agencia española protección datos

Audit Requirements for Personal Data Processing Activities involving AI

v. January 2021

This document has been developed based on a report from Eticas Research and Consulting SL, carried out under the mandate and supervision of the Spanish Data Protection Agency, and reviewed by experts from the Artificial Intelligence Hub of the Spanish National Research Council (CSIC,), from the Artificial Intelligence Social and Ethical Impact Observatory (OdiseIA), from the Professional Association of Public Administration Information Technologies and Systems Engineers (ASTIC), , from the Teaching Innovation Group in Cybersecurity (CiberGID) -ETSI Informática of the National Distance Learning University (UNED), and from the Centre for Industrial and Technological Development (CDTI) .

EXECUTIVE SUMMARY

This document intends to be a first approach to a series of controls liable to be included in audits of processing of personal data which use components based on artificial intelligence (AI). It is important to state that included controls are designed to carry out an analysis on whether a given processing is GDPR. Besides, certain methodological notes are included which may be appropriate and characteristic to this kind of audits.

This document is published in line with the GDPR compliance of processing activities that embed Artificial Intelligence. An introduction published by the AEPD with regard to effective compliance of personal data protection in data processing activities including artificial intelligence solutions.

Auditing personal data processing activities is one of the existing tools to assess compliance. For those processing activities subject to the relevant provisions set forth in the GDPR and including AI-based components, the relevant audit must include specific controls derived from the peculiarities of such components. The control list included in this document intends to be a reference for future auditors to determine, after a previous analysis, which are appropriate to be included when auditing a particular processing procedure.

Although it is true that auditing an specific processing, and even more when it includes AI-based components, could include verification of other aspects of such processing, such as an ethical assessment or technical efficiency, this document shall exclusively focus on those aspects related to personal data protection.

Besides, auditing a processing activity which includes, among others, IA-based components, must not and cannot focus exclusively on specific technical aspects of the technologies used, but must have a much wider scope that includes the nature, scope, context and purposes of processing and the risks for rights and freedoms of persons that it may entail.

This document is mainly addressed to controllers who are in charge of auditing processing activities including AI-based components, as well as processors or developers who wish to provide guarantees with regard to their products and solutions, Data Protection Officers in charge of supervising specific processing activities and acting as advisers to controllers, and audit teams in charge of assessing such activities.

Keywords: Artificial Intelligence, auditing, AI component, algorithms, Machine Learning, automated decisions, profiling, massive data, Big Data, GDPR, Personal Data Protection, accountability, transparency, compliance, bias, explainability

TABLE OF CONTENTS

I. INTRODUCTION

Article 24 of the GDPR sets forth the obligation to *"implement appropriate technical and organisational measures to ensure and to be able to demonstrate that processing is performed in accordance with this Regulation".* Such measures must be chosen *"taking into account the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for the rights and freedoms of natural persons*". When necessary, these measures will also be subject to a process of constant review and updating.

One of the tools used "*to ensure and to be able to demonstrate*" compliance with the provisions of the GDPR is auditing processing activities. In compliance with his or her tasks (article 39), the data protection officer would monitor these audits. All processing activities must be assessed in relation to their purposes, as in the context and scope in which they are to be deployed. Such assessment must also consider the nature of the relevant processing and the need to carry out a data protection impact assessment (article 35) for different reasons, including the processing on a large scale of special categories of data, the use of new technologies, or the systematic and extensive evaluation of personal aspects relating to natural persons based on automated processing, including profiling.

With regard to this aspect, the specificities that may be derived from including in the relevant processing operations implemented by means of elements based on specific technological solutions[1] (article 4.2) must be considered. Such components are specific implementations of data processing techniques which, for reasons of proportionality, collateral effects or other, may include characteristic elements both affecting compliance with GRPD and the risks arising for rights and freedoms of data subjects. Moreover, it is necessary to remember the need to inform data subject about the existence of automated decision-making, including profiling (article 13.2.f)[2], in addition to providing meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

The potential impact of AI component-based processing in the rights and freedoms of citizens proves the need to establish measures of effective control, correction, responsibility, accountability, risk management and transparency with regard to systems and to processing activities in which such AI components are used. Currently, models for auditing processing activities including AI-based components are under development. Holistic models that allow both the practical implementation of the principle of accountability throughout the data life cycle and that the clarification of responsibilities in the different phases of personal data collection and processing are also being developed. In this sense, this document represents only a first approach towards determining the basic elements that future audits could incorporate from the point of view of data protection and towards the development of future standards[3].

