11 November 2025

EUROPEAN DATA PROTECTION SUPERVISOR

The EU's independent data protection authority

Guidance for Risk Management of Artificial Intelligence systems

Contents

Executive summary

The development, procurement and deployment of AI systems involving the processing of personal data by European Union Institutions, Bodies, Offices and Agencies (EUIs) raises significant risks to data subjects' fundamental rights and freedoms, including but not limited to privacy and data protection. As the cornerstone of Regulation 2018/1725 (EUDPR),[1] the principle of accountability enshrined in Article 4(2) (for administrative personal data) and Article 71(4) (for operational personal data) requires EUIs to identify and mitigate these risks, as well as to demonstrate how they did so. This is all the more important for AI systems that are the product of intricate supply chains often involving multiple actors processing personal data in different capacities.

This Guidance aims to guide EUIs acting as data controllers in identifying and mitigating some of these risks. More specifically, they focus on the risk of non-compliance with certain data protection principles elicited in the EUDPR for which the mitigation strategies that controllers must implement can be technical in nature – namely fairness, accuracy, data minimisation, security and data subjects' rights. As such, the technical controls listed in this Guidance are by no means exhaustive, and do not exempt EUIs from conducting their own assessment of the risks raised by their specific processing activities. In doing so, it refrains from ranking their likelihood and severity.

First, this document provides an overview of the risk management methodology according to ISO 31000:2018 (Section 2). Second, it outlines the typical development lifecycle of AI systems as well as the different steps involved in their procurement (Section 3). Third, it explores the notions of interpretability and explainability as cross-cutting concerns that condition compliance with all the provisions covered in this Guidance (Section 4). Lastly, it breaks down the four general principles listed above, namely fairness, accuracy, data minimisation and security into specific risks, each of which is then described and paired with technical measures that controllers can implement to mitigate these risks (Section 5).

The EDPS issues this guidance in his role as a data protection supervisory authority and not in his role as market surveillance authority under the AI Act. This guidance is without prejudice to the Artificial Intelligence Act.

[1] Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC [2018] OJ L295/39 https://data.europa.eu/eli/reg/2018/1725/oj.

1 Introduction

1.1 Objective

This document aims to guide EU Institutions, Bodies, Offices and Agencies (EUIs) acting as controllers within the meaning of Article 3(8) of Regulation (EU) 2018/1725 (EUDPR) in **identifying** and **mitigating** some of the risks for data subjects' fundamental rights raised by the processing of personal data when developing, procuring and deploying AI systems.[2] It is intended to complement Part II of the Accountability on the ground toolkit on Data Protection Impact Assessments & Prior Consultation.[3]

It also complements the EDPS Orientations of the use of Generative AI by EUIs for ensuring data protection compliance when using Generative AI systems issued in June 2024, which provides practical advice on how EUIs can ensure compliance with the EUDPR when developing or using generative AI systems.[4] The present document is both broader, as it encompasses all types of AI systems, but also narrower, as it focuses on technical rather than legal mitigation strategies (see Section 1.2).

This document provides an analytical framework for identifying and treating risks that may arise in AI systems, structured according to the data protection principles potentially affected. It does not constitute and shall not be relied upon as a set of compliance guidelines. The sole purpose of this document is to facilitate a systematic assessment of risks from a data protection perspective. In other words, it does not replace the necessary compliance assessment of each AI system to be done by the controller, who must ensure that the risks identified (with the support of this framework) are managed as necessary to meet all the obligations arising under the EUDPR.

The EDPS issues this guidance in his role as a data protection supervisory authority and not in his new role as market surveillance authority under the AI Act.

1.2 Scope

For the purpose of this Guidance, and building on the terminology used in ISO 31000:2018,[5] the notion of "**risk**" is expressed in term of "risk source", "event", "consequence" and "control", where the "**risk source**" refers to the processing of personal data in the context of the procurement, development or deployment of an AI system, the "**event**" refers to a situation in which that processing would impede on data subjects' fundamental rights and freedoms, the "**consequence**" refers to the material or non-material harm this might cause

---

[2] Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC [2018] OJ L295/39 https://data.europa.eu/eli/reg/2018/1725/oj.

[3] EDPS, , *Accountability on the ground Part II: Data Protection Impact Assessments & Prior Consultation*, February 2018, https://www.edps.europa.eu/sites/default/files/publication/18-02-06_accountability_on_the_ground_part_2_en.pdf

[4] European Data Protection Supervisor, *Generative AI and the EUDPR. Orientations for ensuring data protection compliance when using Generative AI systems. (Version 2)*, 28 October 2025, https://www.edps.europa.eu/system/files/2025-10/25-10_28_revised_genai_orientations_en.pdf

[5] International Organization for Standardization, *ISO 31000:2018 Risk management — Guidelines*, Edition 2, 2018, https://www.iso.org/standard/65694.html.

to data subjects,[6] and the "**control**" refers to the mitigation strategies that controllers can put in place to reduce the likelihood of that risk materialising and/or the impact it has on data subjects should it materialise.[7]

Pursuant to Article 1(2) EUDPR, the objective of the Regulation is to protect the rights and freedoms of natural persons, **including but not limited to** privacy and data protection in the context of the processing of their personal data. According to Articles 4(2), 26(1) and 27(1), EUIs are responsible for identifying and mitigating risks to these rights and freedoms raised by their processing activities, and to demonstrate how they did so. This is particularly important when it comes to the procurement, development and deployment of AI systems, all the adverse impacts of which have not yet been assessed. It is therefore crucial for controllers to properly identify and mitigate, for each of their processing activities, the risks these raise for all data subjects' fundamental rights. Compliance with the provisions explicitly laid down in the EUDPR is a **proxy** to achieve that objective. This Guidance therefore sticks to a conceptualisation of the risk in which the "event" is a situation of **non-compliance** with a provision explicitly laid down in the text.

More specifically, this Guidance focuses on the risk of non-compliance with a **select few data protection principles** for which the "controls" that controllers must implement can be technical in nature – namely fairness, accuracy, data minimisation, security and certain data subjects' rights. The EDPS insists on the fact that the list of risks and countermeasures outlined in this Guidance is **not exhaustive**, but merely reflects some of the most pressing issues that controllers must address when procuring, developing and deploying AI systems.

1.3 Audience

The intended audience for this document is EUIs' staff involved in the procurement, development and deployment of AI systems, including software developers, data scientists, IT engineers, IT project managers, Data Protection Officers and Data Protection Coordinators.

[6] Recital 46 EUDPR specifies that "[t]the risk to the rights and freedoms of natural persons, may result from personal data processing which could lead to physical, material or non-material damage, in particular: where the processing may give rise to discrimination, identity theft or fraud, financial loss, damage to the reputation, loss of confidentiality of personal data protected by professional secrecy, unauthorised reversal of pseudonymisation, or any other significant economic or social disadvantage; where data subjects might be deprived of their rights and freedoms or prevented from exercising control over their personal data; where personal data are processed which reveal racial or ethnic origin, political opinions, religion or philosophical beliefs, trade union membership, and the processing of genetic data, data concerning health or data concerning sex life or criminal convictions and offences or related security measures; where personal aspects are evaluated, in particular analysing or predicting aspects concerning performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, in order to create or use personal profiles; where personal data of vulnerable natural persons, in particular of children, are processed; or where processing involves a large amount of personal data and affects a large number of data subjects".

[7] ISO 31000:2018 (section 3.8) uses the term "control" defined as "measure that maintains and/or modifies risk". The EUDPR uses the term "measure". The remainder of this document will use the term measure.

2 Risk Management methodology

According to the ISO 31000:2018, **risk management** is a process by which an organisation can control the risks (see Figure 1). The core of that activity is the **risk assessment** part, during which an organisation successively identifies, analyses and evaluates the risks.[8]

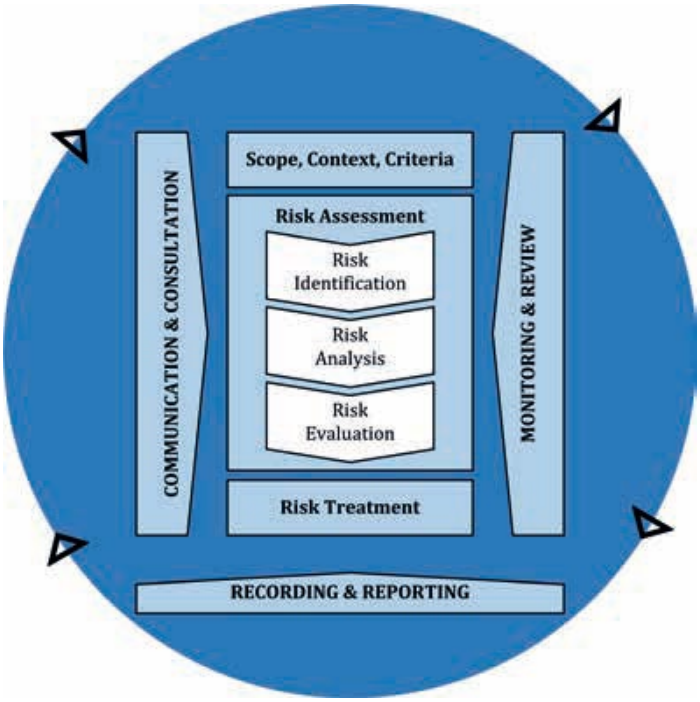


Figure 1: risk assessment[9]

Risk identification involves systematically recognising risks that could potentially affect the organisation's objectives. This step focuses on identifying the sources of risk, the areas of impact, and the events or situations that might lead to uncertainty. The goal is to create a comprehensive risk register that will be further analysed in the following steps. As already hinted at in Section 1.2, this Guidance assumes that the objective pursued by EUIs is to ensure that the processing of personal data with regard to which they act as controllers does not impede on data subjects' fundamental rights.

Risk analysis is the next step, during which the organisation examines the risks identified to understand their nature, their sources, their likelihood, and their potential consequences. This step aims to determine the likelihood of each risk materialising itself, as well as its impact on data subjects should it happen. For qualitative **risk analysis**, the levels of likelihood and impact can be graded on a scale ranging from "Very low", "Low", "Medium",

[8] Isabel Barberá, *AI Possible Risks & Mitigations - Named Entity Recognition,* September *2023,* https://www.edpb.europa.eu/system/files/2024-07/ai-risks_d1named-entity-recognition_edpb-spe-programme_en.pdf
Isabel Barberá, *AI Possible Risks & Mitigations - Optical Character Recognition,* September 2023, https://www.edpb.europa.eu/system/files/2024-06/ai-risks_d2optical-character-recognition_edpb-spe-programme_en_2.pdf
[9] From ISO 31000:2018.

"High", and "Very high".[10] Once each of these elements in the scale is defined,[11] the risk can be evaluated as the product of its likelihood and the severity of its impact (Risk = Likelihood x Impact). This is typically represented via a risk matrix (see Figure 2).

Likelihood	**Very High**	Medium	High	Very high	Very high
	High	Low	High	Very high	Very high
	Low	Low	Medium	High	Very high
	Unlikely	Low	Low	Medium	Very high
		Very limited	**Limited**	**Significant**	**Very significant**
		Severity			

Figure 2: qualitative matrix for risk[12]

Risk evaluation is the final step of the **risk assessment**, where the results of the **risk analysis** are compared against the organisation's risk criteria (such as risk appetite and tolerance) to determine whether each risk is acceptable or requires treatment. The outcome of this evaluation helps the organisation decide whether to avoid, mitigate, transfer, or accept the risks, depending on their severity and organisational goals.

After the **risk assessment** comes the **risk treatment**, the purpose of which is to select and implement measures to mitigate these risks effectively. It is an iterative process that involves several key steps. First, **risk treatment** options are formulated and selected. Then, a plan is developed and implemented to address the identified risks. After implementation, the effectiveness of the measures is assessed to determine whether it has mitigated the risk sufficiently. If the remaining risk is deemed acceptable, no further action is necessary. However, if the risk is still unacceptable, additional measures are taken to further reduce it.

This Guidance focus on **two specific aspects** of the risk management process, namely **risk identification** and **risk treatment**; the **risk analysis** and the **risk evaluation** aspects are too dependent on the specific processing context and their assessment is better left to each organisation in line with their own risk criteria. This means that EUIs should perform a thorough analysis for each AI system they plan to use in order to also evaluate the likelihood and impact of the risks, and decide on the mitigating measures to address them, as well as

[10] In contrast to quantitative risk assessments which require measurements which are often difficult to collect.
[11] Isabel Barberá, *AI Possible Risks & Mitigations - Named Entity Recognition,* September *2023,* https://www.edpb.europa.eu/system/files/2024-07/ai-risks_d1named-entity-recognition_edpb-spe-programme_en.pdf
[12] ibid

on the residual risks.[13] This analysis could even lead to the conclusion that the EUI is unable to mitigate by reasonable means the risks posed by the planned AI system and thus a different solution to the organisation's needs has to be found. In that case, the EUI would have to prior consult the EDPS pursuant to Article 40(1) EUDPR.

3 Understanding the AI lifecycle

3.1 Definition of an AI system

For the purposes of this Guidance, an AI system is understood within the meaning of Article 3(1) of Regulation 2024/1689 (AI Act) as "a machine-based system designed to operate with varying levels of autonomy, that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments".[14] The AI Act, however, does not contain a definition of an "AI model".[15] The terms AI system and AI model are often used as if they were synonyms, when they are not.

