken to correct them.

On the one hand, it is important to ensure the accuracy of the system: through accuracy tests, it is possible to evaluate how the system is performing specific tasks, such as recognizing objects in images or translating languages, and detect if there are any problems that are affecting its accuracy.

In addition, it is also necessary to identify problems in advance: this can be achieved by performing follow-up tests, the performance of the system can be monitored over time, detecting any decrease in accuracy or performance, which allows possible problems to be detected in advance and measures to be taken to correct them before they become serious.

It also helps ensure the scalability of the system: through performance testing, you can assess how the system is handling large amounts of data and detect any issues that may affect its ability to scale.

Example – Evaluation system for staff promotion

Using the example of the personnel promotion system, the intelligent system could lose performance by not retraining the system for years and making promotion decisions that end in layoffs or loose teams. Regular monitoring of performance metrics will make it possible to assess their usefulness and prevent such scenarios.

Measures to carry it out

There are several techniques for evaluating the performance and accuracy of an AI system, some of which include:

- **Performance testing:** These are used to measure the response time of the system and its ability to handle large amounts of data. These are the well-known stress tests that allow the intelligent system to be pushed to the limit in a test environment and assess the response offered in such a scenario.
- **Accuracy tests:** These are used to measure the accuracy of the system when performing specific tasks, such as recognizing objects in images or translating languages. Specifically, the accuracy metrics associated with the AI model must be previously selected (section of the Accuracy Guide: "Accuracy metrics associated with the AI model") and the relevant accuracy levels for the system (section of the Accuracy Guide: "Accuracy Levels, Documentation and Monitoring of AI models.").

However, in some cases, measuring the accuracy of the algorithm can be very complicated due to the use case itself, as can be seen in the example below.

Example – Evaluation system for staff promotion

Continuing with the proposed example of the intelligent promotion system, how could we evaluate whether or not the intelligent system has been correct in its prediction? How can we assess that he has been right not to promote a particular employee? In these cases, it is recommended to maintain a control sample that follows the procedure prior to the implementation of the system in order to compare the results obtained with those offered by the system. To do this, a sample of employees would be left to be evaluated and promoted manually and after this process, the decisions made by the model would be evaluated based on the previous metrics against the manual decisions obtained by the HR team.

In addition to the above tests, the following recommendations should be taken into account:

- **Results history:** As indicated in the Accuracy Guide in its "General Technical Measures" section, a history of the results obtained in previous tests must be maintained to discover performance reduction trends early. As an example, if those in charge of the promotion system will carry out different tests in the first month of deployment in production of an intelligent system, obtaining an accuracy of 99.8%, 99.6% and 99.2% successively, if the minimum accuracy criterion were 98%, this trend of performance loss could go unnoticed. However, by having a history of tests, the trend in the results can be observed.

- **Special care with the improvement in the metrics without variation of the system:** in some cases, the decisions made by the results of the intelligent system may produce a tendency to improve the system metrics by indirectly looking for the inputs that adapt to the improvement of the performance of the algorithm.

Example – Evaluation system for staff promotion

Continuing with our example, if the system of selection of promotion candidates were retrained on data on which it has predicted, it would be biasing itself. In other words, if the algorithm gives a higher probability to employees with certain characteristics, they end up being promoted and the system is retrained on the results obtained, it will progressively bias towards the decisions it made initially, increasing its security. This scenario can be corrected through training on a control sample that has been promoted through HR experts.

4.3 Incident Reports

When establishing a system for collecting and analysing incident reports related to the AI system, we offer the following recommendations:

Define the objectives of the system: Establish the objectives of the incident collection and analysis system. That is, to answer the following question: What information is expected to be collected and how will this information be used?

System design: consisting of the collection and analysis of incident reports. This includes defining the processes for reporting incidents, the structure of the reports, and the mechanisms for data analysis.

System implementation: Includes developing a platform to collect and store reports, as well as creating data analysis tools.

Training: Work should be done on training the people involved in the system, including end users, system administrators, and the data analysis team.

Continuously monitor: This Monitoring includes collecting feedback from users, reviewing incident reports, and analysing trends to identify potential problems and opportunities for improvement.

Example – Smart Insulin Pump

After the detection of abnormal behaviour in the smart insulin pump, a report has been generated with the following information:

- **Characteristics and objectives of the system**: a description of the system in question to facilitate impact assessment by people without knowledge related to the system. This information must be pre-filled in the report.
- **Anomalous behaviour detected:** in this case the anomalous indicator has been the number of predictions per minute of the pump. Contextual information is also described, such as the number of systems affected, the beginning of the anomaly, the impact produced and other related factors that may help to understand the situation.
- **Actions carried out so far**: the monitoring interface and indicator records have been reviewed. The issue occurred on a single device and required an on-site review.
- Other sections that can provide more context of the anomaly detected in the intelligent system:
    - System Settings and Settings at the time the anomalous behaviour occurred.
    - Extract from the history of records and indicators of the intelligent system.
    - Notes from supervisors that can add context to the situation.

