SUPPORT POOL
OF EXPERTS PROGRAMME

AI Auditing

Checklist for AI Auditing

by Dr. Gemma GALDON CLAVELL

As part of the SPE programme, the EDPB may commission contractors to provide reports and tools on specific topics.

The views expressed in the deliverables are those of their authors and they do not necessarily reflect the official position of the EDPB. The EDPB does not guarantee the accuracy of the information included in the deliverables. Neither the EDPB nor any person acting on the EDPB's behalf may be held responsible for any use that may be made of the information contained in the deliverables.

Some excerpts may be redacted or removed from the deliverables as their publication would undermine the protection of legitimate interests, including, inter alia, the privacy and integrity of an individual regarding the protection of personal data in accordance with Regulation (EU) 2018/1725 and/or the commercial interests of a natural or legal person.

Table of Contents

Document submitted in January 2023

1. Introduction

Algorithmic auditing is a way to inspect AI systems in their specific contexts. It is an approach and methodology that allows for a dynamic appraisal of regulation, standards and impacts. If its results are public, it is also a tool for transparency and accountability.

AI audits are key tools for regulators and society, who can use audit reports to assess how systems work and their impacts. But they are also useful for those developing and acquiring AI systems. An end-to-end, socio-technical approach like the one proposed here generates documentation that improves system accountability, organizational memory and compliance with AI and data regulations. For those acquiring and incorporating AI systems into their operations, audits provide crucial evidence that enable due diligence and proper assessment and comparison of the characteristics between different systems and vendors.

The AI audit checklist proposed is specifically focused on AI impacts. This means that while it gathers information on compliance and trust and safety mechanisms both before and after an AI system is launched, the focus of the audit is to validate that AI developers and implementors have taken all necessary measures, at all different stages, to make sure that the impacts of their systems are in line with existing laws, trust and safety best practices and societal expectations.

It should be noted that an audit process in the framework of the controller implementation of the accountability principle and the inspection/investigation carried out by a Supervisory Authority could be different. Such differences rely, among others, in the final purpose of both activities (the SA could do the inspection to get evidence of an infringement), the scope (limited to the GDPR: applies on personal data processing activities but not on technologies) and the national regulations regarding inspection by control authorities.

This document includes a methodology in the form of a check-list to perform an audit of an AI system. We define an AI system as a logic with a specific outcome. An AI system may be composed of several algorithms, and an AI service or product may include several AI systems.

In addition, it should be noted that there are different techniques for developing artificial intelligence[1]. This document is focused on auditing an algorithm for artificial intelligence based on machine learning, where its life cycle is divided in three totally independent stages from the point of view of data processing and these stages are: algorithm training (pre-processing), the operation of the algorithm implementing one (or more than one) operation in the framework of a personal data processing (in-processing – inference) and the decision making and impact of the same in the processing (post-processing – model deployment). It could be a fourth one, that is the algorithm evolution. All of those stages could be different processing activities and could involve different controllers.

2. Scope of algorithmic auditing

An end-to-end, socio-technical algorithmic audit (E2EST/AA) should inspect a system in the actual implementation, processing activity and running context, looking at the specific data used and the data subjects impacted. It is an end-to-end approach because it recognizes that algorithmic systems work with data produced by complex and imperfect individuals and societies, and operate and intervene in complex social and organizational contexts. Thus, AI systems are deeply socio-technical, and a focus on technical issues would fail to incorporate both problems and possibilities for system

[1] There are different approaches to AI-based solutions: neural networks, rule-based systems, fuzzy logic, machine learning, expert systems, adaptive systems, genetic algorithms, multi-agent systems, etc.

improvement and impact testing that go beyond in-processing. In fact, most of the E2EST/AA focuses on pre-processing and post-processing stages of that algorithmic life-cycle. Models and systems that have optimal performance and accuracy rates in-processing may perform in inefficient or harmful ways when audited end-to-end and using social and technical means.

The E2EST/AA process is designed to inspect algorithmic systems used in ranking, image recognition and natural language processing. It works with systems that make decisions on individuals or groups based on known data sources, regardless of whether they use machine learning or classic computing. This definition includes most systems used by the public and private sectors to make decisions on resource allocation, categorization and identification/verification in sectors such as health, education, security, finance and for applications like fraud detection, hiring, operations management, or prediction/risk assessment.

The E2EST/AA is focused on bias assessment, but not limited to it. The methodology to carry out an E2EST/AA incorporates questions related to broader social impact and desirability, as well as the incorporation of end-users in the design process and the existence of recourse mechanisms for those impacted by algorithmic systems. For a system to pass an algorithmic audit, issues of impact, proportionality, participation and resource must be tackled.

A clear limitation of any audit process is that it is based in an existing system. This means that an audit methodology does not prompt a reflection on whether a system should exist in the first place.