AI models are mathematical representations that capture, in a set of parameters, the patterns underlying their training personal data.[16] Although AI models are essential components of AI systems, they do not constitute AI systems on their own, as they will always require other software components to be able to function and interact with users and the virtual or physical environment. In fact, an AI system can be composed of more than one AI model. For example, a voice translator AI system could be composed of a first model transcribing voice data into text, a second model translating the text from one language to another and a third model producing as output voice data from the translated text.

3.2 Lifecycle of an AI system

Risks can appear in different parts of the development lifecycle of an AI system. Thus, it is necessary to understand the specificities of the development lifecycle of an AI system compared to a traditional development lifecycle (for non-AI systems).[17] Different risks may appear in the different phases of the development lifecycle (see Sections 4 and 5). The AI development lifecycle typically comprises the steps detailed in Figure 3.[18]

---

[14] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) EC [2024] OJ L2024/1689 http://data.europa.eu/eli/reg/2024/1689/oj.

[15] Although it defines a general-purpose AI model as "an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications" (Article 3(63) AI Act)

[16] *ISO/IEC 22989:2022 defines an AI model as a "physical, mathematical or otherwise logical representation of a system, entity, phenomenon, process or data".*

[17] See ISO/IEC 15288, ISO/IEC 12207 and ISO 24748:2024.

[18] Details are provided in ISO 22989:2022

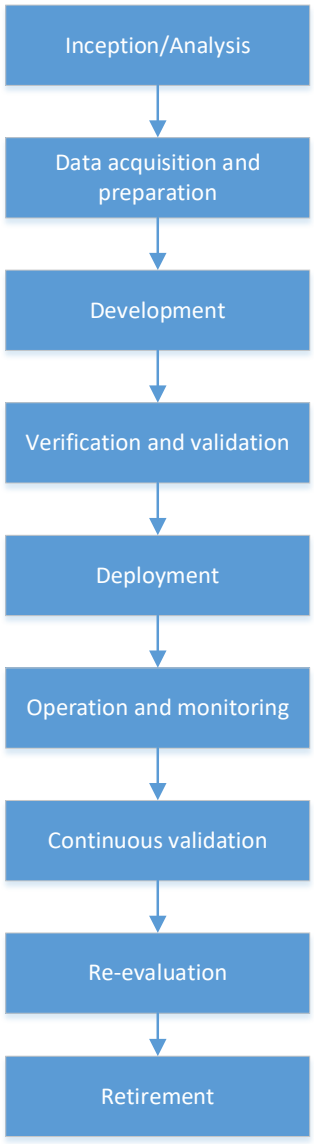


Figure 3: AI development lifecycle

1 Inception/Analysis: This initial stage involves clearly defining the problem the AI system is intended to solve and selecting the AI model architecture.

2 Data acquisition and preparation: The required training personal data depends on the objectives of the AI system. For example, an AI system meant to process images will expect images as training personal data. These images can come from various sources (internet, private databases etc.). The training personal data to be fed to a specific AI system needs to be formatted, checked against applicable quality and legal requirements, and normalised before they can be used.

3 Development: AI systems can be programmed and trained to fulfil specific pre-defined limited functions. This step includes selecting appropriate algorithms, training the AI system on the prepared data, testing (to check if the AI system works and is free of bugs) and tuning its hyperparameters (e.g. learning rate) to improve its performance. Building the AI system might be done with a combination of "libraries" (that could be procured), acquired pre-trained models and internal development.

4 Verification and validation: After the development phase, the AI system is rigorously verified ("Are we building the product correctly?") and validated ("Are we building the right product?") to ensure it meets the functional and non-functional requirements set in the inception stage. This involves checking the AI system's statistical accuracy, robustness, and generalisability using test and validation datasets. Issues related to the AI model during this phase are addressed by retraining the AI system.

5 Deployment: The AI system can be deployed in its final environment (which could be end-user devices, servers, cars, etc.).

6 Operation and monitoring: Once deployed, the AI system is then operated by its users and needs continuous monitoring to ensure it operates as expected. This includes tracking performance, updating the AI system to meet new requirements, and refining it based on feedback.

7 Continuous validation: When an AI system utilises continuous learning,[19] the operation and monitoring phase is extended into an additional phase of continuous validation. In this phase, training is conducted continuously while the system is live in production. The system's performance is regularly assessed using test data to ensure proper functioning. Additionally, the test data may need to be updated periodically to better reflect the current production data, ensuring a more accurate evaluation of the AI system's capabilities.

8 Re-evaluation: After the operation and monitoring phase and the possible continuous validation, it may become necessary to reassess the AI system based on its performance results. The operational results of the system should be thoroughly analysed and compared against the identified risks associated with the AI system to verify if the identified risks have been suitably mitigated. It is possible that, during this phase, risks that have not been previously identified appear. These would need to be treated in the next cycle of the risk management process presented in Section 2.

9 Retirement: An AI system should be responsibly and efficiently decommissioned when it is no longer needed or is replaced by a more advanced solution.

3.3 Procuring an AI system

On top of that, in many cases, building an AI system requires external expertise or the acquisition of a commercial product covering part of the functionality, the data, the security, etc. In those cases it is also important to acknowledge, already in the procurement cycle, making a risk evaluation before any budget is actually committed for a solution that would bring undesirable risks to the organisation.

In these cases, one of several phases are allocated to the AI external provider (who will take care of the provision of the AI system in whole or in part). The EUIs' staff involved in the procurement of AI systems will need to coordinate efforts with the EUIs' staff involved in the deployment of these AI systems (e.g. IT engineers, IT project managers, Data Protection Officers or Data Protection Coordinators) in order to draft the technical part of the call for tender, for the selection of the right product and the execution of the implementation of the AI system.

Several approaches can be taken to integrate a procured AI system into an existing infrastructure. As per Regulation 2024/2509[20], the process followed by EUIs is the following:

1. Publication and transparency (Article 163): Ensure the principles of sound financial management, transparency and equal treatment.
2. Call for tenders (Articles 167-169): Launch an open call for tenders that outlines all necessary specifications. The tender specifications should contain requirements as to the capability of the prospective tenderer to procure the planned AI system and all

[19] Ability to adapt and improve its performance over time by learning from new data without needing to be retrained from scratch.
[20] Regulation (EU, Euratom) 2024/2509 of the European Parliament and of the Council of 23 September 2024 on the financial rules applicable to the general budget of the Union (recast). https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32024R2509

relevant technical and procedural quality guarantees. Among others, information to be required is described in Section 5.3.

3. Selection and award criteria (Article 170): Evaluate offers based on predefined criteria (e.g. price, quality, sustainability).[21]
4. Execution (Article 175): Monitor implementation and ensure compliance.

In this case, the "execution phase" will be comprised of phases similar to some of the phases presented in Section 3.2 namely:

5.1 Verification and validation

5.2 Deployment

5.3 Operation and monitoring

5.4 Continuous validation

5.5 Re-evaluation

5.6 Retirement

Similarly to what happens in the development of AI systems, different risks may appear in the different phases of the procurement lifecycle (see Sections 4 and 5). The different phases where the risks can materialise is indicated for each risk (blue boxes). The corresponding mitigating measures should be devised for each of the indicated phases, if relevant.

[21] Under the accountability principle, it is up to the controller to perform checks with regards to the concerns listed in Sections 4 and 5.

4 Interpretability and explainability as *sine qua non*

Interpretability and **explainability** are cross-cutting concerns when procuring, developing and deploying AI systems. As such, these are prerequisites for EUIs acting as controllers for the processing of personal data in the context of AI systems to ensure compliance with their obligations under the EUDPR. Yet, interpretability and explainability should not be confused with transparency. The former two concepts refer to the extent to which the controller understands the functioning of its AI system. The latter, in turn, refers to the obligation of the controller to provide meaningful information to data subjects. While interpretability and explainability are instrumental for the controller in providing that information, these do not, on their own, suffice to meet the threshold of transparency. The present document addresses the former two notions, as defined below.

Interpretability refers to the degree of human comprehensibility of a given "black box" model or decision. It amounts to the capacity to grasp how an AI model makes its decisions. An interpretable model operates transparently, revealing the connections between its inputs and outputs. When an algorithm is interpretable, a human can explain its workings clearly and understandably. This makes interpretability crucial for ensuring that users can comprehend and trust AI models.

For example, an AI model using linear regression[22] to estimate the price of properties that looks like "Price = 100,000 + (50 × surface_in_square metres) + (10,000 × rooms) + (30,000 × postal_code_score)" is highly interpretable as we can clearly understand the calculation performed.

Explainability in AI concentrates on providing clear and coherent explanations for specific model predictions or decisions. It refers to the ability to clarify how an AI model makes decisions in a way that is accessible to end users. An explainable model offers clear and intuitive explanations for its outputs, helping users understand the reasons behind a particular result. Essentially, explainability emphasises why an algorithm reached a specific decision and how that decision can be justified. Explainability can include post-hoc analysis techniques that summarise or visualise how features influence predictions, even if the model itself is not inherently interpretable.

For example, a convolutional neural network (CNN)[23] used to diagnose pneumonia from chest X-rays is not inherently interpretable due to the complexity of the model's internal workings. However, it is possible to use explainability tools like LIME (Local Interpretable Model-Agnostic Explanations)[24] to generate a heatmap that shows which

[22] In layman's terms, linear regression is a model that estimates the relationship between input and output variables.
[23] Type of deep learning model that uses sliding filters to detect patterns and features in data.
[24] LIME works by perturbing the input data, observing how predictions change, and learning an interpretable model locally around the prediction. This helps building a simple human-understandable AI model. See Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "*Why Should I Trust You?": Explaining the Predictions of Any Classifier.*, 13 August 2016, https://doi.org/10.1145/2939672.2939778

parts of the X-ray the model focused on to make its decision. This explanation helps radiologists understand the model's reasoning, even if the model itself is a black box.

The **difference** between interpretability and explainability is that the former is concerned with understanding the inner workings of AI models, while the latter focuses on explaining the decisions made by those models. Complex AI models, such as deep neural networks, can be difficult to interpret due to their intricate structures and the interactions between different components. In such cases, explainability may be more practical, as it prioritises explaining decisions rather than understanding the model. Finally, interpretability is typically aimed at AI experts and researchers, whereas explainability focuses on effectively communicating model decisions to end users. Therefore, explainability requires a simpler and more intuitive presentation of information. This is necessary to ensure, among other considerations, that:

- The organisation can trust the AI system to perform as expected;
- Errors and biases of the models can be readily identified;
- The organisation can detect when an AI system is being misused;
- The decision-making criteria are in line with the organisation's objectives;
- The AI system is auditable.

4.1.1 Risk 1: Uninterpretable or unexplainable AI system

4.1.1.1 Description

Uninterpretable or unexplainable AI systems pose significant risks because they operate as "black boxes" where the inner-workings and decision-making processes remains opaque to human users, making it difficult to understand how and why certain outputs or decisions were generated.

Note that this risk applies to the following phases of the AI system life cycle:

- Selection (for procuring an AI system)
- Verification and validation (for both developing and procuring an AI system)
- Operation and monitoring (for both developing and procuring an AI system)
- Continuous validation (for both developing and procuring an AI system)
- Re-evaluation (for both developing and procuring an AI system)

4.1.1.2 Possible measures

Possible measures for this risk are:

1. Documentation: Proper documentation should be drafted, including:
    a. What type of AI architecture has been used (decision tree, neural network, etc.) with its specificities (details on the type(s) of AI algorithm(s) used), and an explanation of why this type of model and algorithm have been chosen.
    b. Details on where the training personal data comes from and why it is suitable for the activity at hand.

c. Information on how the AI system acts and how accurate it is across different groups that can be identified in the data.
d. A description of the potential biases, with an explanation of the differences and measures taken to improve overall quality and lower the chances of bias.
e. A description of the limitations of the system, clarifying what expectations should be for what the system can and cannot do.

This documentation is the starting point to explain what the AI does and how it does it; the controller can read the documentation to get information on the workings of the AI system and will be able to see if what the AI system does is fair with regards to the processing of their personal data. The documentation should be relevant, useful and understandable to the user.

2. Consideration of techniques for explainability such as LIME or SHAP (Shapley Additive Explanations)[25].
3. Statistical analysis:[26] Analyse the output of the AI statistically and explain the rationale of the results or lack thereof.

5 Risks associated to main Data Protection Principles

5.1 Principle of fairness

While the EUDPR does not explicitly define fairness, it constitutes a general principle of data protection law enshrined in Article 8(2) of the EU Charter on Fundamental Rights (the Charter).[27] The obligation on controllers to comply with the principle of fair processing is laid down in Article 4(1)(a) EUDPR and, in cases concerning the processing of operational personal data, in Article 71(1)(a). The principle of fairness, while intrinsically linked to that of lawfulness and transparency, has an independent meaning and compliance may be assessed on a standalone basis, irrespective of compliance with other data protection principles.[28]

The EDPB has clarified that fairness is an overarching principle which requires that personal data should not be processed in a way that is unjustifiably detrimental, unlawfully

[25] Method based on cooperative game theory that assigns values to each feature in an AI model. Then, it calculates the contribution of each feature to the prediction for a specific instance, considering all possible feature combinations. This technique provides a unified measure of feature importance and helps explain the AI model's decision. See Lundberg, S. M., & Lee, S. I., *A unified approach to interpreting model predictions. Advances in neural information processing systems*, 25 November 2017, https://arxiv.org/pdf/1705.07874

[26] ICO, "Task 4: Translate the rationale of your system's results into useable and easily understandable reasons", ICO website, 06 August 2025, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/part-2-explaining-ai-in-practice/task-4-translate/

[27] Article 8(2) of the Charter provides that "data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law."