[1] According to article 4.2 of GDPR, the following are considered operations: "collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction."

[2] Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. 2017. "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation." International Data Privacy Law 7 (2):76-99.

[3] See, for example, the standardization efforts being made by the ISO / IEC JTC 1 / SC 42 committee, made up of the International Electrotechnical Commission (IEC) and the International Association for Standardization (ISO) in the area of artificial intelligence. It is worth highlighting the recommendations of this committee to the EU AI strategy: "CEN-CENELEC response to the EC White Paper on AI, version 2020-06". Available at: https://www.cencenelec.eu/News/Policy_Opinions/PolicyOpinions/CEN-CLC%20Response%20to%20EC%20White%20Paper%20on%20AI.pdf

The document "GDPR compliance of processing activities that embed Artificial Intelligence. An introduction", as published by the AEPD, devoted a whole chapter to auditing of processing activities, including AI-based components or phases implemented by means of AI components, from an approach considering personal data protection regulations. Audits were include as one of the potential assessing tools in the framework of IA-comprising processing and an instrument intended to obtain products which are secure, , predictable, controllable and whose internal logic can be explained in a certain way.

In this sense it must be considered that artificial intelligence solutions, especially those classified as "weak AI " (See Annex I "Definitions"), are built and executed using traditional hardware and software components. Weak AI is a new model for application development rather than a new computing model. Maturity of such development procedure shall be critical to ensure traceability, explainability and quality of the final built product.

Traditional development models are supported on good practices, widely known and implemented, such as Software Development Life Cycle Management (SDLCM), Capacity Maturity Model (CMM) or Application Life Cycle Management (ALM). All such models establish guidelines and recommendations for systematic development of products in general and, in particular software applications. However, and despite their shared aspects, these models need to be adapted to specific cases. This is especially so for components whose life cycles are different from traditional systems development models[4]. This is the case, for example, of machine learning, due to its inherent characteristics[5]. In addition, the pace of production means that data sources, software and hardware from both in-house and third parties are integrated and that these data projects can use statistical techniques, machine learning or more advanced artificial intelligence activities. The continuous integration cycle of these three elements is what makes AI product development unique.

Despite these peculiarities, the basic principles, of analysis, design, development, verification and validation also apply to development of AI-based components, albeit they need to be adjusted to the specific characteristics of these technologies. For this reason, it is still of paramount importance to adopt a systematic model of the life cycle of the development of AI-based components, so that its building procedure may be audited from a quality assurance approach.

Assessment of the processing must be performed on the processing as a whole. However, the sheer complexity of certain technological solutions, such as those solutions based on AI, makes it advisable to follow specific guidelines to manage the singular elements [6] [7] [8] [9] [10] incorporated by such technologies. Such guidelines or recommendations

[4] While traditional software applications are deterministic and are programmed to behave according to a specific set of requirements and specifications, applications based on automatic learning are probabilistic, learn from mostly unstructured data, and need to be trained for a variable number of iterations throughout different stages.

[5] Rama Akkiraju, Vibha Sinha, Anbang Xu, Jalal Mahmud, Pritam Gundecha, Zhe Liu, Xiaotong Liu, John Schumacher . Characterizing machine learning process: A maturity framework. IBM Almaden Research Center, San José, California, USA, Nov. 2018. Available at: https://arxiv.org/ftp/arxiv/papers/1811/1811.04871.pdf

[6] Guidelines on Cookie Use. AEPD, 2020. Available at: https://www.aepd.es/sites/default/files/2020-07/guia-cookies.pdf (only in Spanish)

[7] Analysis of information flows in Android. Tools for compliance with accountability. AEPD, 2019. Available at: https://www.aepd.es/sites/default/files/2019-12/estudio-flujos-informacion-android-en.pdf

[8] Guidelines for cloud computing service providers. AEPD, 2018 .Available at: https://www.aepd.es/sites/default/files/2019-09/guia-cloud-prestadores.pdf (only in Spanish)

[9] Code of Good Practices for Projects Involving Big Data. AEPD, 2017. Available at: https://www.aepd.es/media/guias/guia-codigo-de-buenas-practicas-proyectos-de-big-data.pdf (only in Spanish)

[10] Guidelines and guarantees in personal data anonymisation procedures. AEPD, 2016. Available at: https://www.aepd.es/sites/default/files/2019-09/guia-orientaciones-procedimientos-anonimizacion.pdf (only in Spanish)

must be integrated in the set of general controls intended to assess a specific processing, as a specific and necessary subset in the global analysis of such processing. That is, such recommendations are of a general and transversal nature for all processing activities including technological components of this type, but they are not exhaustive for the purposes of assessing a specific processing in its entire scope and dimensions.