3. Auditing process

An E2EST/AA is an iterative process of interaction between the auditor/s and the development team/s. The method provides templates and instructions to guide such interaction, specifying the data inputs that are necessary for auditors to complete the assessment and validate results.

3.1. Model card

Model cards are documents designed to compile information about the training and testing of AI models, as well as the features and the motivations of a given dataset or algorithmic model. A sample model card like the one proposed below can be used and slightly adapted to different systems or compliance concerns.

General information
○ System name/code and version (5.2 GDPR) ○ Leaflet version and version history (5.2 GDPR) ○ System owner and suppliers data ○ Suppliers' role ○ Risk level (AI Act) ○ Governance roles (Chapter IV GDPR) ○ Distribution date (5.2 GDPR) ○ Existing documentation
Information on process
○ Description of intended purposes, uses, context and role/service provided (Article 5.1.b, 5.2 and 24.1 GDPR) ○ Stakeholder involvement ○ Organizational context ○ Human role/s

Information on training/validation data
○ Data sources/collection methodology (Articles 5 and 9 GDPR) ○ Data types and characteristics (Article 5.1.a, b GDPR) ○ Privacy by Design (Article 25 GDPR) ○ Datasets (Article 5.1.a, b GDPR)
Information on the model
○ Method/s used and justification ○ Simplified output/s ○ Decision variables ○ Objective function/s (Article 5.1.d GDPR)
Information on bias and impacts (in lab/operational settings)
○ Metrics (Articles 5.1.a and 5.1.b GDPR) ○ Protected categories (Articles 13.1.e, 14.1.e and 35.9 GDPR) ○ Impact rates per category and profile (Article 5.1.d GDPR) ○ Auditability (Articles 5 and 22 GDPR)
Information on redress:
○ Explainability profiling (Recital 71 GDPR) ○ Redress or review (Articles 13.2.f, 14.2.g and 15 GDPR) ○ Redress metrics, if applicable

The Model Card allows the auditor to have an initial picture of the system, as well as of the available information, and so it is crucial to determine what issues need to be further explored and inquire system developers on how some of the information provided was determined. It is also a useful way to gather all existing documentation on the system.

Specifically, the inspector should record the existence of:

Documents	**Available**	**Non available**	**N/A**
DPIA/HRIA	Link to document or metadata (author/s, date, etc)	Justification	
Data reuse permissions/authorizations	Link to documents	Justification	
Data sharing agreements	Link to documents	Justification	
Ethics/IRB approval	Link to request and approval documents	Justification	
DPA approval	Link to request and approval documents	Justification	

Transparency report	Link to documents	Justification	
Academic paper/s	Full references	Justification	
GitHub/public repositories	Link	Justification	

3.2.System map

The system map puts the algorithm in context, establishing the relationships and interactions between an algorithmic model, a technical system and a decision-making process. A first version can be designed by the auditor/s following the information provided in the MC, to be completed and validated by the development team/s.

Model: The model is the trained algorithm, that is, the rules adapted to a particular domain, which constitute the foundation of the technology we audit. Models are subject to performance evaluation, test, and can be compared to each other via benchmark datasets. The model is the core of an AI system, but it usually relies upon other elements (e.g. data pipelines, visualization platforms,...) for it to work.

System: The system in this case refers to the entire technology. For a mobility service it could be the app that integrates a Machine Learning (ML) model to predict demand and adjust pricing, including the UI, including for example the data pipelines and protocols.

Process: By process we define the entire lifecycle of any unit of work, from the moment it enters into the workflow all the way to the decision and, if part of the process, the actual way the decision is utilized.

Specifically, in the framework of an investigation carried out by a Supervisory Authority should record the existence of:

<u>IDENTIFICATION AND TRANSPARENCY OF THE AI-BASED COMPONENT</u>

- Inventory of the audited AI-based component [Article 5.2]
  Look for evidence to check, at least, the following questions:
  - *Is the AI-based component identified in the documentation by means of a name or code, identification of version and date of creation?*
  - *Do the code and any additional files defined by the version include a digital signature over the entire package to guarantee its integrity?*
  - *Is a version history of the evolution of the AI component available?*
  - *Does every version recorded include the parameters used in the training of the component and everything that ensures the traceability of the evolution/changes in the component?*
- Identification of responsibilities [Chapter IV]
  Look for evidence to check, at least, the following questions:
  - *Is there an identification about the person(s) or institution(s) who manage the life cycle stages of the AI-based component?*
  - *Is there an identification about the associate managers, and the representatives of the controller and of the processor of every life cycle stage?*