[28] European Data Protection Board, *Guidelines 03/2022 on Deceptive design patterns in social media platform interfaces: how to recognise and avoid them*, 14 February 2023, paragraph 9, https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-032022-deceptive-design-patterns-social-media_en. See also European Data Protection Board, *Binding Decision 2/2023 on the dispute submitted by the Irish SA regarding TikTok Technology Limited (Art. 65 GDPR)*, 2 August 2023, paras 100-107, https://www.edpb.europa.eu/our-work-tools/our-documents/binding-decision-board-art-65/binding-decision-22023-dispute-submitted_en.

discriminatory, unexpected or misleading to the data subject.[29] In order for a processing to be fair, there must be a clear understanding on the part of data subjects of the way in which personal data collected from them will be used and the impacts of that processing. Fairness obliges openness on the part of the controller to ensure that processing does not exceed the reasonable expectations of data subjects.

Yet, fairness imposes obligations beyond transparency requirements. In order to comply with the obligation of fair processing, an assessment should be made of how the processing will affect the interests and fundamental rights of those concerned, as a group and individually, and personal data should not be used in ways that could have unjustified adverse effects on them.[30] It also requires controllers to implement procedural safeguards regarding the collection and processing of data as well as the exercise of balancing rights and interests under the data protection framework. In that sense, fairness is also related to the principle of good administration, which requires EUIs to handle people's affairs "impartially, fairly and within a reasonable time" (Article 41 of the Charter).

In this way, the principle of fair processing underpins the entire data protection framework and seeks to address power asymmetries between the controllers and data subjects in order to cancel out the negative effects of such asymmetries and ensure the effective exercise of data subjects' rights. This is particularly important when personal data are processed in AI systems, the functioning and impact of which might be difficult to grasp even for controllers themselves. Relying on complex AI systems to reach decisions about individuals also makes it more challenging for EUIs to justify and motivate these decisions.

One important risk that could lead to non-compliance with the fairness principle in this context is the existence of bias. As already highlighted in the EDPS' Orientations on generative Artificial Intelligence and personal data protection, "artificial intelligence solutions tend to magnify existing human biases and possibly incorporate new ones".[31] EUIs that rely on AI systems therefore also risk replicating the biases contained in the datasets used to train them, which in turn would lead to discriminatory outcomes. This is particularly problematic when such systems are used to take decisions that affect individuals.

For the purpose of this Guidance, the principle of fairness is understood as requiring controllers to identify, measure and mitigate these biases.[32] To ensure that processing is fair, controllers procuring, developing or deploying AI systems that involve the processing of personal data, especially those used to take or assist when taking decisions about

---

[29] European Data Protection Board, *Guidelines on Data Protection by Design and by Default*, 20 October 2020, paragraph 69, https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-42019-article-25-data-protection-design-and_en.

[30] European Data Protection Board, Gui*delines 2/2019 on the processing of personal data under Article 6(1)(b) GDPR in the context of the provision of online services to data subjects*, Version 2.0, 8 October 2019, paragraph 12, https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-22019-processing-personal-data-under-article-61b_en.

[31] European Data Protection Supervisor, *Generative AI and the EUDPR. First EDPS Orientations for ensuring data protection compliance when using Generative AI systems (Version 2)*, 28 October 2025, section 13 (p31), https://www.edps.europa.eu/system/files/2025-10/25-10_28_revised_genai_orientations_en.pdf.

[32] *The AI Act include specific requirements within Article 10 to address similar issues in terms of biases and representative datasets.*

individuals, should identify and measure these biases and implement technical and organisational measures to prevent or correct any form of discriminatory outcome.

There is no single, universally accepted definition of bias in AI. However, it commonly refers to producing unfair, prejudiced, or systematically incorrect results that favour or discriminate against certain groups of people or types of inputs.[33]

Bias in AI systems can produce results that integrate prejudiced viewpoints (e.g. unfairly focusing on a racial or ethnic population in a police context) or unfair preferences (e.g. disproportionately approving loan applications from higher income post codes).

Some of the root causes of bias in an AI context can be:[34]

- Algorithmic bias: The design of the AI system itself can produce biased results. The decision to use certain AI models and algorithms and include certain information in the development of the AI system can lead to unfair results.
- The training personal data: AI systems learn from training personal data. In order to train an AI system effectively, the training personal data typically comes in large quantities.[35] If the training personal data are biased, then the AI system will learn that bias and similarly produce biased results. For example, men have historically occupied certain job posts. An AI system trained with historical data might retain such historical bias and learn that men are the most suitable candidates for those types of jobs. Facial recognition systems trained faces of individuals belonging to a certain demography will similarly struggle to achieve a high statistical accuracy when confronted to faces of individuals un- or underrepresented in the training personal dataset.
- Other human bias: The developers, or people in charge of training or using an AI system can introduce their own conscious or unconscious biases into the design or implementation of AI systems. For example, if some part of the training process requires a human to review parts of the results, the individual may choose to reject or accept some results based on their own explicit or unconscious biases.

5.1.1 Risk 1: Bias caused by the lack of data quality in training personal data

5.1.1.1 Description

AI systems require quality data to be employed during its training phase because AI systems operate on the principle of "garbage in, garbage out".[36] Inaccurate or incomplete training personal data sets can lead to erroneous outputs from AI systems. For instance, training an

[33] IBM, "What is AI bias?", IBM website, 06 August 2025, https://www.ibm.com/think/topics/ai-bias
[34] Other sources of bias exist. See e.g. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). *A survey on bias and fairness in machine learning. ACM computing surveys (CSUR)*, 2022, https://arxiv.org/abs/1908.09635
[35] For example, image classification can require millions of images; Large Language Models are typically trained on billions or trillions of pieces of text called tokens.
[36] This principle refers to the idea that, in any system, the quality of the output is determined by the quality of the input.

image recognition program on a dataset with mislabelled information would result in the program replicating those errors and providing wrong labels.[37]

This risk applies to the following phases of the AI system life cycle:

- Data acquisition and preparation (for developing an AI system)
- Verification and validation (for both developing and procuring an AI system)
- Re-evaluation (for both developing and procuring an AI system)

5.1.1.2 Possible measures

To address data quality risks, it is essential to ensure that the training personal datasets used in the data acquisition and preparation phase are diverse and representative of the population in which the AI system will be used. This requires efforts to collect data from a wide range of sources and to include underrepresented groups. Regular audits and updates of training personal datasets can help maintain their relevance and inclusivity.

1. Define a quality assurance policy for the training personal dataset which:
    a. Describes the types of data to be collected and the methods for data acquisition.
    b. Describes the steps taken during training personal data preparation (cleaning,[38] labelling,[39] normalisation and scaling,[40] splitting[41]).
    c. Gives a definition of the quality criteria and measures.
    d. Defines the quality threshold.
2. Define and implement a procedure for assessing the training personal dataset which, in accordance with the policy, samples the dataset, measures and assesses against the agreed quality threshold.
3. Conduct regular data quality audits of the AI system's training personal data to check for data quality.
4. Employing statistical techniques to detect outliers, which would need to be checked to determine if they are valid (and should be left in the training personal data) or if they are incorrect (and then should be removed). For instance, if we are dealing with training personal data containing dates of birth, dates indicating an individual of more than 100 years old should be scrutinised. Once identified, these errors can be corrected through manual intervention (if needed), statistical techniques (e.g. estimating potential values by calculating the mean or median of the other values, regression techniques (e.g. linear, polynomial, etc.) or more evolved statistical techniques such as K-Nearest Neighbours where similar training personal data are used to correct deviant values), or even removal from the training personal data if deemed necessary.

[37] Process of assigning meaningful tags or annotations to raw data (such as images, text or audio) to indicate the correct output or category, which is used to train supervised machine learning models to make accurate predictions.
[38] Remove or correct errors, duplicates or irrelevant information in the dataset, such as missing values, outliers or inconsistencies.
[39] Verify that the labels are accurate to ensure the AI learns the correct associations.
[40] Standardise the dataset by scaling features to ensure uniformity and prevent certain attributes from disproportionately influencing the model.
[41] Divide the dataset into training, validation, and test sets to ensure the model generalises well to unseen data and doesn't overfit.

5. The training personal data can be verified to confirm the validity of the information and minimise the risk of bias. This involves getting the data from reliable sources and also checking the training personal data against trusted similar sources if available, doing a human review by some individuals with domain expertise, using fuzzy matching techniques (where we are checking for closeness of the training personal data entries), using statistical techniques to detect clusters of training personal data (to identify potential biases) and documenting the provenance of the training personal data, justifying its correctness,[42] validity and traceability.[43]
6. Standardisation and consistency can be checked to ensure that all training personal data entries are formatted and represented identically (e.g. dates of birth using the same DD/MM/YYYY format).

5.1.2 Risk 2: Bias in training personal data

5.1.2.1 Description

The output of a machine learning model can be biased even when trained with accurate and complete training personal data. Common types of training personal data related bias sources are:[44]

- Sampling errors during data collection can lead to population bias. These errors occur when the training personal dataset is not representative of the broader population. For instance, if a healthcare AI system is developed using data predominantly from urban hospitals, it may not perform as well in rural areas where patient demographics and health conditions might differ. The lack of diverse and representative data means that the AI system is not equipped to generalise its predictions across different populations and settings.
- Historical bias refers to pre-existing biases and socio-technical issues present in society that can infiltrate data, even when using techniques to prevent bias. For example, history shows that fewer women Chief Executive Officers (CEOs) were and are in place. An AI that would search for pictures of CEOs could be biased and show mostly pictures of men CEOs.

This risk applies to the following phases of the AI system life cycle:

- Data acquisition and preparation (for developing an AI system)
- Verification and validation (for both developing and procuring an AI system)
- Re-evaluation (for both developing and procuring an AI system)

[42] The AI system's outputs align with the expected or true outcomes.
[43] Whether the AI system is appropriate and functioning as intended within the given context.
[44] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A, *A survey on bias and fairness in machine learning. ACM computing surveys (CSUR)*, 2022, https://arxiv.org/abs/1908.09635
Types of biases are described in Mikołajczyk-Bareła, A., & Grochowski, *A survey on bias in machine learning research*, 2023, https://arxiv.org/abs/2308.11254

5.1.2.2 Possible measures

When the organisation has access to the training personal data, techniques for detecting and correcting bias within the training personal data can be employed. These methods can include statistical adjustments to balance the representation of different groups and algorithmic interventions to mitigate the impact of any detected bias.

1. Representative training personal data is essential for an AI system to produce reliable and unbiased outputs. If the training personal data differs significantly from real-world data, the system is likely to lead to incorrect inferences. This involves:
    a. Distribution matching: Examine if the statistical distributions of key features in both the training and intended input datasets are similar. Tools like histograms, summary statistics, and clustering can identify differences and gaps.
    b. Diversity and coverage: Analyse if the range and diversity of the input data are adequately represented by the training personal data. Representative training personal data should cover all meaningful scenarios, classes and variations present in the real-world operational data.
    c. Validation and cross-validation: Use validation datasets that mimic the expected operational data. Applying cross-validation or rotating through training/validation phases helps to measure if model performance is consistent and robust for unseen data.
    d. Expert review and scenario checks: Subject matter experts should define key variables and test if the training personal data encapsulates all relevant operational aspects. Reviewing for class imbalance or sampling biases is also crucial.
2. Bias-free features:[45] Select features that are less likely to introduce bias. Avoid features that directly encode sensitive attributes like race, gender or socioeconomic status, unless their inclusion is justified and handled carefully. Try to detect features that act as proxy of sensitive attributes (e.g. postal code as proxy of household income).
3. Feature engineering:[46] The features selected for inclusion in the AI model can significantly impact the AI model's behaviour. If certain features are chosen based on biased assumptions, the resulting AI model predictions will reflect these biases. This process requires careful consideration to ensure that the features used are relevant and do not inadvertently introduce bias. In addition, features can be transformed in a way that reduces bias. For example, reweighting or rescaling features can help ensure that no single feature disproportionately influences the AI model's outcomes.[47]
4. Bias audits:[48] Conduct regular audits of the AI system's training personal data to check for biases.

[45] Features are the attributes of the data points in the dataset (e.g. age, post code or height).
[46] Feature engineering is a process of selecting, transforming and creating relevant variables from raw data to improve the performance and interpretability of machine learning models.
[47] Rescaling: Process of readjusting the range or distribution of data values, often before feeding the data into a machine learning model.
Reweighting: Reassignment of numerical values, or "weights," to various inputs, features or connections in an AI model, particularly in machine learning and neural networks.
[48] IBM, "Introducing AI Fairness 360", IBM research website, 06 August 2025, https://research.ibm.com/blog/ai-fairness-360 Aequitas, "The Bias and Fairness Audit Toolkit for Machine Learning", Aequitas website, 06 August 2025. https://dssg.github.io/aequitas/

5. Bias mitigation techniques for data: Bias mitigation techniques for data such as re-weighting,[49] can reduce identified biases in the data used by the AI system.