The European Commission's *White Paper On Artificial Intelligence - A European approach to excellence and trust*"[11], published in February 2020, states that *"Testing centres should enable the independent audit and assessment of AI-systems* in *accordance with the requirements outlined above. Independent assessment will increase trust and ensures objectivity. It could also facilitate the work of relevant competent authorities*." The High-Level Expert Group on Artificial Intelligence set up by the European Commission provides guidance on how Trustworthy AI can be realised, by listing seven requirements that AI systems should meet: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) environmental and societal well-being and (7) accountability[12].

As for this document, it would address a list of control objectives and audit controls which may be included as part of the audit controls of a specific processing implementing at least one AI-based component. The list is extensive, so that not all control objectives and controls listed in this document need to be applied to all processing activities including an AI-based component. Different processing activities shall require different control objectives and controls. The proposed list of controls intends to serve as a reference for auditors, who may then choose those controls which apply to the specific processing being audited. Selection shall be carried out according to different factors: whether they impact on GDPR compliance, the type of AI-based component used, the type of processing (for example, if the development of a component is audited or if the processing that includes a working component is audited) and, most particularly, the risk it entails to rights and freedoms.

On the other hand, the controls presented here do not address other aspects of the processing which are not directly linked to the use of an AI component and arise from a general analysis of the processing.

In this sense, this document cannot be used as a guide to perform a data protection audit on an entire processing procedure supported by an AI-based solution, and does not intend to meet other objectives than those established by the GDPR, such as ethical or efficiency goals. The purpose of this document is not describing the general aspects of auditing methods, which have been widely described in the field's literature, not to recommend tools, of a general or specific nature, which may be implemented in audits. The scope if this document is mainly to provide specific methodological guidelines and a list of control objectives and specific controls which may be selected to be included in the data protection audit of a processing including AI-based components or solutions.

In this context, this document is mainly addressed to controllers who are in charge of auditing processing activities including AI-based components, with the aim of helping them to ensure and be able to demonstrate compliance with the data protection obligations and principles to which they are subject, as well as processors or developers who wish to provide

White Paper On Artificial Intelligence - A European approach to excellence and trust [online] COM(2020) 65 end, 30 [Last consulted: 29 October Available at: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_es.pdf

[12] AI-HLEG. 2019. Policy and investment recommendations for trustworthy Artificial Intelligence. High-Level Expert Group on Artificial Intelligence. Available at: https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence

guarantees with regard to their products and services, to Data Protection Officers in charge of supervising specific processing activities and acting as advisers to controllers, and to audit teams in charge of assessing such activities.

II. AUDITING METHODS FOR PROCESSING ACTIVITIES INCORPORATING AI-BASED COMPONENTS

From the point of view of data protection, it is important to highlight that the approach to AI will always come from the perspective of the use of emerging technologies whose risks to the rights and freedoms of data subjects will have to be assessed. This evaluation will naturally lead to the implementation of technical and organisational measures that will make it possible to minimise these risks and maximise the expected benefits of the processing of personal data. Based on this approach, the value of the methodologies already in use and of the existing standards and certifications is evident.

Regardless that the different existing auditing methods may differ from each other depending on the approach or perspective towards which they are focused, multiple applicable references can be obtained. For example, the general principles in ISO 19011[13], the most specific aspects of software development in standards such as ISO/IEC 15504 SPICE[14], methodologies such as METRICAv3[15], standards such as CMMI[16] or other reference frameworks such as COBIT[17], SOGP[18].The methodological aspects of an audit will not be detailed in this document as they are extensively developed in the previous references. However, there are differentiating characteristics that need to be considered when planning and executing an algorithmic audit of artificial intelligence. There are also numerous standardisation efforts[19] [20] in the field of artificial intelligence that will have to be taken into account when developing specific audit processes that meet the needs of processing activities involving an AI component that is being assessed.