  - *Does every contract associated to each processing stage specify the distribution of responsibilities with regard to personal data protection?*
  - *Has every contract associated to each processing stage been audited?*
  - *Is there a registration in the Records of Processing Activities of the corresponding controllers and processors?*
  - *Is a Data Protection Officer appointed? If not, why?*
  - *Has the Data Protection Officer been identified and communicated his/her identity to the relevant Supervisory Authority?*
- Transparency [Article 5.1.a and Chapter III - Section 1, Articles 13.2.f and 14.2.g of Chapter III - Section 2].
  Look for evidence to check, at least, the following questions:
  - *Are data sources documented?*
  - *Has an information mechanism been implemented?*
  - *Are the characteristics of data used to train the AI component identified, documented and duly justified?*
  - *Is the model chosen for the AI-based component appropriate in terms of simplicity and intelligibility, considering efficiency, quality and accuracy?*
  - *Is the algorithm code explainability documented in order to facilitate its readability, logic comprehension and internal consistency?*
  - *Does the algorithm code documentation include information regarding metadata of the AI-based component, its logic and the consequences that may arise from its use are accessible to data subjects together with the means or mechanisms available to exercise their rights in case of objections to the results?*
  - *Does the algorithm code documentation include information about its behaviour regarding input data sets, data use, intermediate data and output data?*
  - *Can input data sets, data use, intermediate data and output data be traced?*
  - *In case of an erroneous behaviour of the AI-based component that could cause harm to data subjects, have mechanisms been established to minimise such damage?, are communication channels provided to facilitate communication among all stakeholders involved in the process?*

<u>PURPOSE OF THE AI-BASED COMPONENT</u>

- Identification of intended purposes and uses [Article 5.1.b].
  Look for evidence to check, at least, the following questions:
  - *Is the intended purpose of the AI-based component documented both in quantitative and qualitative terms?*
  - *Is there a relation between the use of the AI component with the ultimate purpose of the processing and the conditions guaranteeing the lawfulness of such processing?*
  - *Are the different dynamics, activities and/or processes within the organization in which the life cycle stage of the audited AI component is integrated are identified, delimiting the context of use as much as possible?*
  - *Are potential users of the AI-based component categorized?*
  - *Are other possible uses and secondary users for the AI component? Have been described together with the legal grounds for its use?*
- Definition of the intended context of the AI-based component [Article 24.1]
  Look for evidence to check, at least, the following questions:

- *Are there any legal, social, economic, organizational, technical, scientific or other contexts identified related to the inclusion of the AI-based component in the processing? Are they documented?*
- *Is the organisational and/or contractual structure between the parties defined?*
- *Are the tasks and responsibilities distributed through the structure?*
- *Are the determining factors of the efficacy of the AI component described (including legal guarantees, applicable laws and regulations, organizational and technical resources, available data and internal dynamics that personal data processing needs to implement the audited AI-based component with the appropriate guarantees)?*
- *Are the requirements applicable to human operators in charge of supervising and interpreting the operation of the AI-based component defined?*
- *Is there any interaction between the AI-based component with other own or third-party components, systems or applications? Are responsibilities for maintenance, updating and minimising system privacy issues distributed and documented?*
- *Are levels or thresholds defined for interpreting and using the inference results?*
- *Are defined those contexts for a processing where the AI-based component is not recommended (in terms of its purpose or characteristics, or when it represents an inadequate level of reliability and/or accuracy with regard to the other processing)?*

- Analysis of proportionality and necessity [Article 35.7.b]

  Look for evidence to check, at least, the following questions:
  - *Has the use of the AI component been assessed against other possible options from an approach focusing on the rights and freedoms of data subjects?*
  - *In case of new developments, has a comparative efficiency analysis and adequateness of results of the AI-based component been carried out against other, more thoroughly tested components, which use stricter minimisation criteria or which involve less risks for the rights and freedoms of persons, most especially those that make less intensive use of special data categories?*
  - *In case of addressing a new issue, have the motivations and grounds for addressing this issue by using an AI-based component been documented?*
  - *When addressing a well-known problem, have the grounds for changing the previous operation system that have led to a change in the previous mode of operation been documented including the description of the new control objectives intended by using the AI component in the framework of the procedure?*
  - *Has the risk to the rights and freedoms of data subjects introduced by using an AI-based component in data processing been analysed and managed?*
- Definition of the potential recipients of data [Chapter III; specially Articles 13.1.e and 14.1.e]

  Look for evidence to check, at least, the following questions:
  - *Is the information obligation to data subjects identified regarding data processing arising from the inclusion of the AI-based component?*
  - *Are such obligations identified both for data directly obtained from data subjects and for data obtained from other sources of information?*
  - *When determining such obligations:*
    - *are the recipients or categories of recipients to whom the personal data processed by the AI-based component were or are to be communicated identified (including those who are in third countries or are international organizations)?*