Once the report is completed, it must be transferred through the predefined sending system.

4.4 Transparent communication

The communication of the characteristics of the system, its performance and the consequences of its use in production must be adapted to the recipient of this information to facilitate a correct understanding of all the implications of its use.

Measures to carry it out

It is important to provide clear and transparent information about the performance and security of the AI system to deployers, regulators, and other stakeholders. To this end, it is recommended to establish clear and quantifiable performance and safety indicators to measure the performance and security of the AI system. These indicators should be relevant to deployers, regulators and other stakeholders.

On the other hand, it is relevant to collect data on the performance and security of the AI system and analyse it to obtain information about its performance.

As a measure to highlight, within this section, there is transparency when it comes to offering clear information on the performance and security of the AI system deployers, regulators and other stakeholders. This can include regular reports, tracking dashboards, and a dedicated website.

In this regard, it is advisable to provide details about how the system works, including its algorithm, the data it uses, and the decisions it makes. This will help deployers, regulators, and other stakeholders better understand the system and assess its performance and security.

Example – Evaluation system for staff promotion

An example of these measures, using the example of the employee promotion system, would be the adaptation in the way accuracy metrics are communicated by the system's supervisors when they are reported to the system's decision-makers. In this case, instead of mentioning that:

"The intelligent system has 96% accuracy"

The message could be digested to indicate that:

"The intelligent system has 96% accuracy, that is, out of every 100 employees that the system indicated should be promoted, the system got 96 of them right and made a mistake in indicating a promotion for 4 of them. However, this metric does not consider errors that occurred among employees who were not promoted, and other metrics should be considered along with the current one."

4.5 Training

When training supervisors, basic training should be provided on how the AI system works and how it is used. This includes information about the algorithms used, the data being used, and how decisions are made. It is also important to provide examples of anomalous operation, including examples of errors, failures, and unexpected behaviour.

Example – Evaluation system for staff promotion

In the case of the example, the provider of the intelligent personnel selection system for internal promotions must have supervisors with the following requirements:

- Training about the context in which the system is being used and the consequences of the predictions generated for employees.
- Knowing the data that is being taken into consideration to carry out the predictions (employee experience, projects completed, availability of transfer, among others) and how they can be affected by external factors (latency in the introduction of new updates in employees' CVs by the human resources team).

  Understanding of the algorithm or algorithms used in the system and what system is used to generalize knowledge about employees. As an example, knowing that it is a rule-based algorithm will help.
- Experience through examples of audiovisual or practical scenarios where anomalous behaviour of the intelligent system in which it does not generate an adequate prediction for employees or does not allow predictions to be made due to an overload of the infrastructure and how they should act in such a case.

On the part of the users of the system themselves, human resources specialists, should also be trained in the detection of this type of scenario and the appropriate measures to report any anomalous situation to the service provider.

4.6 Flexibility

Maintaining a flexible and scalable plan is critical to improving system Monitoring. On the one hand, there is the fact that AI systems are constantly evolving and improving, so this feature can allow you to adopt new technologies and techniques to improve performance. Likewise, in terms of security, flexibility allows us to adapt to new threats and vulnerabilities, as well as guarantee the protection of data and user privacy, and adapt to new changes in regulation.

In short, flexibility is the adaptation of the intelligent system to internal and external changes that may affect its operation.

Measures to carry it out

a) **Identify applicable regulations:** Includes sectoral regulations and data protection regulations.
b) **Assess system performance and security:** The use of clear and quantifiable performance and safety indicators is recommended.
c) **Identify risks:** Includes performance risks and security risks.
d) **Establish procedures to prevent risks:** Includes procedures for reporting incidents, investigating incidents, and implementing solutions.
e) **Monitor compliance with existing regulation:** Includes compliance with industry regulations and data protection regulations.

f) **Establish a contingency plan:** This issue is important in order to prevent critical incidents or failures in the AI system.

g) **Implement a continuous review system:** Its objective is to evaluate the performance and security of the AI system, identify problems and take measures to improve it.

h) **Specific training on existing** regulations on AI, and data protection, as well as '**best practices**' (Ethics in AI).

Example – Evaluation system for staff promotion

The company providing the intelligent employee promotion system has a change in its management and in its employee incorporation and promotion policy, after a merger with another company whose data has not been incorporated into the system. The provider shall adapt the Monitoring plan and Monitoring system to the new changes:

- Evaluate whether there are new applicable regulations after incorporation.
- Review that the indicators are still adequate for adequate Monitoring based on the new risks assessment.
- Carry out the necessary training activities in the event that members of the new company join the supervisory team.
- Review notification procedures if it is necessary to include new decision-makers in the communication chain.
- Evaluate whether the infrastructure on which the Monitoring system is housed should be updated with new connections.
- All those measures aimed at keeping the system and the Monitoring plan active.