5.1.3 Risk 3: Overfitting to the training personal data

5.1.3.1 Description

For an AI system, overfitting refers to a tendency to learn the details and noise in the training personal data to such an extent that it tends to reproduce the data it was trained on and negatively affects the AI system's performance on new, unseen data. In other words, overfitting occurs when the model learns the specific details and noise in the training personal data so thoroughly that it effectively memorises the data. This occurs because the AI system becomes overly complex, capturing patterns that are specific to the training personal dataset but do not generalise well to other data, or when the training personal dataset does not have the right minimum size.

This risk applies to the following phases of the AI system life cycle:

- Data acquisition and preparation (for developing an AI system)
- Verification and validation (for both developing and procuring an AI system)
- Operation and monitoring (for both developing and procuring an AI system)
- Continuous validation (for both developing and procuring an AI system)
- Re-evaluation (for both developing and procuring an AI system)

5.1.3.2 Possible measures

Possible measures for this risk are:[50]

1. Early stopping: Early stopping is a technique where the training process is halted as soon as the AI model's performance on the validation set starts to degrade, indicating potential overfitting to the training personal data.
2. Simplification: Simplifying the AI model is also a practical approach to mitigate overfitting. This can involve selecting fewer, more relevant features or pruning the AI model by removing less important parameters or neurons.[51] By reducing the complexity of the AI model, it becomes less likely to overfit to the noise in the training personal data, thereby improving its generalisation capability.

---

[49] Assign different weights to samples from underrepresented groups to ensure they have a fair influence on the model. See Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. *Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-aware Classification. In Proceedings of the 2018 World Wide Web Conference (WWW '18). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE,* 23 April 2018, https://doi.org/10.1145/3178876.3186133

[50] Ying, Xue. (2019), *An Overview of Overfitting and its Solutions. Journal of Physics: Conference Series*, 2019, https://iopscience.iop.org/article/10.1088/1742-6596/1168/2/022022/pdf

[51] Individual measurable property or characteristic of the data that is used by a model to make predictions or classifications.
Domino.ai, "What is a Feature in Machine Learning and Data Science?", Domino.ai website, 06 August 2025, https://domino.ai/data-science-dictionary/feature

3. Regularisation techniques: Regularisation techniques are methods used in machine learning to prevent overfitting by adding a penalty to the model's complexity, ensuring it generalises well to new data. The two most common techniques are L1 and L2 regularisation. L1 regularisation (Lasso) adds a penalty based on the absolute values of the model's weights, encouraging sparsity by driving some weights to zero, which effectively performs feature selection. L2 regularisation (Ridge) adds a penalty based on the square of the weights, shrinking them towards zero without eliminating any, promoting smoothness and preventing overfitting without discarding features. Elastic Net combines both L1 and L2 regularisation to balance feature selection and weight stability, which is especially useful when dealing with highly correlated features.
4. Dropout: Dropout is another regularisation technique commonly used in neural networks, where randomly selected neurons are ignored during training. This prevents the AI model from becoming overly dependent on specific neurons and encourages the network to learn more robust features.

5.1.4 Risk 4: Algorithmic bias

5.1.4.1 Description

Algorithmic bias is defined as bias emerging from the design of the AI system itself, independent of the input and training personal data used.[52]

The way an algorithm is designed can also lead to biased decisions. For example, the choice of mathematical functions used to optimise the algorithm's performance, the methods used to prevent overfitting, and the decision to apply statistical models to the entire dataset or to specific subgroups can all introduce bias. The COMPAS model,[53] used to predict recidivism rates in the US justice system, was found to have a bias against African American defendants. One of the causes was that the model wrongly assumed a linear relationship between some features and the prediction.[54] Additionally, using statistical methods that are prone to bias can also affect the algorithm's outcome. These design choices can influence the algorithm's decisions, leading to biased results that may unfairly favour or disadvantage certain groups.

This risk applies to the following phases of the AI system life cycle:

- Inception/analysis (for developing an AI system)
- Verification and validation (for both developing and procuring an AI system)
- Continuous validation (for both developing and procuring an AI system)
- Re-evaluation (for both developing and procuring an AI system)

[52] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). *A survey on bias and fairness in machine learning. ACM computing surveys (CSUR)*, 25 January 2022, https://arxiv.org/pdf/1908.09635
[53] Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)
[54] Cynthia Rudin, Caroline Wang, and Beau Coker, *The Age of Secrecy and Unfairness in Recidivism Prediction*, 31 May 2020, https://hdsr.mitpress.mit.edu/pub/7z10o269/release/7

5.1.4.2 Possible measures

It is important to include fairness-aware algorithms, balanced objective functions, careful feature engineering, regular audits, transparency, fairness metrics,[55] and inclusive development practices to create AI systems that are equitable and perform fairly across diverse populations.[56]

1. Fairness-aware algorithms:[57] Choose algorithms that are designed with fairness constraints in mind. Some algorithms, such as Two Naive Bayes,[58] are specifically developed to address fairness and can help mitigate biases during the inception/analysis phase.
2. Selection of the objective function:[59] Algorithms are typically optimised to achieve specific goals defined by objective functions, such as maximising accuracy or minimising error. However, these functions can introduce bias if they do not account for fairness across different groups. For example, an algorithm optimised solely for overall accuracy might neglect the performance disparity between majority and minority groups, leading to biased outcomes.
3. Bias audits:[60] Conduct regular audits of the AI system to check for biases. This involves evaluating the algorithm's performance across different demographic groups and identifying and assessing any disparities.
4. Testing with diverse data: Test the algorithm on diverse datasets that reflect the variety of real-world scenarios it will encounter. This helps ensure that the AI model performs fairly across different populations.
5. AI model interpretability:[61] Use interpretable AI models or techniques that make complex AI models more understandable. This allows identifying and addressing sources of bias within the AI model.
6. Verify whether the problem at stake could be solved by effectively and efficiently using algorithms other than machine learning or deep learning ones, or integrating them with other approaches, including neurosymbolic AI.[62]

---

[55] Metrics are quantitative measures used to evaluate an AI system's performance and effectiveness across various tasks. Different AI models (e.g. classification, regression, clustering) require different metrics. Often, no single metric provides a complete picture of performance so it is recommended to calculate multiple metrics to evaluate different aspects of the AI system's performance comprehensively.

[56] Building diverse and inclusive teams. See Dr. Moeed Yusuf, *Algorithmic Justice: Bias in Code, Bias in Society*, 2024, https://journalpsa.com/index.php/JPSA/article/view/14/16, and Hernández, E. G, *Towards an ethical and inclusive implementation of artificial intelligence in organizations: a multidimensional framework*, 02 May 2024, https://arxiv.org/abs/2405.01697

[57] Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D, *A comparative study of fairness-enhancing interventions in machine learning. In Proceedings of the conference on fairness, accountability, and transparency* (pp. 329-338), January 2019, https://arxiv.org/pdf/1802.04422

[58] Toon Calders and Sicco Verwer, *Three Naive Bayes Approaches for Discrimination-Free Classification. Data Mining journal; special issue with selected papers from ECML/PKDD*, 2010, https://www.cs.ru.nl/~sicco/papers/dmkd10.pdf

[59] Mathematical formulation used to measure the difference between predicted and actual outcomes, guiding the model's optimisation

[60] IBM, "Introducing AI Fairness 360", IBM research website, 06 August 2025, https://research.ibm.com/blog/ai-fairness-360 Aequitas, "The Bias and Fairness Audit Toolkit for Machine Learning", Aequitas website, 06 August 2025. https://dssg.github.io/aequitas/

[61] Carvalho DV, Pereira EM, Cardoso JS., *Machine Learning Interpretability: A Survey on Methods and Metrics,* 26 July 2019, https://doi.org/10.3390/electronics8080832

[62] Wan, Z., Liu, C. K., Yang, H., Li, C., You, H., Fu, Y., ... & Raychowdhury, A. *Towards cognitive ai systems: a survey and prospective on neuro-symbolic ai*, 02 January 2024, https://arxiv.org/abs/2401.01040

5.1.5 Risk 5: Interpretation bias

5.1.5.1 Description

Interpretation bias occurs when analysts draw incorrect or skewed conclusions from the training personal data and AI model outputs, often influenced by preconceptions or incomplete understanding. Additionally, selective interpretation of performance metrics can obscure issues related to fairness, potentially leading to the deployment of biased AI systems. This can lead to flawed considerations that may have consequences when fixing, fine-tuning or re-training the model.

For example, a healthcare organisation develops an AI-powered diagnostic tool to predict the likelihood of patients having a specific disease based on their medical history, symptoms, and test results. The tool outputs a probability score between 0 and 1, where 1 represents a high likelihood of having the disease. Interpretation bias would be if healthcare providers using the tool misinterpret the output as a definitive diagnosis rather than a probability score. They could assume that a score of 0.8 means the patient definitely has the disease, while a score of 0.2 means the patient is disease-free.

This risk applies to the following phases of the AI system life cycle:

- Verification and validation (for both developing and procuring an AI system)
- Operations and monitoring (for both developing and procuring an AI system)
- Re-evaluation (for both developing and procuring an AI system)

5.1.5.2 Possible measures

Possible measures for this risk are:

1. Diverse team involvement: Involving a diverse team of data scientists, domain experts and stakeholders can provide multiple perspectives on the training personal data and AI model outputs.
2. Clear documentation and communication: Maintaining clear and comprehensive documentation of the data sources, feature selection, pre-processing steps and modelling decisions helps ensure that the analysis process is transparent.
3. AI model explainability techniques: Incorporating AI model explainability techniques such as SHAP, LIME and feature importance analysis can provide insights into how AI models make decisions.[63]
4. Training and awareness: Providing training and awareness on bias, fairness and interpretability to the team members involved in the AI development process can improve the team's ability to identify and address interpretability biases.

[63] EDPS, *TechDispatch #2/2023 on Explainable Artificial Intelligence*, 16 November 2023, https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2023-11-16-techdispatch-22023-explainable-artificial-intelligence_en

5. Bias audits:[64] Conduct regular audits of the interpretation of the output by analysts to check for biases.

5.2 Principle of accuracy

5.2.1 Legal meaning of accuracy in the EUDPR

According to Article 4(1)(d) EUDPR, personal data must be accurate and, where necessary, kept up to date. Every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay. Accuracy under the EUDPR requires controllers to ensure that *the personal data itself* is not incorrect or misleading as to any matter of fact.

5.2.2 Statistical meaning of accuracy in AI development

Contrary to the meaning of the data protection principle of accuracy, in the AI context, accuracy is a performance metric that measures how often an AI system guesses the correct answer divided by the total number of predictions.

Accuracy in AI does not refer to the accuracy of the input personal data or to the predicted personal data itself, but to the performance of the *AI system*.

In this Guidance, we will from now on use the term "accuracy" to refer to the corresponding data protection principle and "statistical accuracy" to refer to the accuracy of an AI system.

5.2.3 Risk 1: Inaccurate personal data output

5.2.3.1 Description

Failing to assess statistical accuracy can create a compliance risk with the principle of data accuracy when it leads to the deployment of AI models that produce inaccurate personal data. When an AI system's outputs are not thoroughly validated, errors can go undetected. Given the wide variety of AI models, there is also a wide variety of metrics that can be used to assess AI models' statistical accuracy. Some examples are provided in Annex 1.

An AI model can generate incorrect or nonsensical information (including personal data) that was neither present in its training personal data or the input it received. This can occur in models like Large Language Models (LLMs), which may "invent" facts or provide confident but false answers. These "hallucinations" arise from the probabilistic nature of AI models, which attempt to predict the most likely output rather than make calculations based on deterministic rules detected and validated in advance.

[64] IBM, "Introducing AI Fairness 360", IBM research website, 06 August 2025, https://research.ibm.com/blog/ai-fairness-360 Aequitas, "The Bias and Fairness Audit Toolkit for Machine Learning", Aequitas website, 06 August 2025. https://dssg.github.io/aequitas/

Furthermore, the statistical accuracy of AI systems is heavily dependent on the quality of the datasets used for training.[65] If the training personal data is inaccurate, incomplete, or biased, the AI system may produce unreliable or flawed results. Since machine learning algorithms learn patterns, behaviours, and associations from the data they are trained on, any errors or misrepresentations in this data can be perpetuated in the AI system's predictions. Note that even AI systems trained with good quality datasets can hallucinate.

This risk applies to the following phases of the AI system life cycle:

- Data acquisition and preparation (for developing an AI system)
- Verification and validation (for both developing and procuring an AI system)
- Operation and monitoring (for both developing and procuring an AI system)
- Re-evaluation (for both developing and procuring an AI system)

5.2.3.2 Possible measures

Possible measures for this risk are:

1. High quality training personal data: High quality training personal data is fundamental to developing accurate and reliable AI models. Since AI systems learn from the data they are trained on, ensuring that the data is well-prepared and clean can significantly improve model statistical accuracy.
2. Edge cases: The AI system should be verified and validated with edge cases (outliers) and adversarial examples to evaluate its resilience and reliability under unusual or challenging conditions.[66]
3. Diverse and representative data: It's essential to collect data from diverse sources and ensure it represents all the possible scenarios the AI system will encounter in production. For example, if you're developing an AI system for facial recognition in an airport, the AI system should be trained on representative images (lighting conditions, facial expressions) and not solely on high-resolution well-lit frontal images.
4. Balanced dataset: A balanced dataset ensures that each category or class in a classification problem is equally represented. For instance, in a medical diagnosis model, there should be an adequate number of both positive and negative cases to prevent the model from becoming biased towards one outcome.
5. Hyperparameter optimization (HPO):[67] HPO involves finding the best set of hyperparameters to improve a model's performance on unseen data. Hyperparameters are configuration settings in machine learning models that are set before training and control aspects of the learning process, such as the model's complexity, learning rate and

[65] Zhou, Y., Tu, F., Sha, K., Ding, J., & Chen, H., *A Survey on Data Quality Dimensions and Tools for Machine Learning*, 28 June 2024, https://arxiv.org/abs/2406.19614v1
Budach, Lukas & Feuerpfeil, Moritz & Ihde, Nina & Nathansen, Andrea & Noack, Nele & Patzlaff, Hendrik & Harmouch, Hazar & Naumann, Felix, *The Effects of Data Quality on ML-Model Performance*, July 2022, https://www.researchgate.net/publication/362386427_The_Effects_of_Data_Quality_on_ML-Model_Performance
[66] And edge case is a problem or situation that occurs only at an extreme (maximum or minimum) operating parameter.
[67] Morales-Hernández, A., Van Nieuwenhuyse, I. & Rojas Gonzalez, S. *A survey on multi-objective hyperparameter optimization algorithms for machine learning*, 24 December 2022, https://doi.org/10.1007/s10462-022-10359-2

regularisation (constraining large weights or parameters), but are not learned from the data.