        - *are the intentions of the controller of transferring personal data to a recipient in a third country or international organization and the existence or absence of a Commission decision on adequacy identified?*
    - *Are data recipients –including those from third countries or international organizations– identified under the activity or activities recorded in the Records of Processing Activities in which the relevant AI-based component is included?*
- Limitation of data storage [Article 5.1.e, exceptions Article 89.1]
  Look for evidence to check, at least, the following questions:
    - *Are the legal grounds to store personal data used by the AI-based component for a period of time that exceeds the period established for processing purposes identified (especially when it is related with compatible purposes or included in any of the exceptions provided in the regulations)?*
    - *Is it justified to store personal data once it is processed in any life cycle stages of the AI-based component?*
    - *Have appropriate technical and organizational measures and criteria been defined to storage personal data?*
    - *Are the time limits for erasure of stored personal data defined?*
    - *Has a conservation policy been defined to keep a sample of training data for the purpose of auditing the AI component? Does it consider the minimum or assumable risks for the data subjects?*
    - *Are there procedures to verify storage periods, criteria and implemented measures?*
    - *For those cases where an excessive pattern of data storage has been detected, either in terms of time or quantity, has a procedure for reviewing the analysis of the need and the proportionality of data storage been defined?*
    - *Has a storage policy for personal data included in the activity records of the AI-based component and privacy strategies (minimisation, hiding, separation or abstraction) been defined? Has it been implemented for operation purposes?*
- Analysis of categories of data subjects [Article 35.9]
  Look for evidence to check, at least, the following questions:
    - *Are the categories of data subjects affected by the development of the AI component and its use in the framework of the intended processing identified?*
    - *Are the short- and long-term consequences that the implementation of the AI component may have on the categories of data subjects identified?*
    - *Is there any procedure that analyses the social context in which the AI component is used, collecting information from people, groups or organizations affected by such AI component for the purposes of knowing their levels of satisfaction, position, concerns and uncertainties regarding the application of this technique for processing their data?*

<u>BASES OF THE AI COMPONENT</u>

- Identification of the AI-based component development policy [Article 24.1]
  Look for evidence to check, at least, the following questions:
    - *Do the documents with development policies of products and systems consider the data protection policy?*
    - *Are the policies reviewed and version controlled?*
- Involvement of the Data Protection Officer (DPO) [Section 4 of Chapter IV]
  Look for evidence to check, at least, the following questions:

  - *Does the DPO have the necessary professional qualifications and, particularly, the legal and technical expertise, as well as data protection practice appropriate to the project?*
  - *Is the DPO assisted and advised by experts on specific matters relating to the AI component?*
  - *Are there internal procedures defined within the organisation for correct communication between the DPOs and the people in charge of those projects that may have an impact in data processing, in order to obtain assistance, particularly when developing the data protection impact assessment for those processing activities which include AI-based components?*
  - *Has the DPO played an active role in the stages being audited? Has his or her independence of judgment within the organisation and his or her obligations to cooperate with the supervisory agencies been respected and his or her opinions, remarks and considerations taken into account?*
- Adjustment of basic theoretical models [Article 5.1.a]

  Look for evidence to check, at least, the following questions:
  - *Has an analysis been carried out regarding the theoretical framework and previous similar experience on which the development of the AI component is based?*
  - *Have the basic hypotheses and premises considered in order to create and develop the relevant model been accurately described, justified and documented?*
  - *Is a critical and verified procedural revision defined for the reasoning arising from acceptance of important hypotheses for the development of the AI-based component (i.e. examining which are the arguments for a causal relationship that models an algorithm, such as the selection of variables defining a certain phenomenon)?*
  - *Have appropriate premises been established regarding the potential proxy variables intervening in the AI-based components after carrying out a careful analysis?*
- Appropriateness of the methodological framework [Article 5.1.a]

  Look for evidence to check, at least, the following questions:
  - *Is there proper documentation that include the methodological framework for defining the model and creating the AI component in the audited stages, such as the methods for selecting, collecting and preparing component's training data, labelling, model building, using intermediate data, selecting the test/validation data subset or measuring deviations for improvement purposes?*
  - *Is the development model to be used properly determined depending on the results of the analysis of the problem to be solved and in a justified way (i.e. supervised, unsupervised or others)? In case of supervised models, does it specify the procedure for supervising the learning process of the algorithm, the degree of supervision and the basis for such supervision?*
  - *Are the metrics for measuring the behaviour of the AI component duly selected and measured?*
  - *Has a procedure been implemented for recording and monitoring the deviations in the behaviour of the AI component with respect to the defined metrics that allows to identify the circumstances which may arise in an erroneous or biased behaviour?*
- Identification of the basic architecture of the AI-based component [Article 5.2]

  Look for evidence to check, at least, the following questions:
  - *Does the project analysis phase of the AI-based component include, as part of the requirement catalogue, a series of specific requirements too guarantee privacy and personal data protection?*