A. GENERAL CONTROL OBJECTIVES OF AUDITING AN AI-BASED COMPONENT WITH REGARD TO DATA PROTECTION

Auditing an AI-based component in the context of data protection needs to be a systematic, independent and well documented process focused on obtaining and objectively assessing evidence, for the purposes of establishing the level of compliance with the relevant chosen auditing criteria which, in this case, are related to processing principles and other requirements established by the GDPR and other applicable data protection regulations.

---

[13] Guidelines for auditing management systems. ISO 19011.2018. Available at: https://www.iso.org/standard/70017.html y https://www.en.une.org///encuentra-tu-norma/busca-tu-norma/norma?c=N0060855

[14] ISO/IEC 15504-1:2004 Information technology – Process assessment – Part 1: Concepts and vocabulary. Available at: https://www.iso.org/standard/38932.html

[15] Métrica v3. Consejo Superior de Informática, 2001. Available at: https://administracionelectronica.gob.es/pae_Home/pae_Documentacion/pae_Metodolog/pae_Metrica_v3.html?idioma=en

[16] The Capability Maturity Model Integration (CMMI). Carnegie Mellon University (CMU). Available at: https://cmmiinstitute.com/cmmi/intro

[17] Control Objectives for Information and Related Technologies (COBIT). ISACA. Available at: https://www.isaca.org/resources/cobit

[18] Standard of Good Practice for Information Security 2020. Information Security Forum (ISF). Available at: https://www.securityforum.org/tool/standard-of-good-practice-for-information-security-2020/

[19] See the position of the European Telecommunications Standards Institute (ETSI) and the International Telecommunication Union (ITU) on this subject ("White Paper No. #34 Artificial Intelligence and future directions for ETSI". ETSI. 1st edition: June 2020. Available at: https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp34_Artificial_Intellignce_and_future_directions_for_ETSI.pdf. "Artificial Intelligence (AI) for Development Series Module on Setting the Stage for AI Governance: Interfaces, Infrastructures, and Institutions for Policymakers and Regulators" ITU. July 2018. Available at: https://www.itu.int/en/ITU-D/Conferences/GSR/Documents/GSR2018/documents/AISeries_GovernanceModule_GSR18.pdf),

[20] In this respect it is relevant to take into consideration the action of the following standardisation committees: CTN27 (ISO/IEC 27037:2012); ISO 27050 electronic Discovery, CTN320, and TC 46 SC11 (Record's Management), which is the ISO standardisation committee on Artificial Intelligence and Big Data.

The life cycle of an AI-based component can be broken into several phases: design and analysis, development (which includes research, selection, analysis and data cleansing, prototyping, design, training, testing, implementation in software and/or hardware, integration as part of a global processing and validation), operations and maintenance (including evolutionary maintenance) or final disposal. The above stages admit iteration, depending on the selected development model.

As in any audit procedure, the first step is to establish its scope. Such scope may cover all aspects of the life cycle of the component development and the processing in which it is to be implemented. In this case, the scope of the audit would be total. It is also possible to limit the audit scope to certain specific phases and aspects of the AI-based components. For example, an audit limited to the data sets used, aspects of the development methodology followed, implementation of the component design in a hardware element of software library in a particular environment, or the way in which the component is integrated in a processing deployed in a given context.

Within the vast range of quality audits, both process and product audits can be carried out. The goal of the relevant audit may be to ensure compliance with personal data regulations and thus prevent situations of non-compliance. In the same way, potential risks associated with the use of personal data in AI-based components can be identified, anticipated and corrected. This allows to reinforce accountability mechanisms implemented by the controller in order to prove compliance of their obligations regarding data protection. Audits can also contribute to carry out an in-depth analysis of the functioning of the AI component, in order to implement transparency mechanisms to allow the controller to be aware of and justify any decisions regarding system design, development and/or selection.

B. SINGULAR CHARACTERISTICS OF THE METHODOLOGY OF THE IA-BASED COMPONENT AUDIT WITH REGARD TO DATA PROTECTION

As in any audit, the first decision to be made is to define its purpose and scope. These may be determined either externally (certifications, evidence of compliance for a third party of other obligations included in the relevant regulations) or by the interests of the organization.