5. Other elements to consider

5.1 Connections to other guides

Given the cross-cutting nature of post-market monitoring systems, it is essential to consider the Monitoring measures considered in the rest of the guides.

Specifically:

- **Risk Management Guide:** This guide is the first one that we must consider in our implementation and therefore the first that we must address. Its relationship with the Monitoring Guide is derived from three aspects:
    - Design: After carrying out the risks assessment, we will be able to define what information and data we should collect from the system to design our post-market monitoring system indicators (See the section on: "What elements should I implement and how should I do it to develop an adequate risk management system?" of the Risks Management Guide).
    - Supervision: The post-market monitoring system should offer us indicators to know if our system is approaching a risk situation.
    - Feedback: The activity of the post-market monitoring system itself will provide *feedback* on when the risks assessment developed should be updated based on changes in the selected indicators. (Article 9.2.c)
- **Records Guide:** The list is found in the Records Guide, where it is indicated that we must define the information and data that we will need to collect from the system. After this definition, we will move on to designing the necessary records. Therefore, this Monitoring Guide provides us with the minimum indicators that we must collect from the system (See the annex of this guide: "Monitoring indicators") and the Records Retention Guide tells us how to generate these records at a technical level (See section of the Records Guide: "What elements should I implement and how should I do it to develop an adequate records management system?").
- **Human Oversight Guide:** The relationship is within the applicable measures of human Monitoring. Specifically, the measures in the Human oversight Guide: "Applicable measures" must be completed in order to, for example, have the users in charge of monitoring the system.
- **Accuracy Guide:** Reference is made to the need to evaluate the degradation of the model through monitoring with dashboards and accuracy visualization tools. To do this, the accuracy metrics associated with the AI model must be previously selected (section of the Accuracy Guide: "Accuracy metrics associated with the AI model") and the relevant accuracy levels for the system (section of the Accuracy Guide: "Accuracy Levels, Documentation and Monitoring of AI models").
- **Robustness Guide** The development of this Guide must be prior to post-marketing Monitoring since it will specify:
    - The metrics selected for system robustness monitoring to be monitored (section of the Robustness Guide: "Selecting Metrics").

  - Validation and verification methods (section of the Robustness Guide: "Validation and verification").
  - Aspects related to the monitoring of robustness indicators (section of the Robustness Guide: "Robustness monitoring").
- **Cybersecurity Guide:** It is indicated that the provider of the intelligent system must apply the cybersecurity measures of said Guide throughout the life cycle of the system. For this reason, monitoring measures must be established for the aspects described in the headings of the Cybersecurity Guide in the form of indicators (See Annex A.IV "Security indicators" of this guide) or through the review of records manually or automatically.
- **Conformity Assessment Guide:** The conformity assessment process extends throughout the life cycle of the intelligent system to ensure that these criteria are maintained over time. In particular, the provider shall verify that the post-market monitoring system is consistent with the technical documentation in order to be able to complete the conformity assessment.
- **Technical Documentation Guide:** In the "Post-market Evaluation System" section of the Technical Documentation Guide, it is indicated that the following must be generated:
  - Detailed documentation of the post-market monitoring system.
  - Documentation on the measures taken to monitor and cover the risks detected must be adequately documented.

6. Technical documentation

According to Annex IV of the European Regulation on Artificial Intelligence, in Article 9, the technical documentation must include a detailed description of the post-market monitoring system and the monitoring plan:

1. **Detailed documentation on the monitoring system**. It must include at least the following aspects:
    a. Selected indicators: data from which it is obtained, normality scale and associated monitored risk if applicable.
    b. Capture and sending systems: framework or technology used for collection and sending, correspondence of indicators with the collection and sending system and periodicity of shipments.
    c. Records of indicators: selected recording system, recording format and storage times.
    d. Alert system and analysis interface for supervisors: list with all defined alerts including the indicator or indicators on which they base their trigger and description of the analysis interface implemented.

2. **Detailed documentation on the measures of the post-market monitoring plan**. It must include at least:
    a. List of actions carried out in the continuous monitoring of the system: description of the task, its objective, periodicity of execution, related responsible and method of recording/communicating the result.
    b. List of actions carried out in the periodic monitoring of the system: description of the task, its objective, periodicity of execution, person in charge and method of recording/communicating the result.
    c. Draft incident report: Include a copy of the draft incident.
    d. List of training activities of those involved in the Monitoring plan: date, description of the activity, objective, duration, roles involved and method of recording/communicating the result.