- *Is there documentation which assure that, when programming AI-based components, the coding principles, codes and coding, best practices applied are followed in order to guarantee that the code is readable, secure, low-maintenance and robust?*
- *Is the basic architecture of the AI component identified and documented? It must include information on the chosen machine learning technique, the type(s) of tested and, when appropriate, dismissed algorithms at the learning and training stages, and other data on the functioning of the relevant component, such as the model loss function or cost function.*
- *Does a systematic procedure for documenting the component implementation procedure exist? Is it implemented? It is necessary to guarantee registration and subsequent acquisition of all necessary information to identify such component, its elements and its environment, understanding what it does and why it does it, and enables to verify code quality and legibility for auditing purposes: description of the programming language(s) used, most recent code version, commented-out code, necessary packages and libraries, and interfaces with other components, when appropriate, used APIs and useful documents such as requirements specifications, functional and organic analyses, guidelines, etc.*
- *Is the AI component code impossible to access? If yes, is a reverse-engineering process or other alternative method used (i.e., a zero-knowledge proof (ZKP))? A reverse-engineering process enables to know more about the component function and to establish the logic of rules applied in order to detect inconsistencies, direct manipulations and underestimation or overestimation of the variables used in the original component.*

3.3. Moments and sources of bias

Bias refers to a deviation from the standard. As such, and in technical terms, bias may be needed and desirable. In the context of AI accountability, however, "bias" has become an hypernym or umbrella term for lack of fairness and discrimination in data processes which result in individual and/or collective harms. By identifying and mitigating bias, we can ensure or protect fairness in AI systems. Bias is the result of many factors, social and technical: from systematic errors introduced by algorithmic design choices, dirty data, sampling procedures, reporting protocols, or wrong assumptions that cause a mismatch between the input features and the target outputs. To date, *most studies on bias have focused on historical and aggregation bias*, that is, the need to identify protected groups and calculate disparate treatment and impact. This is at the heart of the E2EST/AA methodology. However, bias and inefficiencies can emerge at other times, and a focus on historical and aggregation bias alone will lead to incomplete and therefore harmful assessments of bias. This will result in rights violations, stereotyping, bad or inefficient decisions, discrimination of individuals and groups, and the reproduction of processes of inequality and dispossession. Partial or wrongful identification of bias sources and inadequate mitigation measures will lead to unacceptable societal harms and compliance risks.

The E2EST/AA distinguishing between moments and sources of bias. This provides the auditor with an overview of the possible causes of a given *disparate impact*, understood not only as an individual function of accuracy or performance but also as a general measure of (lack of) fairness in an *algorithmic process*. The E2EST/AA method defines and identifies moments and sources of bias, establishes the documents and tests needed to assess compliance with legal and social requirements, provides an opportunity to address and mitigate inefficiencies and harms, and provides a measure for overall system fairness and impact.

AI life-cycle	Pre-processing	In-processing (model inference)	Post-processing (model deployment)
Moments of bias	World → Data Data → Population Population → Sample Sample → Variables + Values	Variables + Values → Patterns Patterns → Predictions	Predictions → Decisions Decisions → World
Sources of bias	Techno-solutionist bias Selection bias Historical bias Label bias Generalization bias Statistical bias Oversimplification, partial or biased featurization Omitted variable	Over and underfitting Measurement bias Hot hand fallacy Privacy bias Aggregation bias	Benchmark test bias Data visualization Automation bias Deployment bias

Specifically, in the framework of an investigation carried out by a Supervisory Authority should record the existence of:

DATA MANAGEMENT

- Data quality assurance [Article 5.1]
  Look for evidence to check, at least, the following questions:
  - *Is there a documented procedure to manage and ensure proper data governance, which allows to verify and provide guarantees of the accuracy, integrity, accuracy, veracity, update and adequacy of the datasets used for training, testing and operation?*
  - *Are there supervisory mechanisms for data collection, processing, storage and use processes?*
  - *Has a previous analysis been carried out together with a measurement of the sample used for training the relevant model? Has the sample size been verified as adequate? Has the frequency and distribution of each feature been verified, their intersection or the relevant groups for the study are appropriate regarding defined parameters or to reality?*
  - *Has the learning process been analysed, both at the beginning and in each iteration of the global learning process, and on the sample used to train the model? Has it been verified that the final dataset is representative with respect to the population of the context to which the AI-based component is oriented and that the groups defined by said AI component are appropriate?*
  - *Is the feature distribution appropriate and make de IA component not especially sensitive or ignores any of them?*

- *Are there procedures to analyse, measure and detect any possible imbalances between the amount of data that the component collects on a certain feature with respect to another and which may lead to behaviour deviations?*
- *Has an accurate compensation analysis been carried out in order to establish the relationship between the amount and type of data to be collected/discarded and those who are necessary to guarantee the that the AI component is effective and efficient?*
- *Has a sample size analysis has been carried out regarding data storage for audit purposes?*

- Definition of the origin of the data sources [Articles 5 and 9]

  Look for evidence to check, at least, the following questions:
  - *Has the origin and the data sources context used for training and validating the model been identified?*
  - *Is there documentation that justify the selection process of data sources used to train the relevant AI-based component?*
  - *Are legal grounds to used personal data in the different stages of the AI-based component life cycle identified?*
  - *Is there a justification to collect and use personal data when such data are not necessary in the training stage, in order to test the model behaviour in the subsequent stages of component verification and validation?*
  - *If sensitive personal data are processed, has the need for their use been assessed and certain circumstances justify to lift the general prohibition to process such data?*
- Preprocessing of personal data [Article 5]