The criteria that must guide a data protection audit that includes an AI-based component must be the principles relating to processing of personal data defined in the Regulation: lawfulness, fairness and transparency, purpose limitation, data minimisation, accuracy, storage limitation, integrity and confidentiality, and accountability (art. 5 of GDPR).

Selection of control objectives and controls to be considered in the audit process, the extension of the analysis carried out by the auditor and the level of formality required by the auditor when implementing each control shall depend, as in any audit, on the goal and scope defined for such audit, as well as of the risk analysis carried out by the auditor. Since the goal of the audit described in this document is compliance with the GDPR, the corresponding risk analysis must be limited to the scope of the rights and freedoms of the persons whose data are processed.

Besides, the risk analysis of the audit process itself must be considered with regard to complying with the control objectives stated in the context of the specific AI-based component. A risk analysis of the process must be focused in identifying those aspects which may have an impact on achieving the desired control objectives and which shall be related to planning, available resources, audit team, control of documented information, availability and cooperation of persons in charge, the conditions regarding both the AI-based

component itself and its context, monitoring and revised procedures related to the designed audit plan, legal and administrative issues and any other circumstances.

Therefore, as in any audit recommendation document, the listing in this document of a series of potential controls does not imply the obligation to systematically apply each and every one of them, but rather to rationally select those relevant to comply with the corresponding control objectives and scope defined for the audit. For example, in case of a component which, based on an analysis of input data, makes decisions which may significantly affect a person, denying him or her access to essential services or restricting his or her freedoms, obviously the scope of the corresponding audit and the degree of exhaustivity when analysing the proposed controls must be higher than would be for a component whose function is, for example, limited to classifying certain e-mails in the Spam folder.

For the control objectives and controls described in the following chapter to be applicable, they must be based on the premise that the AI-based component is to perform personal data processing at some stage of its life cycle, or that the related personal data processing is profiling or making automated decisions regarding natural persons which may involve legal consequences or significantly affect those persons. In some cases, this may involve a previous analysis of the degree of anonymisation performed on the data used in processing, both in general terms and by the AI-based component in particular, the calculation or estimation of the possible risk of re-identification that exists, or the calculation of the risk of data loss in the cloud[21], among others.

Given the type of audits considered here, the audit team must have personnel with sufficient knowledge both about the AI solution being audited and the data protection regulations. Besides, if the controller's organization has appointed a Data Protection Officer or DPO, such DPO should be available for the audit team so that they can answer any question regarding the purpose, nature, scope and context of the processing affecting the behaviour of the used AI component. In the same manner, and depending on the development model, it could be advisable to include a data scientist in the audit team.

[21] Cloud Data Loss Prevention (DLP) can compute four re-identification risk metrics: k-anonymity, l-diversity, k-map, and δ-presence. A dataset is k-anonymous if quasi-identifiers for each person in the dataset are identical to at least k – 1 other people also in the dataset. A dataset has l-diversity if, for every set of rows with identical quasi-identifiers, there are at least l distinct values for each sensitive attribute. Computes re-identifiability risk by comparing a given de-identified dataset of subjects with a larger re-identification—or "attack"—dataset. Delta-presence (δ-presence) estimates the probability that a given user in a larger population is present in the dataset and it is used when membership in the dataset is itself sensitive information. More information is available at: https://cloud.google.com/dlp/docs/compute-risk-analysis

III. CONTROL OBJECTIVES AND CONTROLS

This section includes a set of control objectives and a list of controls to be considered when auditing AI-based components included in personal data processing and/or which make automated decisions affecting natural persons.

As stated before, the selection of controls to be audited, the extension of the relevant analysis and the level of formality required in their implementation shall depend, as in all audits, on the objectives and scope defined for the corresponding audit, as well as of the risk analysis carried out. Therefore, the auditor must select, out of the list of proposed controls, those which are adjusted to the specific audit being conducted and add those that they deem appropriate.

A. IDENTIFICATION AND TRANSPARENCY OF THE AI-BASED COMPONENT

Control objective: Inventory of the audited AI-based component

Compliance with the principle of accountability[22] requires traceability, and therefore, a correct identification of the AI-based component included in the audited processing.