7. Annexes

7.1 Annex A - Monitoring indicators

The definition of an indicator depends largely on the context of its application. However, within the scope of a monitoring system, an indicator can be considered as a piece of data within a scale that allows its impact to be measured with respect to one or more specific consequences. As an example, our intelligent system responsible for administering insulin to the user records a **data of** the current blood sugar level: 115 mg/dL. How is this data transformed into an indicator? In this case, to be an **indicator**, you must know:

1. The data scale: the minimum (70 mg/dL) and maximum (99 mg/dL) acceptable sugar level.
2. The impact of this data: evaluate what implications the current measurement (115 mg/dL) has on the consequences or scenarios we want to control: from 99 mg/dL the sugar level is unusually high, and an in-depth evaluation of the user's condition must be carried out.

Another example related to the same system would be the number of predictions indicator. The **data** would be the number of predictions generated by the system during the last minute: 128 predictions. To transform it into **an indicator** we must establish:

1. The scale of the data: a minimum of 90 predictions and a maximum of 180 predictions per minute are expected.
2. The impact of this data: In this case, the data is within the usual range of activity of the system. However, if it were higher than 180 it could alert us to potential unreported system errors or if it was lower than 90 it could alert us to a possible unusually high system workload.

Below is a **list of different examples of indicators** that should be selected based on the risks assessment carried out (See section the Risks Management Guide: "What elements should I implement and how should I do it to develop an adequate risks management system?") and transformed into records (See section of the Records Guide: "What elements should I implement and how should I do so in order to develop an adequate records management system?"):

Some explanatory notes on the proposed indicator lists:

- It is not an exhaustive list but a minimum list on which to expand indicators depending on the specific intelligent system.
- In cases where the indicator obtains an average data, it is also advisable to obtain the minimum and maximum value to record anomalous point values of the indicators. For example, if you get the average number of predictions per minute, you also want to get the minimum number of predictions made per minute and the maximum. In short, it is advisable to include all the statistical measures necessary for proper monitoring of the indicator.

- It is especially important to highlight that cybersecurity indicators are a list of examples that must be adapted and expanded by those responsible for the security of the intelligent system based on their application context. Therefore, these are not minimum indicators but general examples.
- The minimum and maximum thresholds established in the scale must be accompanied by a system for sending alerts to act in real time in the event of any type of contingency. It is also advisable to have a *real*-time monitoring dashboard of the measurements obtained from the system.
- In some cases, having indicators alone is not enough and a manual review of the records should be considered depending on their complexity. For example, the number of different session sources for a user (let's say 15 logins in different places) should not only be evaluated through an indicator but in many cases through a manual or automated review of those sources. The aim is for the indicator to act as an alert or trigger for such a review.

7.1.1 Annex A.I - Intelligent System Indicators

INDICATOR	DATA	SCALE
Predictions made per unit of time	Number of predictions made. Example: 1,567 predictions at the last minute.	Minimum expected predictions and maximum number of predictions without performance loss. Example: minimum 200 and maximum 22,000.
Average Prediction Time	Time between the introduction of input into the intelligent system and its response. Example: 0.65 seconds.	Minimum expected time for a prediction and maximum acceptable time. Example: 0.2 seconds / 2.5 seconds
Processing queue	Number of tasks waiting to be processed by the model in real time. Example: 10 tasks in queue.	Acceptable processing queue range. Example: 0 - 50 tasks.
Prediction Error Rate*	Percentage of incorrect predictions out of the total predictions made. Example: 3%	Tolerance range for errors. Example: 0% - 10%

Model accuracy*	Percentage of correct predictions out of the total predictions made. Example: 97%	Desired accuracy range. Example: 85% - 100%
Recall *	Percentage of true positives over the total number of actual positive cases (true positives and false negatives). Especially important in the case of unbalanced problems. Example: 90%	Acceptable Recall *Range*. Example: 80% - 100%
F1 Model Score*	Harmonic average of accuracy and completeness. Example: 0.85	Acceptable range of F1 Score. Example: 0.7 - 1.0
Variation in statistical measures of inputs	It depends on the particular input, but conceptually it is the recording of statistical measurements such as the mean, median and standard deviation of the input values to detect considerable variations in real time. Example for a system that receives text as input: Mean of 6.8 tokens / Standard deviation of 2.1	The acceptable range of statistical measures selected for the particular input. Example: Average of minimum 1 and maximum 100.

* In many cases, measurements related to system prediction errors cannot be obtained in real time and cannot therefore be applied in continuous monitoring but in periodic monitoring described in section 4.2 of this Guide "Periodic monitoring". As an example, we cannot know in real time if the employee promotion system has made a mistake in the assignment of a new position until we check the results and KPIs obtained by the worker over time. In this case, it does not make sense to obtain error rate indicators from the system on an ongoing basis, but to carry out periodic evaluations of the system.