  Look for evidence to check, at least, the following questions:
  - *Is the criteria to carry out previous cleansing of original data sets and any other tasks needed throughout the different iterations of the AI-based training process duly identified and documented?*
  - *Are data cleaning techniques and best practices used in the data cleansing process properly selected and documented?*
  - *Do classifying features define clearly distinguishable and identifiable types?*
  - *Is the structure and properties of the processed data set documented, including the number of data subjects and the extension of used data?*
  - *Have data been previously classified into categories, organizing them in non personal and personal data, and, for the latter, identifying which fields constitute identifiers, quasi-identifiers and special data categories?*
  - *Have the relevant model features for the model been determined (identifying the those associated with special data categories and proxy variables, including the necessary information for their interpretation)?*
  - *Has data minimisation criteria been 14determined and applied to the different stages of the AI component, using strategies such as data hiding, separation, abstraction, anonymisation and pseudonymisation that might apply for the purposes of maximising privacy in the operation of the relevant AI-based component?*
  - *Do databases have an associated data-dictionary for the analysis and understanding?*
  - *Have segregation and de-identification strategies been implemented on additional information that is not required for training purposes but shall be required in the verification and validation processes of the model's behaviour? It is needed to analyse correlations between variables, measure the degree of accuracy of the AI component with regard to certain attributes and ensure that no biases are introduced.*

  - *Have data selection and assessment been carried out with the involvement of an expert in modelling techniques and data science?*
  - *Have training and validating data been previously pre-processed and cleaned in order to detect any possible abnormality which may require previous processing (i.e., boundary values, incomplete records, etc.) and to convert any heterogeneous data sources to a homogeneous format?*
  - *In case of input data are not appropriate with regard to the functioning of the AI component or because it is not representative of the reality it intends to reflect, have the necessary modifications been introduced in the format of these data?*
  - *When necessary, has a data anonymisation analysis, including the possible risk of re-identification, been carried out?*
  - *When necessary, if data imputation techniques have been used to complete the information of the data set, have the procedures and algorithms used for such imputation been documented?*
- Bias control [Article 5.1.d]

  Look for evidence to check, at least, the following questions:
  - *Have appropriate procedures been defined in order to identify and remove, or at least limit, any bias in the data used to train the relevant model?*
  - *Has it been verified that in training data did not have previous biases?*
  - *Is there a procedure to assess the need to have additional data for improving precision or removing any possible bias?*
  - *Are there human supervision mechanisms implemented in order to control and ensure that results are bias-free?*
  - *Are mechanisms implemented to enable data subjects to request human intervention, provide feedback or refute the results obtained by means of automated decision-making algorithms?*

VERIFICATION AND VALIDATION

- Adapting the verification and validation process of the AI based component [Articles 5.1.b and 5.2]

  Look for evidence to check, at least, the following questions:
  - *Is there documentation that duly describe the verification and validation process, the techniques used, the verification and test assembly carried out, the results obtained, and the proposed actions?*
  - *Have there been established or followed guidelines, standards or regulations in order to carry out a systematic procedure to verify and validate the AI-based component and its behaviour once integrated in the processing activities it supports?*
  - *Are control and supervision mechanisms in place to ensure that the AI-based component efficiently complies with its intended goals and purposes?*
  - *Are metrics and criteria, on which verifications within the verification and validation process shall be carried out, defined and justified?*
  - *Is a testing strategy defined? Related to this strategy, is there a testing plan to assess the correction of the AI component both from structural and functional terms?*
  - *Are the personnel involved in AI-component verification and validation tasks qualified to carry out the necessary checks in order to ensure that the component has been correctly built and behaves as expected?*
- Verification and validation of the AI-based component [Articles 5.1.a and 5.1.b]

  Look for evidence to check, at least, the following questions:

- *Does the testing plan include reviews and, when appropriate, inspections for the purposes of early identification and remedy of defects in requirements or design, incorrect specifications or deviations from applicable criteria during development?*
- *Is white-box testing of the network design or the AI component considered as part of the testing plan?*
- *Is white-box testing at code and implementation levels included in the testing plan?*
- *Is black-box testing considered as part of the testing plan in order to ensure that functionality of AI-based component is guaranteed, it behaves as expected and the information integrity is preserved?*
- *Are security check tests provided as part of the test plan in relation to the protection of rights and freedoms, in its holistic definition (physical and IT) in the case of AI components implemented in robotic systems, industry 4.0, or the Internet of Things?*
- *Does the validation test plan include verification of boundary values and extreme test cases which might make the component functioning in an unexpected manner?*
- *Is there a cleaning procedure to correct any errors, shortcomings or inconsistencies detected during the verification and validation process?*