6. Human Oversight (Human-AI collaboration (HAIC) and Human in-the-loop (HITL)):[68] incorporating human review into the AI decision-making process ensures that the model's predictions are double-checked, reducing the chances of errors. Human review of AI systems can take various forms, depending on the context, the complexity of the AI application, and the level of risk associated with its decisions.[69]
7. Verify whether the problem at stake could be solved by effectively and efficiently using algorithms other than machine learning or deep learning ones, or integrating them with other approaches, including neurosymbolic AI.[70]

5.2.4 Specific example: Inaccurate output due to data drift and deterioration of input personal data quality

5.2.4.1 Description

Data drift is understood as changes over time to the statistical properties of input data.[71] It can occur due to various factors such as shifts in user behaviour or changes in AI system operating context. Data drift can cause the AI model to make inaccurate predictions or decisions. The quality of input data can degrade due to issues such as increased noise, missing values or inaccuracies. For instance, a credit scoring AI model trained with data of credits requested during a stable economic situation used in the context of an economic crisis with substantial changes in inflation and unemployment is representative of data drift.

This risk applies to the following phases of the AI system life cycle:

- Operations and monitoring (for both developing and procuring an AI system)
- Continuous validation (for both developing and procuring an AI system)
- Re-evaluation (for both developing and procuring an AI system)

5.2.4.2 Possible measures

Possible measures for this risk are:

---

[68] Fragiadakis, G., Diou, C., Kousiouris, G., & Nikolaidou, M., *Evaluating human-ai collaboration: A review and methodological framework*, 07 March 2025, https://arxiv.org/abs/2407.19098
Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, Liang He, *A survey of human-in-the-loop for machine learning, Future Generation Computer Systems*, October 2022, https://doi.org/10.1016/j.future.2022.05.014

[69] Pre-deployment review involves scrutinising the AI system during its development phase, including validating the quality of training personal data, testing for biases, and ensuring compliance with the legal framework. Human In The Loop (HITL) oversight, incorporates human intervention during the AI's operational phase, allowing humans to monitor, modify, or approve decisions in real time. Post-decision review focuses on evaluating decisions made by the AI after they have been executed, identifying errors or areas for improvement.

[70] EDPS, *Techsonar 2025*, 15 November 2024, https://www.edps.europa.eu/data-protection/our-work/publications/reports/2024-11-15-techsonar-report-2025_en
Wan, Z., Liu, C. K., Yang, H., Li, C., You, H., Fu, Y., ... & Raychowdhury, A, *Towards cognitive ai systems: a survey and prospective on neuro-symbolic ai*, 02 January 2024, https://arxiv.org/abs/2401.01040

[71] GeeksforGeeks. *Data drift in Machine Learning*, 23 July 2025, https://www.geeksforgeeks.org/machine-learning/data-drift-in-machine-learning/

1. Data drift detection methods:[72] Implement data drift detection methods that monitor changes in the distribution of data over time.
2. Data quality monitoring: Implement data quality monitoring systems that track metrics such as completeness, statistical accuracy and consistency of incoming data. Regularly review these metrics to ensure that the data being used for validation remains of high quality.
3. Regular model retraining: This is a process in which machine learning models are periodically updated with new data to ensure their performance remains high as data patterns evolve over time. Retraining can be done on a fixed schedule (e.g. weekly or monthly) or triggered by performance drops or detected drift. The retraining process typically involves collecting new data, pre-processing it, updating the model, and then validating the new version to ensure improved performance.
4. User feedback: Create feedback loops where users can report issues or anomalies in inferences. Use this feedback to identify potential data quality problems or drift and make necessary adjustments to the AI model.

5.2.5 Risk 2: Unclear information from the AI system provider

5.2.5.1 Description

To effectively manage data protection risks when procuring a pre-trained AI system, organisations should focus on understanding the development processes employed by the AI provider.

This involves asking questions regarding the AI system as a whole and digging into the details with regards as to how risks relevant for the inception/analysis, data acquisition and preparation, development and verification, and validation phases (see Section 3.2) were properly managed in order to understand if the final product will meet the organisation's needs.

This risk applies to the following phases of the AI system life cycle:

- Call for tenders (for procuring an AI system)
- Selection (for procuring an AI system)

5.2.5.2 Possible measures

The organisation should invite the provider to produce:

1. General documentation, which covers:
    a. What the AI system does and how it does it: Technical specifications and architecture documentation that detail how the AI system operates, including its

[72] Gemaque RN, Costa AFJ, Giusti R, dos Santos EM. *An overview of unsupervised drift detection methods*, 21 July 2020, https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1381
Andrés L. Suárez-Cetrulo, David Quintana, Alejandro Cervantes, *A survey on machine learning for recurring concept drifting data streams, Expert Systems with Applications*, 01 March 2023, https://doi.org/10.1016/j.eswa.2022.118934

underlying algorithms, data processing methods, features and integration capabilities with other existing systems.

b. User interface and Application Programming Interface (API)[73] details: How the organisation accesses and uses the AI system from a user's and a developer's perspective.

2. Documentation on how the AI system deals with transparency, interpretability and explainability: To avoid operating the AI system as a "black box," where outcomes are generated without clear visibility into how they are derived, the organisation should get information which relates to understanding and explaining how the AI reaches its conclusions, and what factors are driving the AI's decisions, and providing clear, understandable reasons behind specific outcomes.
3. Documentation on cybersecurity measures related to the model integrity: How the integrity of the model was ensured during development and what measures are in place/should be put in place to ensure its continued integrity.
4. Documentation regarding the provider's personal data governance practices, including personal data collection and processing: What kind of data was collected? How was the personal data sourced? How was the personal data used to train the AI model, as well as the methods employed to ensure fairness and accuracy? Many AI providers might provide vague or limited information on the matter. However, some level of transparency is fundamental, including the statistical properties of the training personal data set. The controllers would need to assess whether the demographics of training personal dataset are close or far from the demographics of the personal data that the AI system will ingest during its operation.
5. Validation and testing procedures, and results: How was the model tested and validated across various scenarios and what were the results? What data was used and how were the edge cases handled? Furthermore, the organisation should request from the provider a set of metrics that will allow for the evaluation of the AI system against the organisation's objectives.

   Although metrics are context-dependant,[74] some common metrics are:

   a. False Positive Rate (FPR) Parity: The number of incorrect positive cases can be compared across different groups. This metric should be similar for these groups.
   b. False Negative Rate (FNR) Parity: The number of missed true positives can be compared across different groups. This metric should be similar for these groups.
   c. Calibration Fairness: This metric compares the output of the model with the reality across different groups. The model should perform with similar statistical accuracy across groups.

---

[73] Set of protocols and tools that enables different software applications to communicate and exchange data seamlessly. It acts as an intermediary, allowing various software components to interact without needing to understand the underlying implementation details. APIs simplify the development process by providing predefined methods for accessing specific functionalities or data. For example, the OpenAI API provides access to advanced models for natural language processing, image generation, and other AI tasks. Developers can use this API to integrate functionalities like text generation, summarisation, and conversation capabilities into their applications.

[74] Isabel Barberá, *AI Possible Risks & Mitigations - Named Entity Recognition,* September *2023,* https://www.edpb.europa.eu/system/files/2024-07/ai-risks_d1named-entity-recognition_edpb-spe-programme_en.pdf
Isabel Barberá, *AI Possible Risks & Mitigations - Optical Character Recognition,* September 2023, https://www.edpb.europa.eu/system/files/2024-06/ai-risks_d2optical-character-recognition_edpb-spe-programme_en_2.pdf

d. Equality of Opportunity (EOP): Individual inputs with similar characteristics should have the same chance to get a positive outcome to the model's prediction.

Apart from these common metrics, there are also task-specific benchmarks (e.g. natural language understanding or mathematical problem solving) that could allow for the evaluation of the AI system against the organisation's objectives. Annex I includes a list of some of the most well-known.

5.3 Principle of data minimisation

In order to accurately learn patterns, make reliable predictions, and generalise well to new, unseen data, AI systems are often trained on large datasets. This is necessary to ensure that the AI system is given enough information to learn patterns and create outputs that have statistical properties close to those of the training personal data it received. If training personal data are not representative enough of the input data that it will receive when deployed, the AI system will not be able to produce accurate outputs for some of its input data.

Furthermore, training personal data can come from various sources, be distributed across the globe and belong to different entities/organisations/individuals with different data quality requirements. EUIs must ensure they have a valid legal basis before using personal data to train AI systems.[75]

EUIs must also ensure compliance with the principle of data minimisation. Article 4(1)(c) states that personal data shall be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed (data minimisation)". Thus, a balance is required to provide the AI system with sufficient personal data to function accurately while at the same time limiting the amount of personal data to what is necessary to achieve the purpose pursued by the controller.

5.3.1 Risk 1: Indiscriminate collection and storage of personal data

5.3.1.1 Description

Given that machine-learning models depend on their training personal data, there is a tendency to collect and process as many training personal data as is possible to collect.

[75] European Data Protection Supervisor, *Generative AI and the EUDPR. Orientations for ensuring data protection compliance when using Generative AI systems. (Version 2)*, 28 October 2025, https://www.edps.europa.eu/system/files/2025-10/25-10_28_revised_genai_orientations_en.pdf.
"The EDPS has already cautioned against the use of web scraping techniques to collect personal data, through which individuals may lose control of their personal information when these are collected without their knowledge, against their expectations, and for purposes that are different from those of the original collection. The EDPS has also stressed that the processing of personal data that is publicly available remains subject to EU data protection legislation. In that regard, the use of web scraping techniques to collect data from websites and their use for training purposes might not comply with relevant data protection principles, including data minimisation and the principle of accuracy, insofar as there is no assessment on the reliability of the sources."

Gathering large volumes of data including any possible piece of information without clear criteria or relevance can lead to the accumulation of information that may not be necessary for the AI system's objectives and might be contrary to the data minimisation principle. Data minimisation ensures that only relevant, up-to-date information is stored, preventing the retention of outdated or inaccurate data that could skew decision-making processes or harm individuals.

This risk applies to the following phases of the AI system life cycle:

- Data acquisition and preparation (for developing an AI system)
- Verification and validation (for both developing and procuring an AI system)
- Re-evaluation (for both developing and procuring an AI system)

5.3.1.2 Possible measures

Possible measures for this risk are:

1. Use existing information on the subject matter for a pre-assessment of the type of training personal data that can be useful to draw the needed inferences. Validate the relevance of the planned training personal data types before full training and operation.
2. Data sampling:[76] Sample a representative subset of the training personal data instead of using the full dataset. This approach, known as data sampling, involves selecting a smaller, well-balanced portion of the data that accurately reflects the diversity and key characteristics of the entire dataset. By carefully designing the sample to include all relevant categories and avoid overrepresentation or bias, organisations can train AI models effectively, reducing to the minimum the amount of data they process.
3. Anonymisation/pseudonymisation: The AI system should be developed with anonymised data wherever possible. If personal data are necessary, pseudonymised data should be considered.

5.4 Principle of security

The obligation to ensure the security of personal data is enshrined in Article 4(1)(f) of the EUDPR: "Personal data shall be [...] processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures (integrity and confidentiality)".

IT systems integrating AI components have to consider security threats related to IT systems in general (such as phishing attacks, malware attacks) but also specific security threats related to these AI components.

[76] Daitaku, "ML Models on a Data Diet: How Training Set Size Impacts Performance", Daitaku website, 19 July 2023, 06 August 2025, https://blog.dataiku.com/ml-models-on-a-data-diet
Andrea Montanari, 4 March 2024, "Improving AI via optimal selection of training samples", Granica.ai website, https://granica.ai/blog/improving-ai-via-optimal-selection-of-training-samples

As mentioned previously, this document addresses data protection concerns, risks and measures specifically stemming from the development and use of AI systems. Thus, it will only cover data protection AI-specific security risks (and not general IT systems security risks).[77]

For example, given the quantity of data required to train the AI system, these training personal data are valuable and, if its confidentiality is compromised, they could be used to target individuals whose data was leaked. Confidentiality could be compromised by exploiting specific AI vulnerabilities such as model inversion attacks.[78]

Another example could be to manipulate the training personal data (data poisoning) or the AI model itself (model poisoning) in order to introduce errors into the AI system, such as introducing bias or having the AI produce nonsensical results.[79] Furthermore, the AI model itself could be stolen and then used for malicious purposes.