7.1.2 Annex A.II - Infrastructure indicators

INDICATOR	DATA	SCALE
CPU usage	Percentage of CPU utilization in the system. Example: 60%	Acceptable range of CPU usage. Example: 20% - 90%
Average of processes in the system	Average number of processes in the system per unit of time. Example: 433 processes on average/minute.	Acceptable range of active processes on average. Example: 53 - 2500 processes/minute.
RAM usage	Percentage of memory utilization in the system. Example: 70%	Acceptable range of memory usage. Example: 30% - 95%
Using GPUs and similar systems	GPU utilization percentage. Example: 55%	Acceptable range of GPU usage. Example: 20% - 85%
Storage usage	Percentage of storage utilization in the system. Example: 80%	Acceptable range of storage use. Example: 10% - 90%
Uptime	Percentage of time that the system remains operational without interruption. Example: 26 hours.	Desired range of uptime. Example: Maximum of 72 hours.
Network bandwidth	The ability of the network to transmit data per second. Example: 1.62 Gbps	Desired range of network bandwidth. Example: 100 Mbps - 10 Gbps
System Temperature	Average system temperature. Example: 22°C	Acceptable ambient temperature range. Example: 15°C - 30°C

7.1.3 Annex A.III - Indicators on user actions

Indicators related to user actions can be applied both generally for all users and at the user level depending on the monitoring and risks needs of the intelligent system. For example, the number of logins can be applied as a metric for all users in the system, and it can also be applied for each of the users in the system.

INDICATOR	DATA	SCALE
Logins	The number of logins made by users in a given period. Example: 500 daily logins.	Acceptable range of logins. Example: 100 - 1,000 per day.
Number of interactions per user per unit of time	Average number of predictions requested by each user in a given period. Example: 50 predictions per user per day.	Acceptable range of interactions per user. Example: 0 - 1000 a day.
Average time between interactions	Average duration of user sessions on the platform. Example: 15 minutes	Desired range of average time on the platform. Example: 5 - 60 minutes.
Interface interactions	The number of interactions made by users with the interface in a given period. Example: 2,000 interactions per day	Desired range of interactions with the interface. Example: 500 - 5,000 per day.
Statistical measures of inputs entered per user	Same data as "Variation in statistical measures of inputs" from the intelligent system scoreboard but associating these metrics with a particular user.	Same scale as "Variation in statistical measures of inputs" in the table of indicators on the intelligent system but associating these metrics with a particular user.
Number of different session sources	In some cases, it is essential to control the dispersion of logins, both from a technical and cybersecurity perspective. Example: 3 different origins.	Acceptable range of different logon sources per user. Example: 1 - 20.

7.1.4 Annex A.IV - Security indicators

The following cybersecurity indicators are some examples that should be expanded by those responsible for the security of the intelligent system based on their application context. In addition, the indicators should be combined with manual analysis of the records obtained for a complete understanding of each scenario. Following the recommendations of the Cybersecurity Guide:

INDICATOR	DATA	SCALE
Failed login attempts	Number of failed logins attempts in a given period. Example: 50 failures per day.	Acceptable range of failed attempts. Example: 0 - 100 daily
Unauthorized changes to files	The number of unauthorized modifications to system files in a given period. Example: 1 change detected.	High importance from a single change.
Number of active users in the system	Number of users logged into the system at any given time. Example: 10 users.	Acceptable range of logged-in users. Example: 2 - 3 users.
Volume of data transferred on the network	The amount of data transferred per unit of time on the network. Example: 103.2 MB/second.	Acceptable range of data quantity. Example: 10 - 750 MB/second.
Number of open communication s	The number of network connections established per unit of time. Example: 2,498 connections/hour.	Acceptable range of connections. Example: 200 - 20,000 connections/hour.
This list only represents an example with different indicators and must be completed according to the characteristics of the intelligent system and the risks analysis carried out.		

8. References

[1] Akenine-Möller, T., & Johnsson, B. (2012). Performance per what? Journal of Computer Graphics Techniques, 1, 37-41.

[2] Amazon AI. (2021). Sagemaker Clarify: Amazon AI Fairness and Explainability Whitepaper. https://pages.awscloud.com/rs/112-TZM-766/images/Amazon.AI.Fairness.and.Explainability.Whitepaper.pdf

[3] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety.

[4] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115.

[5] Artelt, L. et al. (2021). Evaluating Robustness of Counterfactual Explanations.

[6] Bella, A., Ferri, C., Hernández-Orallo, J., & Ramírez-Quintana, M. J. (2010). Calibration of Machine Learning Models. Department of Computer Systems and Computing, Polytechnic University of Valencia.

[7] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2007). Analysis of representations for domain adaptation. Advances in neural information processing systems.

[8] Bennetot, A. (2022). A Neural-Symbolic learning framework to produce interpretable predictions for image classification (Doctoral dissertation).

[9] Besmira, N., Ece, K., & Eric, H. (2018). Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure. HCOM.