- Performance [Article 5.1.d]

  Look for evidence to check, at least, the following questions:
  - *Are metrics or sets of aggregated metrics used to determine the precision, accuracy, sensitivity or other performance parameters of the relevant component in consideration of the principle of data accuracy established?*
  - *Are the rate values of false positives and false negatives yielded by the AI component known and analysed and interpreted in order to establish their accuracy, specificity and sensitivity of the component behaviour?*
  - *Has the level and definition of the performance parameters required for the AI-based component in the framework of the processing been assessed?*
  - *Have the performance values between different options of AI components been compared in the context of a process of selection of the most appropriate component for a specific processing?*
  - *Are output variables defined and determined with special consideration to those that constitute special data categories?*
  - *Have measures, which ensure that data used are exhaustive and updated, adopted?*
  - *Have relevant parameters and their cut-off values been determined (so the model considers certain variables in order to obtain significant results)?*
  - *Are there procedures to detect whether the response of the AI-based component to input data is erroneous or exceeds a predetermined error threshold, or whether there are different error thresholds associated with different categories of data subjects in the data set?*
  - *Has a dimension reduction been carried out in order to achieve a balance between complexity and generalization?*
- Consistency [Article 5.1.d]

  Look for evidence to check, at least, the following questions:
  - *Is there a procedure to verify whether the obtained results present significant changes with respect to the results expected, and to act accordingly?*
  - *Has a threshold been established to determine when an obtained result deviates from the expected result based on identical or similar data inputs (significant deviations)?*

- *Has it been analysed whether the AI-based component behaves differently when processing data from individuals who differ in their personal characteristic associated to special data categories or in the values of the proxy variables?*
- *Has the effect of changes in low prevalence variables within the training dataset in output results of the AI-based component been assessed?*
- *Are there measures adopted to ensure component independence?*
- *Is it verified that there is no correlation between the results and the additional variables associated to data subjects that are not a part of the process variables and which may evidence the existence of biases?*

- Stability and robustness [Article 5.2]

  Look for evidence to check, at least, the following questions:
  - *Within the possible or actual context of function of the relevant component, are the factors, whose variation may impact the properties of the AI component and may establish the need to manage its readjustment, identified?*
  - *Has the AI component behaviour in unexpected environments been assessed?*
  - *Has a time estimation for reassessment, readjustment or reboot of the component in order to have it adjusted to input data deviation or changes in decision-making criteria is required been analysed?*
  - *Is there any documentation that show whether the AI component has been built with a static approach, a dynamic approach or a continuous learning approach by design?*
  - *In case of continuous learning AI component, has the degree of adaptability to new input data or types of data been assessed? have monitoring procedures and mechanisms been defined in order to verify that conclusions obtained remain valid, that the AI component is capable of acquiring new knowledge and/or previous associations learned have not been lost?*
- Traceability [Articles 5 and 22]

  Look for evidence to check, at least, the following questions:
  - *Is there a version control system in place for all elements of the AI-based component: used datasets, AI-based component code, libraries used and any other element associated with the component?*
  - *Is there a formal and documented procedure, subject to reassessment as appropriate, of risk assessment depending on such changes that may occur on the implementation of the AI-based component throughout its life cycle?*
  - *Have monitoring and supervision mechanisms (such as log files and results records) been implemented to properly assess the behaviour of the component in interaction with environment, to measure that the relevant outputs are adjusted to the responses of real-life processes that they model and to any potential inconsistency between expected behaviour and the automated one?*
  - *Is there a record of incidences and previous abnormal behaviours detected and remedied?*
  - *Are there monitoring mechanisms available for human operators for monitoring and verification purposes?*
  - *Has a procedure been implemented and documented to ensure human intervention in decision-making, either on its own initiative, when results deviate from expected behaviour, or on request of data subjects affected by the AI-based component's output?*
  - *Are mechanisms adopted within the framework of the processing so that the results and decisions taken may be entirely the responsibility of human operators?*

- Security [Articles 5.1.f, 25 and 32]
  Look for evidence to check, at least, the following questions:
  - *Has a risk analysis developed with regard to the risks for rights and freedoms of persons? Have the results of this risk analysis been used to determine the security and privacy requirements of the AI-based component in the framework of the processing?*
  - *Are data protection and security requirements related defined at the origin and together with any other requirements, regardless of whether they are to be applied to the design of a new AI-based component or to the modification of an existing one?*
  - *Have standards and best practices been taken in consideration for secure configuration and development of the AI relevant component?*
  - *Are measures to ensure protection of the processed data implemented? particularly those oriented to guarantee confidentiality by means of data anonymisation or pseudonymisation, and integrity to protect component implementation from accidental or intentional manipulation.*
  - *Are measures to guarantee component resilience and its capacity to withstand an attack implemented?*
  - *Have procedures implemented in order to properly monitor the functioning of the component and early detect any potential data leak, unauthorised access or other security breaches?*
  - *Do component users and operators have sufficient information and are able to be aware of their security duties and responsibilities regarding data protection and safeguarding data subjects' rights and freedoms?*