Thus, in terms of confidentiality and integrity (to ensure that the AI system functions as intended and to protect individuals' personal data), it is necessary to protect the training personal data, the input data, the output data and the AI model itself.

5.4.1 Risk 1: AI system output disclosure of training personal data

5.4.1.1 Description

When AI models are trained on datasets containing personal information, there is a possibility that the model's outputs could unintentionally reveal details about individuals included in the training set. This phenomenon can occur through various privacy attacks,[80] such as model inversion, membership inference or regurgitation of training personal data. In model inversion attacks, an adversary can reconstruct sensitive information by analysing the outputs of the model, effectively revealing personal data associated with individuals in the training personal dataset. Membership inference attacks exploit the confidence scores generated by AI models.[81] If a model exhibits higher confidence in predictions related to specific individuals who were part of the training personal data, attackers can infer that those individuals were included in the dataset. Excerpts from training personal datasets or data included in them could also be regurgitated verbatim when an AI model inadvertently

[77] Yupeng Hu, Wenxin Kuang, Zheng Qin, Kenli Li, Jiliang Zhang, Yansong Gao, Wenjia Li, and Keqin Li, *Artificial Intelligence Security: Threats and Countermeasures*, 23 November 2021, https://doi.org/10.1145/3487890

[78] Type of AI security threat where an attacker exploits the outputs of a machine learning model to infer sensitive information about its training personal data, effectively reverse-engineering the model to reveal confidential attributes of the data it was trained on.

[79] Data poisoning: A cyberattack where an adversary intentionally manipulates a training personal dataset used by an AI or machine learning model, aiming to degrade its performance or alter its behaviour by injecting false information, modifying existing data, or deleting critical data points.
Model poisoning: Intentional manipulation of an AI model's parameters or architecture by adversaries to achieve specific malicious outcomes during the AI model's inference phase.

[80] Fang, H., Qiu, Y., Yu, H., Yu, W., Kong, J., Chong, B., ... & Xia, S. T., *Privacy leakage on DNNs: A survey of model inversion attacks and defenses*, 11 September 2024, https://arxiv.org/abs/2402.04013

[81] Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., & Zhang, X., *Membership inference attacks on machine learning: A survey*, 03 February 2022, https://arxiv.org/abs/2103.07853

reproduces in its output examples or identifiable details from individuals included in its training personal data.

This risk applies to the following phases of the AI system life cycle:

- Operation and monitoring (for both developing and procuring an AI system)
- Continuous validation (for both developing and procuring an AI system)
- Re-evaluation (for both developing and procuring an AI system)

5.4.1.2 Possible measures

Possible measures for this risk are:

1. Training personal data minimisation: Only the necessary personal data should be collected and used. This minimises the risk that identities can be pieced together.
2. Data perturbation techniques: Several techniques can be used to modify the training personal data in order to make re-identification harder, while keeping the training personal data sufficiently accurate for the AI system's purposes:
    a. Generalisation: Some input can be generalised with broader ranges to give fewer possibilities for re-identification. For example, post codes scan be used instead of street addresses, counties instead of city/town names.
    b. Aggregation: Data points can be grouped. For example, large age brackets can be used instead of specific age, income brackets instead of exact income.
    c. Differential privacy/adding noise: Controlled randomness can be introduced into the training personal data (keeping the statistical properties of the training personal data).
3. Synthetic data generation:[82] AI systems could be, at least partially, trained using artificially generated training personal data. These synthetic data reflect the real-world data's statistical properties while not being attributable to an individual.[83] If deemed suitable, this measure should be implemented with due care as it may introduce additional challenges.[84] Thus, if this measure is envisaged, it should be implemented in combination with the additional measures presented above given the possibility of additional attacks (e.g. membership inference attacks).[85]
4. Implement measure to prevent exact replication of training personal data when producing an output such as by using MEMFREE decoding.[86]

---

[82] Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., & Wei, W., *Machine learning for synthetic data generation: a review*, 04 April 2025, https://arxiv.org/abs/2302.04062v9

[83] There is a privacy-utility tradeoff when creating synthetic data. A practical framework to evaluate this tradeoff can be found at reslbesl, tandriamil Nampoina Andriamilanto, emidec Emiliano De Cristofaro, bristena-op "Privacy evaluation framework for synthetic data publishing", 23 June 2021, https://github.com/spring-epfl/synthetic_data_release

[84] Hao, S., Han, W., Jiang, T., Li, Y., Wu, H., Zhong, C., ... & Tang, H., *Synthetic data in AI: Challenges, applications, and ethical implications*, 03 January 2024, https://arxiv.org/abs/2401.01629v1

[85] Van Breugel, B., Sun, H., Qian, Z., & van der Schaar, M., *Membership inference attacks against synthetic data through overfitting detection.*, 24 February 2023, https://arxiv.org/abs/2302.12580

[86] . Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A. Choquette-Choo, Nicholas Carlini, *Preventing Generation of Verbatim Memorization in Language Models Gives a False Sense of Privacy*, 11 September 2023, https://arxiv.org/abs/2210.17546

5.4.2 Risk 2: Personal data storage and personal data breaches

5.4.2.1 Description

The vast amounts of data required for AI systems increases security risks. If the training personal data are somehow compromised (in terms of confidentiality and/or integrity), then the AI system can be severely affected and lead to a data breach within the meaning of Article 3(16) EUDPR.[87] The effect on the overall processing operation will depend on how the data are affected. For example, if the integrity of the data is affected (e.g. data poisoning or evasion attacks), then the AI system might malfunction or provide incorrect results, whereas, if confidentiality is affected, then the personal data that is compromised will affect the individuals (depending on what personal data are compromised, the effects can be financial, health-related, etc.).

This risk applies to the following phases of the AI system life cycle:

- Data acquisition and preparation (for developing an AI system)
- Verification and validation (for both developing and procuring an AI system)
- Operation and monitoring (for both developing and procuring an AI system)
- Continuous validation (for both developing and procuring an AI system)
- Re-evaluation (for both developing and procuring an AI system)

5.4.2.2 Possible measures

Possible measures for this risk are:

1. Using anonymisation and/or pseudonymisation where possible: This will ensure that if a confidentiality data breach occurs, the impact on individuals is minimised. This needs to be balanced with the fact that an AI system needs sufficient quality data to be effective.
2. Encryption: Encrypting the data while it is not being actively used by the AI system to avoid leaking information and protecting the integrity of the data.
3. Synthetic training personal data:[88] Use of synthetic training personal data (in opposition to real data) will ensure that no real data can be compromised in terms of confidentiality. The synthetic training personal data should be built in such a way as to be representative of the real data (e.g. same statistical characteristics) in order for the development phase to align as much as possible to the real use of the final product (ensuring quality of output). This measure should be implemented with due care as it may introduce additional challenges.[89] Thus, if this measure is envisaged, it should be implemented in combination with the

[87] That is, "a breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data transmitted, stored or otherwise processed".
[88] Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., & Wei, W., *Machine learning for synthetic data generation: a review*, 04 April 2025, https://arxiv.org/abs/2302.04062v9
[89] Hao, S., Han, W., Jiang, T., Li, Y., Wu, H., Zhong, C., ... & Tang, H., *Synthetic data in AI: Challenges, applications, and ethical implications*, 03 January 2024, https://arxiv.org/abs/2401.01629v1

additional measures presented above given the possibility of additional attacks (e.g. membership inference attacks).[90]

4. Secure development practices: Follow secure coding practices when developing AI models to prevent attackers from exploiting vulnerabilities in AI code or infrastructure.
5. Multi-Factor Authentication (MFA): Implement MFA for access to sensitive AI systems to prevent unauthorised users from manipulating or stealing models.

5.4.3 Risk 3: Personal data leakage through application programming interfaces

5.4.3.1 Description

Many AI systems are built by using third-party provided AI models accessible through API calls. APIs can be vulnerable to exploitation if not properly secured. Unauthorised access to these APIs can lead to data breaches. For instance, an API might inadvertently expose more data than intended if access controls are not properly defined, or if debugging endpoints that provide detailed system information are left enabled in a production environment.

This risk applies to the following phase of the AI system life cycle:

- Operation and monitoring (for both developing and procuring an AI system)

5.4.3.2 Possible measures

Possible measures for this risk are:

1. Access to APIs: Implement strong authentication mechanisms, such as MFA, to ensure that only authorised users and systems can access the APIs.
2. Role-Based Access Control (RBAC): Enforce RBAC to limit who can access, modify or interact with the AI system based on their role in the organisation.
3. Throttling:[91] This is a technique used to control the number of requests a client can make to an API within a specified time frame to prevent their abuse and mitigate the risk of automated attacks, such as brute force attempts.
4. Communication encryption: Use HTTPS (TLS) to encrypt data transmitted between clients and APIs, ensuring that data are protected from interception and eavesdropping during transit.
5. Logging and monitoring: Implement logging and monitoring for API calls. Use Security Information and Event Management (SIEM) systems to analyse and respond to potential threats in near real-time.

[90] Van Breugel, B., Sun, H., Qian, Z., & van der Schaar, M., *Membership inference attacks against synthetic data through overfitting detection*, 24 February 2023, https://arxiv.org/abs/2302.12580

[91] Technique used to control the number of requests a client can make to an API within a specified time frame, effectively managing traffic and preventing server overload. When a client exceeds the allowed request rate, throttling temporarily blocks or slows down their requests, ensuring fair resource allocation and maintaining overall system performance.

6. Security audits: Conduct regular security audits and penetration testing of APIs to identify and address vulnerabilities. These audits should include code reviews, configuration checks and vulnerability scans, and should use automated tools to continuously scan for common vulnerabilities.
7. Secure development: Follow security best practices in API design such as validating and sanitising input.
8. Patching: Keep API software and underlying infrastructure up to date with the latest security patches and updates.

5.5 Data subject's rights

The EUDPR provides data subjects with various individual rights, namely the right of access (Article 17), to rectification (Article 16), to erasure (Article 19), to restriction (Article 20), to data portability (Article 22) and to object (Article 23). The complex nature of AI systems might make it more challenging for EUIs to act upon these rights, especially when data subjects exercise these rights with regard to the personal data contained in training personal datasets, which the model might then "memorise" and regurgitate at the inference stage. Rather than addressing all the risks related to non-compliance with the provisions governing data subject's rights, this section focuses on cross-cutting technical issues that condition the very possibility for controllers to act upon these rights.

First, the implementation of these rights requires the identification of the processed personal data. For instance, to allow data subjects to access their personal data, these personal data first need to be located in the system. Similarly, to erase some personal data, there is need to identify them first. It is only when the controller has located the personal data in the system that it can provide access, rectify, erase, restrict, and port that data. Second, and looking specifically at the rights to rectification and erasure, the controller must implement measures to actually rectify or erase the personal data it identified. This is particularly challenging in the context of AI systems, where the training personal datasets have been absorbed in the parameters of the model.

5.5.1 Risk 1: Incomplete identification of the personal data processed

5.5.1.1 Description

When personal data is processed by AI systems, data subjects have the right to access their personal data, including the personal data contained in the training personal datasets and potentially retained in the resulting model. When replying to a data subject's request to exercise their rights, the controller must both identify whether and where personal data are used during the training phase, and assess whether the model, when answering certain prompts, might leak that data when making inferences.

For structured datasets,[92] identifying where the data subject's personal data is processed in the training personal dataset can be relatively easy to achieve if the training personal data was retained. For unstructured datasets,[93] the situation will be more complex since there is no fixed schema or format that can be leveraged.

When considering personal data contained in AI models themselves, the complexity of many AI models (notably deep learning models, where data is represented and stored in such a complex way) makes accessing specific data points difficult. Furthermore, due to their very nature where outputs can change with the same request (AI models are, at their core, statistical machines that provide an output selected from a set of possible answers), extracting the complete information cannot be guaranteed.

For example, if Jane Doe wants her data deleted from a model built using deep neural networks, it will be impossible to identify which are the model parameters representing Ms. Doe's related data. Even if those parameters could be located, they might also represent data about other individuals who share some characteristics with Ms. Doe; modifying the parameters to erase Ms. Doe's data might not be possible.

This risk applies to the following phases of the AI system life cycle:

- Development (for developing an AI system)
- Operation and monitoring (for both developing and procuring an AI system)

5.5.1.2 Possible measures

Possible measures for this risk are:[94]

1. Keeping metadata to facilitate the identification of personal data: The training personal datasets should include metadata so that the relevant records or files including personal data can be more easily identified in case a data subject invokes their right to access their personal data. The metadata should include detailed information about the sources of the data and the methods used for their collection. Additionally, they should document any pre-processing steps, such as data cleaning, pseudonymisation or augmentation, to ensure a clear trail of how the data was prepared for training.
2. Data retrieval tools: Data retrieval tools should be created by the AI developers or training personal dataset providers in order to provide a data subject with their personal data when invoking their right to access. These tools should allow data subjects to request and obtain a clear and comprehensive view of their personal data in the training

[92] Data that is organised and formatted in a predefined way, making it easily searchable, stored, and processed by computers.
IBM, "Structured vs. unstructured data: What's the difference?", IBM website, 07 February 2025, 06 August 2025, https://www.ibm.com/think/topics/structured-vs-unstructured-data?utm_source=chatgpt.com
[93] Data that does not have a predefined data model or is not organised in a structured manner (like rows and columns).
IBM, "Structured vs. unstructured data: What's the difference?", IBM website, 07 February 2025, 06 August 2025, https://www.ibm.com/think/topics/structured-vs-unstructured-data?utm_source=chatgpt.com
[94] Dr. Kris Shrishak, "AI: Complex Algorithms and effective Data Protection Supervision", EDPB website, March 2024, https://www.edpb.europa.eu/our-work-tools/our-documents/support-pool-experts-projects/ai-complex-algorithms-and-effective-data_en

personal datasets. Such tools must be designed to securely identify and retrieve individual data records without exposing other users' information. The personal data provided by these tools should be in a machine-readable and user-friendly format.
3. Tools such as MemHunter (automated tool to detect LLM memorisation) could be implemented.[95]

5.5.2 Risk 2: Incomplete rectification or erasure

5.5.2.1 Description

When personal data is processed by AI systems, data subjects have the right to request rectification of incorrect personal data processed by the AI system and/or erasure of their personal data from the AI system. Where data subjects are of the opinion that an AI system has incorrect or incomplete data about them, be it in the training personal dataset or in the output of the AI system, they can request that the organisation corrects it.