[10] Blouw, P., Choo, X., Hunsberger, E., & Eliasmith, C. (2019). Benchmarking keyword spotting efficiency on neuromorphic hardware. Proceedings of the 7th Annual Neuro-inspired Computational Elements Workshop, 1-8.

[11] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners.

[12] Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In S. A. Friedler & C. Wilson (Eds.), Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Vol. 81). Proceedings of Machine Learn Research.

[13] Burns, K., Hendricks, L. A., Saenko, K., Darrell, T., & Rohrbach, A. (2018). Women also Snowboard: Overcoming Bias in Captioning Models. ECCV'18, 771-787.

[14] Catalog of Bias. (n.d.). Retrieved from https://catalogofbias.org

[15] Chen, N. (2018). Metrics for Deep Generative Models.

[16] Chen, S. F., Beeferman, D., & Rosenfeld, R. (1998). Evaluation Metrics for Language Models.

[17] De Laat, P. B. (2021). Companies Committed to Responsible AI: From Principles towards Implementation and Regulation? Philosophy & Technology, 34, 1135–1193.

[18] Del Ser, J. et al. (2022). Exploring the Trade-off between Plausibility, Change Intensity and Adversarial Power in Counterfactual Explanations using Multi-objective Optimization.

[19] Deloitte. (2020, August 26). Deloitte introduces trustworthy AI framework to guide organizations in ethical application of technology.

[20] Díaz-Rodríguez, N., Lomonaco, V., Filliat, D., & Maltoni, D. (2018). Don't forget, there is more than forgetting: new metrics for Continual Learning. NeurIPS workshop on Continual Learning.

[21] Díaz-Rodríguez, N., Vellido, A., & Moreno, A. (2021). Questioning causality on sex, gender and COVID-19, and identifying bias in large-scale data-driven analyses: the Bias Priority

[22] Dieterich, D. et al. (2016). COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity.

[23] Dietterich, T. G. (2017). Steps Toward Robust Artificial Intelligence.

[24] Dietterich, T. G., & others. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation, 10(7), 1895-1923.

[25] DiveDeep AI. (2022). Data drift vs concept drift. https://divedeep.ai/2022/03/17/data-drift-vs-concept-drift/

[26] Eigner, P. (2021). Towards Resilient Artificial Intelligence: Survey and Research Issues.

[27] ENISA. (2021). SECURING MACHINE LEARNING ALGORITHMS.

[28] Epstein, Z., et al. (2018). Turingbox: an experimental platform for the evaluation of AI systems. IJCAI International Joint Conference on Artificial Intelligence, 2018-July, 5826-5828.

[29] Fabrizzi, L. et al. (2022). A survey on bias in visual datasets.

[30] Ferrario, A. et al. (2022). The Robustness of Counterfactual Explanations Over Time.

[31] Franklin, et al. (2022). An Ontology for Fairness Metrics. Retrieved from https://dl.acm.org/doi/pdf/10.1145/3514094.3534137

[32] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM computing surveys (CSUR), 46(4), 1-37.

[33] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for Datasets.

[34] Gendered Innovations. (2018). Facial Recognition: Analyzing Gender and Intersectionality in Machine Learning. Retrieved from http://genderedinnovations.stanford.edu/case-studies/facial.html#tabs-2

[35] Gloor, L. (2016). Suffering-focused AI safety: In favor of "fail-safe" measures. Center on Long-Term Risk Report.

[36] Google. (2021). Machine Learning Glossary: Fairness. Retrieved November 29, 2021, from https://developers.google.com/machine-learning/glossary/fairness.

[37] HCAI. (2022). Human-centred artificial intelligence. Retrieved from https://scilog.fwf.ac.at/en/environment-and-technology/15317/human-centred-artificial-intelligence

[38] Henderson, P. (2017). Deep Reinforcement Learning that Matters.

[39] Hertweck, C., & Räz, T. (2022). Gradual (In)Compatibility of Fairness Criteria. arXiv preprint arXiv:2109.04399.

[40] Heuillet, A., Couthouis, F., & Díaz-Rodríguez, N. (2021). Explainability in Deep Reinforcement Learning. Knowledge-Based Systems, 214, 106685.

[41] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network.

[42] Hockert, T. (2010). Safeguard By Design: Lessons Learned from DOE Experience Integrating Safety in Design.

[43] Holzinger, A. (2016). Interactive Machine Learning for Health Informatics: When do we need the human-in-the-loop? Brain Informatics, 3(2), 119-131. DOI:10.1007/S40708-016-0042-6.

[44] Holzinger, A. et al. (2022). Digital Transformation in Smart Farm and Forest Operations Needs Human-Centered AI: Challenges and Future Directions.

[45] Holzinger, A., Carrington, A., & Müller, H. (2020). Measuring the quality of explanations: the system causability scale (SCS). KI-Künstliche Intelligenz, 34(2), 193-198.