3.4. Bias testing

Based on the documentation provided and access to team developers and the data available, different types of tests can be designed to determine whether different types of bias are impacting systems in ways that may cause harm to individuals, groups, society or the efficient functioning of an AI system. In all cases, bias testing involves a documentation and literature review, interviews with developers/implementors and a good understanding on who is impacted by AI systems and how. Bias testing involves statistical analysis and checking, and auditors have a choice of fairness definitions and metrics to choose from. Statistical notions of fairness such as those described by Verma & Rubin (2018) are a good starting point and can be the basis for more advanced approaches such as similarity-based measures and causal reasoning. In some cases, bias testing requires reaching out to end users or those impacted by systems.

As it may not be feasible for an inspector to go through all moments and sources of bias, the main step of the inspection exercise must include:

a) Definition of protected groups: in the context of artificial intelligence, a protected group is a group of people who are historically disadvantaged or marginalized, and who may be at risk of discrimination or negative impacts from the development and deployment of AI. Protected groups may be defined by characteristics such as race, ethnicity, gender, sexual orientation, religion, age, ability, and socio-economic status.
b) Testing the output of the AI system: one way to measure bias is to test the output of the AI system and compare it to a benchmark or ground truth. For example, if an AI system is intended to classify objects in images, the inspector could test its performance on a dataset that includes a diverse range of objects and see how accurately it classifies them.
c) Examining the training data: another way to measure bias is to examine the training data that was used to develop the AI system. If the training data is not representative of the population

that the AI system will be used on, or if it contains biased examples, then the AI system may also be biased.

	decision		
group	-	+	
protected	a	b	n_1
unprotected	c	d	n_2
	m_1	m_2	n

d) Using fairness metrics: fairness metrics are used to determine whether a protected group has sufficient presence, receives consistent treatment and is properly represented in the system. A good place to start is calculating the Risk Difference, where RD is p1-p2 and the Risk Ratio (where p1/p2). The inspector can also measure demographic parity, equal opportunity, equalized odds, and seek to measure both direct and indirect bias through different means. The best means to utilize will be determined by the system's transparence, complexity and the inspection point (pre-processing, in-processing or post-processing). Any inconsistency detected will point to issues that need to be discussed with the development team and further explored to ensure that all necessary precautions and measures to ensure a fair functioning of the system have been taken and documented.

3.5. Adversarial audit (optional)

The most thorough auditing methodology can still miss things. Omitted variables or proxies that only become visible once an algorithmic system is functioning in real-life production settings will result in unfair treatment and harmful impacts. For unsupervised ML models, reverse-engineering may be the only way to trace back model attributes. For high-risk and unsupervised ML systems, performing an adversarial audit once a system is implemented is highly recommended. Adversarial audits are also useful to verify that the information provided during the auditing process is complete and accurate.

Adversarial auditing can reveal the existence of the moments of bias listed above, but also additional sources of bias such as learning bias, which occurs when an unsupervised ML system incorporates new variables and labels that emerge from the training data without human intervention or control, leading to potential harms that are only identified when the auditor can access impact data at scale.

To conduct an adversarial audit, the auditor needs to gather impact data at scale. This can be done through *scrapping* web sources (in the case of web-based systems), by *interviewing* end users, by *crowdsourcing* end-user data or by *sockpuppeting* a system (creating fake profiles or input data with specific characteristics to trigger mode outcomes and analyze them).

Adversarial audits can complement a E2EST/AA or be conducted as a stand-alone when impacted communities or regulators do not have access to an algorithmic system.

4. The audit report

Audits should always result in a public document. However, this is not the only report that will be produced during the audit process. A crucial part of auditing is documentation, and so all interactions and documents exchanged must be compiled and either kept on file by system owners (and, if both parties agree, by auditors). There are three main audit reports:

a) Internal E2EST/AA report with mitigation measures and annexes

This document captures the process followed, the issues identified and the mitigation measures that have been applied or can be applied. Contrary to financial auditors, algorithmic auditors do engage in proposing solutions, monitoring their implementation and reporting on the final results. The internal audit report need not be published.

b) Public E2EST/AA report

Final version of the audit process, where auditors describe the system, the auditing methodology, the mitigation and improvement measures implemented and further recommendations, if any. The public audit report must also include a proposal for the periodicity and methodology/metrics to be used in follow-up audits.

c) Periodic E2EST/AA reports

Follow-up audit reports. These must always refer and provide access to the initial audit report, if it is still relevant, and provide guarantees that the system developers have continued to test for bias, implement mitigation measures and control for impact. Depending on the complexity of the system/s, both parties may agree to produce an internal and a public version of each periodic audit.


edpb
European Data Protection Board