The exercise of these two rights suffers from the same difficulties as the exercise of the right of access. Data cannot be erased or rectified if they cannot be identified in the datasets or the model first. For structured datasets, rectifying or erasing data subjects' personal data in the training personal dataset can be relatively easy to achieve if the training personal data was retained (given the structured nature of the data). For unstructured datasets, the situation will be more complex since there is no fixed schema or format that can be leveraged.

Correcting the output of the AI system presents challenges, especially if the data has already been incorporated into complex models like deep neural networks or LLMs. The right to erasure poses similar complications for AI systems. Ensuring that these rectification/erasure requests are fully implemented in the AI models can be complex due to the nature of these AI models. AI models often learn from vast datasets, and once trained, they may retain patterns or information that can be difficult to isolate, rectify and/or erase.

This risk applies to the following phases of the AI system life cycle:

- Development (for developing an AI system)
- Operation and monitoring (for both developing and procuring an AI system)

5.5.2.2 Possible measures

Possible measures for this risk are:

1. Data retrieval tools: Similarly to Section 5.6.1.2, data rectification and erasure tools should be created by the developers or suppliers in order to be able to rectify or erase personal data in the training personal dataset.

[95] Zhenpeng Wu, Jian Lou, Zibin Zheng, Chuan Chen, *MemHunter: Automated and Verifiable Memorization Detection at Dataset-scale in LLMs,* 10 December 2024, https://arxiv.org/html/2412.07261v1

2. Machine unlearning:[96] Machine unlearning is a process that allows machine learning models to selectively forget specific data points that were previously learned, effectively enabling the model to behave as if it had never been trained on that data. Machine unlearning can be categorised into two main approaches: exact unlearning, which involves retraining the model from scratch to completely remove the influence of the specified data, and approximate unlearning, which seeks to minimise the impact of the data through limited updates to the model's parameters. While exact unlearning provides strong guarantees of data removal, it is often expensive (computationally or monetarily because of required changes). In contrast, approximate unlearning offers a more efficient alternative but may not entirely eliminate the data's influence.

3. When machine un-learning is not viable, output filtering can be used. Output filtering involves real-time scanning of AI model responses to detect and block personal information before reaching users. The system could employ pattern recognition or named entity detection to identify personal data and block them before they reach the user.

6 Conclusion

Putting into operation AI systems that process personal data entails significant risks for data subjects, which EUIs, acting as controllers, have a legal and ethical duty to identify, assess and mitigate. The stakes are high: AI systems can amplify risks to fundamental rights at an unprecedented scale and speed if proper safeguards are not embedded from the outset. For this reason, this document has applied a risk management methodology aligned with ISO 31000:2018, contextualised to the specific requirements of the EUDPR, to help ensure that risks are addressed in a systematic and accountable way.

Chapter 2 introduced the risk management methodology as the cornerstone for handling data protection challenges, providing a structured basis for analysing threats and implementing proportionate safeguards. Chapter 3 explored the definition and lifecycle of AI systems, highlighting procurement as a decisive stage where risks can and should be anticipated before systems are put into operation. Chapter 4 examined five data protection principles – transparency, fairness, accuracy, data minimisation, and security – and analysed how risks can manifest to ensure compliance with these principles, together with some risks linked to the effective exercise of data subjects' rights. For each principle, the document identified specific risk scenarios and suggested non-exhaustive technical countermeasures to illustrate how risks can be managed in practice.

The analysis does not attempt to present an exhaustive catalogue of all risks and mitigation strategies. Instead, it provides a practical framework to help EUIs build a systematic approach to risk management that might need to be complemented in the view of other risks and compliance requirements. Therefore, it is intended to be a framework that can be adapted to the diversity of contexts in which AI systems are put in operations. Ultimately, controllers remain fully responsible for performing their own comprehensive compliance

[96] EDPS, *Techsonar 2025*, 15 November 2024, https://www.edps.europa.eu/data-protection/our-work/publications/reports/2024-11-15-techsonar-report-2025_en

assessments, complementing this framework with the necessary legal analysis, and ensuring that their decisions are consistent with the EUDPR.

In conclusion, addressing the risks raised by AI systems is not a peripheral task but a central obligation for EUIs. Compliance with the EUDPR, the protection of fundamental rights, and the preservation of public trust all depend on the controllers' ability to proactively identify, evaluate, and mitigate risks throughout the AI lifecycle. This requires more than a one-off assessment: it demands a culture of accountability, continuous monitoring and adaptive improvement. By embedding these practices into their AI governance, EUIs can not only navigate the complexities of AI development and use AI responsibly but also demonstrate leadership in ensuring that innovation is firmly anchored in the respect for fundamental rights and data protection principles.

Annex 1: Metrics

Metrics for evaluating AI performance vary significantly across different types of AI systems, reflecting the diverse nature of AI applications and their specific goals.[97]

For instance, in natural language processing tasks, metrics like BLEU, ROUGE, GLEU and METEOR are commonly used to assess the quality of generated text or translations. These metrics focus on comparing the AI-generated output to reference texts, measuring aspects such as precision, recall and semantic similarity.

In contrast, classification and identification tasks in AI models often rely on metrics such as statistical accuracy, precision, recall, and F1 score. These metrics evaluate how well an AI model can categorize data into predefined classes or individuals, which is crucial for applications like spam detection or image recognition.

For regression problems, where the AI predicts continuous values, different metrics come into play, such as Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

Benchmarks provide standardised methods to evaluate and compare different AI models.[98] By offering consistent datasets, predefined tasks, and evaluation metrics, benchmarks enable researchers and developers to objectively assess the performance, efficiency and statistical accuracy of AI solutions. Moreover, benchmarks play a vital role in identifying potential risks and limitations of AI models before deployment by identifying areas of concern that need to be tackled during the whole lifecycle of the use of the AI system.

The table below provides a list of some benchmarks available at the publishing of this report that can be used to evaluate AI systems. This list is not intended to be exhaustive but offers a good starting point.

Type of AI	Possible benchmark	Short description
Natural Language Processing (NLP) and Large	Recall-Oriented Understudy for Gisting Evaluation (ROUGE)[100]	Set of metrics and a software package designed to evaluate the quality of automatically generated summaries and machine translations in natural language processing.
	Bilingual Evaluation	Benchmark evaluating the quality of machine-generated text, particularly in machine

[97] OECD.ai, "Catalogue of Tools & Metrics for Trustworthy AI", 06 August 2025, OECD.ai website, https://oecd.ai/en/catalogue/metrics
[98] Github, "Papers with code", Github website, 06 August 2025, https://paperswithcode.com/sota
[100] Neri Van Otten, "ROUGE Metric In NLP: Complete Guide & How To Tutorial In Python", 12 August 2024, 06 August 2025, https://spotintelligence.com/2024/08/12/rouge-metric-in-nlp/

Language Models (LLM)[99]	Understudy (BLEU)[101]	translation tasks. It calculates this similarity by measuring the precision of n-grams (sequences of n consecutive words) that appear in both the generated text and the reference texts.
	General Language Understanding (GLUE)[102] and SuperGLUE[103]	Benchmark datasets designed to evaluate the performance of NLP models across various language understanding tasks. GLUE, introduced first, consists of nine diverse NLP tasks, including sentence classification, sentiment analysis, and textual entailment. SuperGLUE, developed as a more challenging successor to GLUE, builds upon its predecessor by introducing more complex tasks that require advanced reasoning, common-sense knowledge, and contextual understanding. While GLUE focuses on simpler linguistic challenges, SuperGLUE incorporates tasks like question answering, co-reference[104] resolution, and reading comprehension, pushing models to demonstrate higher-level cognitive abilities
	Holistic Evaluation of Language Models (HELM)[105]	Benchmark framework designed to assess the capabilities, limitations, and potential risks of language models across a wide range of scenarios and metrics. The framework evaluates models on 16 core scenarios and 26 targeted scenarios, measuring seven key metrics: statistical accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency
	Massive Multitask Language Understanding	MMLU consists of approximately 16,000 multiple-choice questions spanning 57 academic subjects, including mathematics, philosophy, law, and medicine. The benchmark

[99] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. *A Survey on Evaluation of Large Language Models*, 29 March 2024, https://doi.org/10.1145/3641289

[101] https://spotintelligence.com/2024/08/13/bleu-score-in-nlp/

[102] Wang, A., *Glue: A multi-task benchmark and analysis platform for natural language understanding*, 22 February 2019, https://arxiv.org/abs/1804.07461

[103] Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... & Bowman, S., *Superglue: A stickier benchmark for general-purpose language understanding systems. Advances in neural information processing systems*, 13 February 2020, https://arxiv.org/abs/1905.00537

[104] Identifying and linking expressions in text that refer to the same entity or event

[105] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., ... & Koreeda, Y.. *Holistic evaluation of language models*, 01 October 2023, https://arxiv.org/abs/2211.09110

	(MMLU)[106] and MMLU-Pro[107]	aims to assess AI models' general knowledge and problem-solving abilities, with difficulty levels ranging from elementary to professional. MMLU-Pro features more challenging, reasoning-focused questions and increases the choice set from four to ten options.
Image recognition[108]	ImageNet[109]	Visual database designed for use in visual object recognition research. It contains over 14 million labelled images covering thousands of object categories. The dataset is structured into 1,000 distinct classes, with approximately 1.2 million images used for training, 50,000 for validation, and 100,000 for testing.
	CIFAR-10 and CIFAR-100[110]	CIFAR-10 consists of 60,000 32x32 colour images divided into 10 mutually exclusive classes. The dataset is split into 50,000 training images and 10,000 test images, with each class represented equally. CIFAR-100, while maintaining the same total number of images and image dimensions as CIFAR-10, expands the classification challenge by dividing the dataset into 100 classes. These classes are further grouped into 20 super-classes, providing an additional layer of categorisation. Each image in CIFAR-100 is associated with both a "fine" label (specific class) and a "coarse" label (superclass).

[106] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J., *Measuring massive multitask language understanding*, 12 January 2021, https://arxiv.org/abs/2009.03300
[107] Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., ... & Chen, W., *Mmlu-pro: A more robust and challenging multi-task language understanding benchmark*, 06 November 2024, https://arxiv.org/abs/2406.01574v4
[108] Li, L., Chen, G., Shi, H., Xiao, J., & Chen, L. (2024). *A survey on multimodal benchmarks: In the era of large ai models*, 21 September 2024, https://arxiv.org/abs/2409.18142
Rangel, Gabriela, Cuevas-Tello, Juan C., Nunez-Varela, Jose, Puente, Cesar, Silva-Trujillo, Alejandra G., *A Survey on Convolutional Neural Networks and Their Performance Limitations in Image Recognition Tasks*, 12 July 2024, https://doi.org/10.1155/2024/2797320
[109] Image Net, "Imagenet Database", Image Net website, 11 March 2021, 06 August 2025, https://www.image-net.org/
[110] Alex Krizhevsky, "The CIFAR-10 dataset", Alex Krizhevsky's home page, 06 August 2025, https://www.cs.toronto.edu/~kriz/cifar.html

	MNIST[111]	The dataset consists of 70,000 grayscale images of handwritten digits, each sized 28x28 pixels. It is divided into two subsets: a training set of 60,000 images and a testing set of 10,000 images.