[46] Holzinger, A., Dehmer, M., Emmert-Streib, F., Cucchiara, R., Augenstein, I., ... (2022). Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. Information Fusion, 79, 263-278.

[47] Holzinger, A., Kargl, M., Kipperer, B., Regitnig, P., Plass, M., & Müller, H. (2022). Personas for Artificial Intelligence (AI): An Open Source Toolbox. IEEE Access, 10, 23732-23747.

[48] Holzinger, A., Malle, B., Kieseberg, P., Roth, P. M., Müller, H., Reihs, R., & Zatloukal, K. (2017). Machine Learning and Knowledge Extraction in Digital Pathology needs an integrative approach. In Springer Lecture Notes in Artificial Intelligence (Vol. LNAI 10344). Springer International. doi: 10.1007/978-3-319-69775-8_2

[49] Holzinger, A., Malle, B., Kieseberg, P., Roth, P.M., Müller, H., Reihs, R. & Zatloukal, K. (2017). Towards the Augmented Pathologist: Challenges of Explainable-AI in Digital Pathology. arXiv:1712.06657.

[50] Hullermeier, E., Waegeman, W., Pölsterl, S., & Szepesvári, C. (2019). Aleatoric and Epistemic Uncertainty in Machine Learning: A Tutorial Introduction.

[51] Hurwicz, L., & Reiter, S. (2006). Designing Economic Mechanisms.

[52] Huyen, C. (2019). Evaluation Metrics for Language Modeling. The Gradient. https://thegradient.pub/understanding-evaluation-metrics-for-language-models/

[53] IBM. (2021). Uncertainty Quantification 360 Toolkit. Retrieved from https://uq360.mybluemix.net

[54] Information Commissioner's Office. (2020). Guidance on the AI auditing framework: draft guidance for consultation.

[55] ISO/IEC 25000. (2021). Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) and ISO/IEC WD 25059:2021, Software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) Quality Model for AI systems.

[56] Kaczmarek-Majer, K., Casalino, G., Castellano, G. et al. (2022). PLENARY: Explaining black-box models in natural language through fuzzy linguistic summaries. Information, Elsevier.

[57] Kalifou, R. T., Caselles-Dupré, H., Lesort, T., Sun, T., Diaz-Rodriguez, N., & Filliat, D. (2019). Continual reinforcement learning deployed in real-life using policy distillation and sim2real transfer. ICML Workshop on Multi-Task and Lifelong Learning.

[58] Kusters, R., Misevic, D., Berry, H., Cully, A., Le Cunff, Y., Dandoy, L., ... (2020). Interdisciplinary Research in Artificial Intelligence: Challenges and Opportunities. Frontiers in Big Data, 3, 45.

[59] Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., & Ntoutsi, E. (2022). A survey on datasets for fairness-aware machine learning.

[60] Lesort, T., Díaz-Rodríguez, N., Goudet, O., & Filliat, D. (2022). Understanding Continual Learning Settings with Data Distribution Drift Analysis. https://www.youtube.com/watch?v=WFhozvAgnsU

[61] Lesort, T., Lomonaco, V., Stoian, A., Maltoni, D., Filliat, D., & Díaz-Rodríguez, N. (2020). Continual Learning for Robotics: Definition, Framework, Learning Strategies, Opportunities and Challenges. Information Fusion, 220.

[62] Lesort, T., Seurin, M., Li, X., Díaz-Rodríguez, N., & Filliat, D. (2019). Deep Unsupervised state representation learning with robotic priors: a robustness analysis. 2019 International Joint Conference on Neural Networks (IJCNN).

[63] Lopez-Paz, D., Muandet, K., Scholkopf, B., & Tolstikhin, I. O. (2015). Towards a learning theory of cause-effect inference. In F. R. Bach & D. M. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, ICML 2015 (Vol. 37, pp. 1452-1461). JMLR.org. http://proceedings.mlr.press/v37/lopez-paz15.html

[64] Lopez-Paz, D., Nishihara, R., Chintala, S., Scholkopf, B., & Bottou, L. (2017). Discovering causal signals in images. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 58-66). doi: 10.1109/CVPR.2017.14

[65] Lutjens, B., Sutanudjaja, E., Straatsma, M., & Maris, M. (2021). Physically-Consistent Generative Adversarial Networks for Coastal Flood Visualization.

[66] Mallya, A. (2018). Piggyback: Adapting a Single Network to Multiple Tasks by Learning to Mask Weights.

[67] Mauri, L. et al. (2021). STRIDE-AI: An Approach to Identifying Vulnerabilities of Machine Learning Assets. IEEE CSR.

[68] McSherry, F. (2022). Materialize: a platform for building scalable event based systems.

[69] McSherry, F., & Talwar, K. (2008). Mechanism design via Differential Privacy.