Category	Benchmark	Description	Models Tested
Natural Language Processing (NLP)	**GLUE (General Language Understanding Evaluation)**	A collection of tasks designed to test the general language understanding ability of models.	Text-based models, language models (e.g., BERT, GPT)
	SuperGLUE	An extension of GLUE with more challenging tasks.	Advanced NLP models (e.g., T5, RoBERTa)
	SQuAD (Stanford Question Answering Dataset)	Tests reading comprehension and the ability to answer questions based on a passage.	QA models (e.g., BERT, T5)
	CoNLL-03[112]	Named Entity Recognition (NER) dataset for evaluating entity recognition performance.	NER models (e.g., LSTMs, CRFs)
	MNLI (Multi-Genre Natural Language Inference)	Evaluates models' ability to determine if a premise entails, contradicts, or is neutral to a hypothesis.	Inference models (e.g., BERT, RoBERTa, XLNet)
	TREC (Text REtrieval Conference)	Focuses on question classification tasks.	Text classifiers, intent recognition models
Computer Vision (CV)	**ImageNet**	Large-scale image classification benchmark with a vast number of categories (1000).	CNNs, vision transformers (e.g., ResNet, EfficientNet)

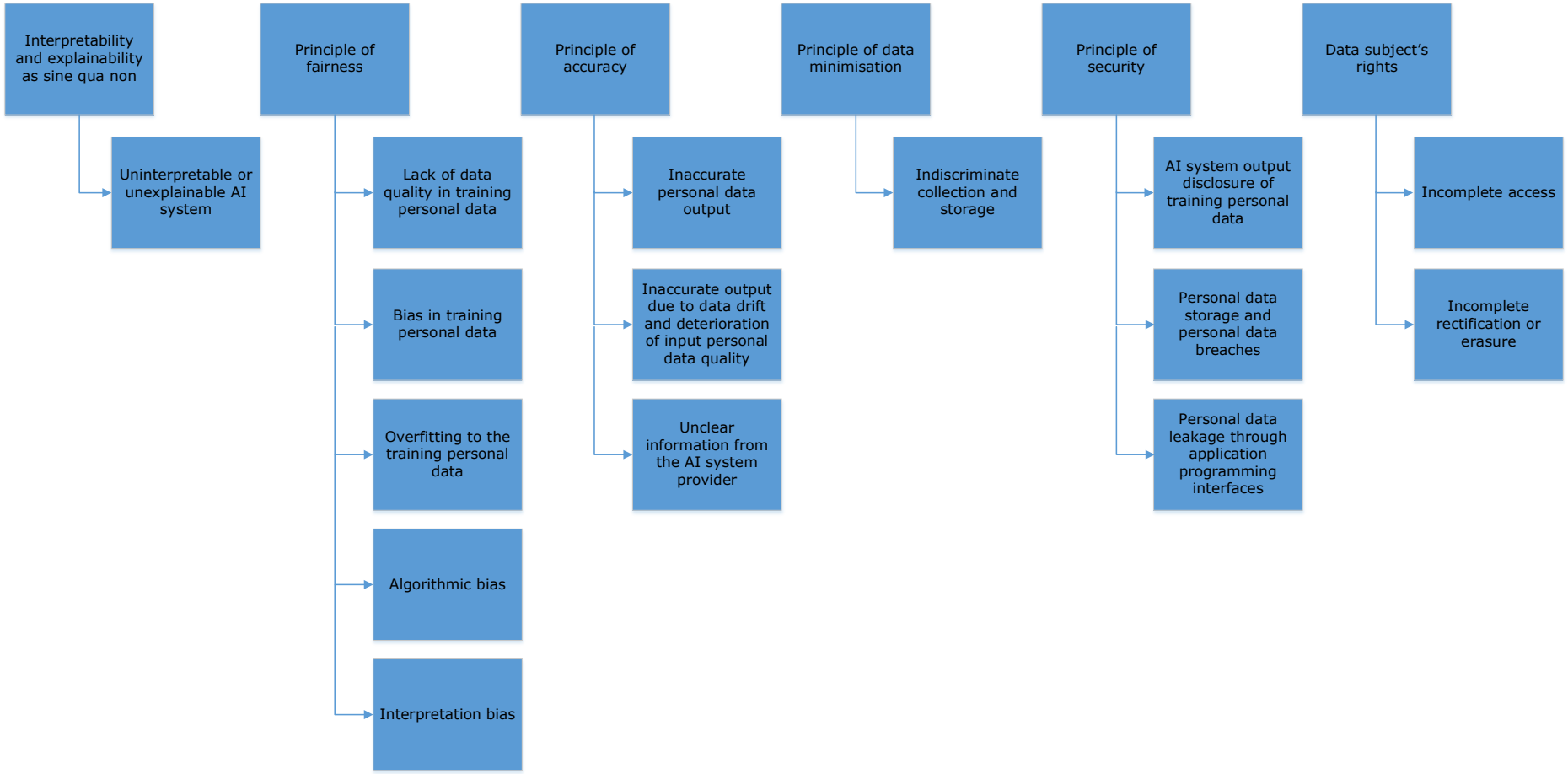
[111] Hojjat Khodabakhsh, "MNIST Dataset", Kaggle website, 06 August 2025, https://www.kaggle.com/datasets/hojjatk/mnist-dataset
[112] Github, "Papers with code", Github website, 06 August 2025, https://paperswithcode.com/cobll-2023

	COCO (Common Objects in Context)[113]	A benchmark for object detection, segmentation, and captioning tasks.	Object detection and segmentation models (e.g., YOLO, Mask R-CNN, Faster R-CNN)
	PASCAL VOC	Focuses on image classification, object detection, and segmentation tasks.	Object detection, segmentation models
	ADE20K	A semantic segmentation benchmark for dense pixel-level annotation.	Segmentation models (e.g., DeepLabV3+, U-Net)
	KITTI	Focused on autonomous driving, including tasks like stereo matching, optical flow, and object detection.	Object detection, optical flow models (e.g., CNNs, RNNs)
Speech and Audio	**LibriSpeech**	Speech recognition benchmark focused on transcribing English audiobooks.	Speech-to-text models (e.g., DeepSpeech, Wav2Vec)
	VoxCeleb	Speaker recognition and identification in audio clips.	Speaker recognition models (e.g., ECAPA-TDNN, VGGVox)
	TIMIT	A corpus for acoustic-phonetic continuous speech recognition.	Speech recognition models
	CHiME	Evaluates speech recognition in noisy environments.	Robust speech recognition models (e.g., RNNs, LSTMs)
Reinforcement Learning	**OpenAI Gym**	A platform for developing and comparing reinforcement learning algorithms across a wide range of environments.	RL agents (e.g., PPO, DDPG, A2C)

[113] Cocodataset, "Cocodataset", Cocodataset website, 06 August 2025, https://cocodataset.org

	MuJoCo	Physics engine used to simulate environments for continuous control tasks.	RL agents for continuous control (e.g., DDPG, TRPO)
	Visual Question Answering (VQA)	Tests a model's ability to answer natural language questions about images.	Vision + NLP models (e.g., LXMERT, ViLBERT)
Multimodal AI	**MS COCO Captioning**	Benchmarks the generation of natural language descriptions of images.	Image captioning models (e.g., Show and Tell, Image Transformer)
	AI2 Reasoning Challenge (ARC)	A benchmark for testing general reasoning skills in multiple-choice questions on a broad range of topics.	General reasoning AI (e.g., GPT, T5)
General AI Performance	**CLIP (Contrastive Language-Image Pretraining)**	Measures how well models can align text and image inputs for tasks like zero-shot classification.	Multimodal AI models (e.g., CLIP, Flamingo)
	WinoBias	Measures biases in language models by analysing their responses to gendered pronouns.	NLP models (e.g., GPT, BERT)
Fairness and Bias	**FairFace**	Evaluates facial recognition models for fairness across different demographics (e.g. skin tone, age, gender).	Face recognition models (e.g., FaceNet, ArcFace)
	MedNLI	A benchmark for evaluating natural language inference in the healthcare domain.	Healthcare NLP models (e.g., BioBERT, ClinicalBERT)
AI for Healthcare	**ChestX-ray14**	Used to evaluate models on detecting 14 common thoracic diseases from X-ray images.	Medical image classification models (e.g., CNNs)
	MIMIC-CXR	A large-scale chest X-ray dataset for evaluating diagnostic statistical accuracy.	Medical image models (e.g., ResNet, DenseNet)

Annex 2: Overview of concerns and risks



Annex 3: Checklist per phase of the AI lifecycle development

Developing and AI system

Phase of the AI lifecycle development	Principle	Risk
1. Inception/Analysis	5.1 Principle of fairness	5.1.4 Algorithmic bias
2. Data acquisition and preparation	5.1 Principle of fairness	5.1.1 Lack of data quality in training personal data
		5.1.2 Bias in training personal data
		5.1.3 Overfitting to the training personal data
	5.2 Principle of accuracy	5.2.3 Inaccurate personal data output
	5.3 Principle of data minimisation	5.3.1 Indiscriminate collection and storage
	5.4 Principle of security	5.4.2 Personal data storage and personal data breaches
3. Development	5.5 Data subject's rights	5.5.1 Incomplete access
		5.5.2 Incomplete rectification or erasure

4. Verification and validation	4 Interpretability and explainability	4.1 Uninterpretable or unexplainable AI system
	5.1 Principle of fairness	5.1.1 Lack of data quality in training personal data
		5.1.2 Bias in training personal data
		5.1.3 Overfitting to the training personal data
		5.1.4 Algorithmic bias
		5.1.5 Interpretation bias
	5.2 Principle of accuracy	5.2.3 Inaccurate personal data output
	5.3 Principle of data minimisation	5.3.1 Indiscriminate collection and storage
	5.4 Principle of security	5.4.2 Personal data storage and personal data breaches
5. Deployment	None	None
6. Operation and monitoring	4 Interpretability and explainability	4.1 Uninterpretable or unexplainable AI system
	5.1 Principle of fairness	5.1.3 Overfitting to the training personal data
		5.1.5 Interpretation bias

	5.2 Principle of accuracy	5.2.3 Inaccurate personal data output
		5.2.4 Inaccurate output due to data drift and deterioration of input personal data quality
	5.4 Principle of security	5.4.1 AI system output disclosure of training personal data
		5.4.2 Personal data storage and personal data breaches
		5.4.3 Personal data leakage through application programming interfaces
	5.5 Data subject's rights	5.5.1 Incomplete access
		5.5.2 Incomplete rectification or erasure
7. Continuous validation	4 Interpretability and explainability	4.1 Uninterpretable or unexplainable AI system
	5.1 Principle of fairness	5.1.3 Overfitting to the training personal data
		5.1.4 Algorithmic bias
	5.2 Principle of accuracy	5.2.4 Inaccurate output due to data drift and deterioration of input personal data quality
	5.4 Principle of security	5.4.1 AI system output disclosure of training personal data

		5.4.2 Personal data storage and personal data breaches
8. Re-evaluation	4 Interpretability and explainability	4.1 Uninterpretable or unexplainable AI system
	5.1 Principle of fairness	5.1.1 Lack of data quality in training personal data
		5.1.2 Bias in training personal data
		5.1.3 Overfitting to the training personal data
		5.1.4 Algorithmic bias
		5.1.5 Interpretation bias
	5.2 Principle of accuracy	5.2.3 Inaccurate personal data output
		5.2.4 Inaccurate output due to data drift and deterioration of input personal data quality
	5.3 Principle of data minimisation	5.3.1 Indiscriminate collection and storage
	5.4 Principle of security	5.4.1 AI system output disclosure of training personal data
		5.4.2 Personal data storage and personal data breaches

9. Retirement	None	None

Procuring an AI system

Phase of the AI lifecycle development	Concern	Risk
1. Preparation	None	None
2. Call for tenders	5.2 Principle of accuracy	5.2.5 Unclear information from the AI system provider
3. Selection	4 Interpretability and explainability	4.1 Uninterpretable or unexplainable AI system
	5.2 Principle of accuracy	Unclear information from the AI system provider
4. Award and Contract	None	None
5. Execution	None	None
6 Verification and validation	4 Interpretability and explainability	4.1 Uninterpretable or unexplainable AI system
	5.1 Principle of fairness	5.1.1 Lack of data quality in training personal data
		5.1.2 Bias in training personal data
		5.1.3 Overfitting to the training personal data
		5.1.4 Algorithmic bias

		5.1.5 Interpretation bias
	5.2 Principle of accuracy	5.2.3 Inaccurate personal data output
	5.3 Principle of data minimisation	5.3.1 Indiscriminate collection and storage
	5.4 Principle of security	5.4.2 Personal data storage and personal data breaches
7 Deployment	None	None
8 Operation and monitoring	4 Interpretability and explainability	4.1 Uninterpretable or unexplainable AI system
	5.1 Principle of fairness	5.1.3 Overfitting to the training personal data
		5.1.5 Interpretation bias
	5.2 Principle of accuracy	5.2.3 Inaccurate personal data output
		5.2.4 Inaccurate output due to data drift and deterioration of input personal data quality
	5.4 Principle of security	5.4.1 AI system output disclosure of training personal data
		5.4.2 Personal data storage and personal data breaches

		5.4.3 Personal data leakage through application programming interfaces
	5.5 Data subject's rights	5.5.1 Incomplete access
		5.5.2 Incomplete rectification or erasure
9 Continuous validation	4 Interpretability and explainability	4.1 Uninterpretable or unexplainable AI system
	5.1 Principle of fairness	5.1.3 Overfitting to the training personal data
		5.1.4 Algorithmic bias
		5.2.4 Inaccurate output due to data drift and deterioration of input personal data quality
	5.4 Principle of security	5.4.1 AI system output disclosure of training personal data
		5.4.2 Personal data storage and personal data breaches
10 Re-evaluation	4 Interpretability and explainability	4.1 Uninterpretable or unexplainable AI system
	5.1 Principle of fairness	5.1.1 Lack of data quality in training personal data
		5.1.2 Bias in training personal data

		5.1.3 Overfitting to the training personal data
		5.1.4 Algorithmic bias
		5.1.5 Interpretation bias
	5.2 Principle of accuracy	5.2.3 Inaccurate personal data output
		5.2.4 Inaccurate output due to data drift and deterioration of input personal data quality
	5.3 Principle of data minimisation	5.3.1 Indiscriminate collection and storage
	5.4 Principle of security	5.4.1 AI system output disclosure of training personal data
		5.4.2 Personal data storage and personal data breaches
11 Retirement	None	None