[70] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Statt, C., ... & Gebru, T. (2019). Model cards for model reporting. Proceedings of the 2020 conference on fairness, accountability, and transparency.

[71] Morris, J. et al. (2020). TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.

[72] Nobel Prize Committee. (2007). Mechanism Design Theory. The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel.

[73] Orcaa. (2020). It's the age of the algorithm and we have arrived unprepared.

[74] Papernot, N. (2016). Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks.

[75] Pisoni, G., Díaz-Rodríguez, N., Gijlers, H., & Tonolli, L. (2021). Human-Centered Artificial Intelligence for Designing Accessible Cultural Heritage. Applied Sciences, 11(2), 870.

[76] PwC. (2020). PwC Ethical AI Framework.

[77] Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2022). Dataset Shift in Machine Learning.

[78] Raffin, A. (2021). Stable-Baselines3 Reliable Reinforcement Learning Implementations. https://stable-baselines3.readthedocs.io/en/master/

[79] Raffin, A., Hill, A., Lesort, T., Traoré, R., & Díaz-Rodríguez, N. (2018). S-RL Toolbox: Environments, Datasets and Evaluation Metrics for State Representation Learning.

[80] Raffin, A., Hill, A., Traoré, R., Lesort, T., Díaz-Rodríguez, N., & Filliat, D. (2018). NeurIPS workshop on Deep Reinforcement Learning.

[81] Raffin, A., Hill, A., Traoré, R., Lesort, T., Díaz-Rodríguez, N., & Filliat, D. (2018). S-RL Toolbox: Environments, Datasets and Evaluation Metrics for State Representation Learning. NeurIPS workshop on Deep Reinforcement Learning.

[82] Raffin, A., Hill, A., Traoré, R., Lesort, T., Díaz-Rodríguez, N., & Filliat, D. (2019). Decoupling feature extraction from policy learning: assessing benefits of state representation learning in goal based robotics. ICLR 2019 Workshop on Structure & Priors in Reinforcement Learning (SPIRL).

[83] Rahtz, D. (2022). Safe Deep RL in 3D environments using human feedback.

[84] Rodríguez, N.D., Cuéllar, M.P., Lilius, J., & Calvo-Flores, M.D. (2014). A fuzzy ontology for semantic modelling and recognition of human behaviour. Knowledge-Based Systems, 66, 46-60.

[85] Rodríguez, N.D., Cuéllar, M.P., Lilius, J., & Calvo-Flores, M.D. (2014). A survey on ontologies for human behavior recognition. ACM Computing Surveys (CSUR), 46(4), 1-33.

[86] Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., & Saenko, K. (2018). Object Hallucination in Image Captioning. Proceedings of the EMNLP'18.

[87] Ross, A.S., Hughes, M.C., & Doshi-Velez, F. (2017). Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. Proceedings of the IJCAI'17.

[88] Russel, S. (2015). Research priorities for robust and beneficial artificial intelligence.

[89] Schwartz, R. S., Dodge, J., Smith, N. A., & Ettinger, M. (2019). Green AI.

[90] Sena, L. H., et al. (2019). Incremental Bounded Model Checking of Artificial Neural Networks in CUDA.

[91] Shi, et al. (2020). Robustness Verification for Transformers. International Conference on Learning Representations. arXiv:2002.06622

[92] Sotala, K., & Gloor, L. (2017). Superintelligence as a Cause or Cure for Risks of Astronomical Suffering. Informatics, 41, 389–400.

[93] Stanton, S., Izmailov, P., Kirichenko, P., Alemi, A. A., & Wilson, A. G. (2021). Does knowledge distillation really work? Retrieved from https://arxiv.org/abs/2106.05945

[94] Strobel, B., Yoo, S., Papernot, N., & Kumar, S. (2022). Data Privacy and Trustworthy Machine Learning.

[95] Tomkins, S., Isley, S., London, B., & Getoor, L. (2018). Sustainability at scale: towards bridging the intention-behavior gap with sustainable recommendations. Proceedings of the 12th ACM conference on.

[96] Widmer, G. et al. (2022). Towards Human-Compatible XAI: Explaining Data Differentials with Concept Induction over Background Knowledge.

[97] Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. Biometrics Bulletin, 1(6), 80-83.

[98] Wilms, I. et al. (2021). Omitted variable bias: A threat to estimating causal relationships.

[99] Yang, D., Rangwala, H., Johri, A., & Rose, C. P. (2022). Generalized out-of-distribution detection: A survey.

[100] Yoo, S., Yang, E., & Zhang, Y. (2020). Blackbox NLP Workshop track proceedings. EMNLP. Retrieved from https://github.com/QData/TextAttack

[101] Zech, J. et al. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study.

[102] Zheng, S. (2016). Improving the Robustness of Deep Neural Networks via Stability Training

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