Controls:

- The AI-based component is identified in the documentation by means of a name or code, identification of version[23] and date of creation.
- Both the code and any additional files defined by the version must include a digital signature over the entire package to guarantee its integrity.
- A version history of the evolution of the AI component used must be available and documented. It will include the parameters used in the training of the component and everything that ensures the traceability of the evolution/changes in the component.

Control objective: Identification of responsibilities

Compliance with the principle of accountability requires a clear identification of roles related to the processing being audited, which includes the relevant AI-based component, and the responsibilities of each of the involved parties[24].[25]

Controls:

- Identification and control data of the person(s) or institution(s) who manage the life cycle stages of the AI-based component, as well as their associate managers, and the representatives of the controller and of the processor.
- Contracts associated to processing stages being audited must specify the distribution of responsibilities with regard to personal data protection.

---

[22] Article 5.2 of the GDPR – Principles relating to processing of personal data Accountability. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e1807-1-1

[23] When it is a component developed based on previous versions of the same component, it will be useful to know how this version differs from previous ones.

[24] Chapter IV of GDPR – Controller and processor . Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e3022-1-1

[25] The process of debugging responsibilities includes the entire supply chain of the system in which the AI component is implemented. In the current procedures of hardware and software continuous integration these cycles have associated chains of dependencies where it is necessary to achieve an adequate convergence between user confidence and machine reliability. See Olav Lysne's book "The Huawei and Snowden Questions: Can Electronic Equipment from Untrusted Vendors be Verified? Can an Untrusted Vendor Build Trust into Electronic Equipment?" 2018. Springer Nature

- Registration in the Records of Processing Activities of the corresponding controllers and processors with regard to the personal data processing being audited.
- Establishing whether a Data Protection Officer must be necessarily be appointed, and, in such case, identification and communication of the identity of such Data Protection Officer to the relevant Supervisory Authority.

Control objective: Transparency

Compliance with the principle of transparency[26] and the obligation to provide information about the processing procedure to data subjects[27], requires that both the data source and the logic of the AI-based component are accessible, understandable and can be explained.

Controls:

- Data sources are documented, and an information mechanism has been implemented.
- The characteristics of data used to train the AI component are identified, documented, and duly justified.
- Considering efficiency, quality and accuracy of the AI-based component, the most appropriate model has been chosen (using criteria of simplicity and intelligibility), among several concurrent components[28], and taking into account its codification to facilitate readability, logic comprehension, internal consistency and explainability[29].
- Information regarding metadata[30] 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.
- The logic of the relevant AI-based component is well documented so it can be understood. Its behaviour regarding input data sets, data use, intermediate data and output data can be traced.
- If an erroneous behaviour of the AI-based component can cause harm to data subjects, mechanisms have been established to minimise such damage, provide communication channels to relevant stakeholders and facilitate communication among all stakeholders involved in the process.

[26] Article 5.1.a and Chapter III - Section 1 of the GDPR - Principles relating to processing of personal data Lawfulness, fairness and transparency Available at: (https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e1807-1-1) and Transparent information, communication and modalities (https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e2182-1-1) respectively.

[27] Articles 13.2.f and 14.2.g of Chapter III - Section 2 of GDPR Information and access to personal data Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e2244-1-1

[28] The current state of development of the AI component will have to be taken into consideration and advice will have to be sought from experts and researchers in artificial intelligence (deep learning, natural language processing, etc.) Regarding deep learning see as an example the review by Xie, N., Ras, G., van Gerven, M. and Doran, D., 2020, "Explainable deep learning: A field guide for the uninitiated" arXiv preprint arXiv:2004.14545.

[29] Recently, there has been a growing concern to understand and explain how Artificial Intelligence models make decisions based on data gathered in their application environment, particularly where such decisions have consequences for humans (e.g. medical diagnostics). There are several reasons for this growing concern. http://blogs.tecnalia.com/inspiring-blog/2019/11/14/explicabilidad-transparencia-trazabilidad-equidad-no-precision-uso-responsable-la-inteligencia-artificial/ (only in Spanish)

[30] Metadata of the AI component are the parameters used in learning processes.

B. PURPOSE OF THE AI-BASED COMPONENT

Control objective: Identification of intended purposes and uses

In compliance with the principle of purpose limitation[31], the purposes for which the data processed by and for the AI component are used must be established, explicit and lawful and not use in an incompatible manner.

Controls:

- The intended purpose of the relevant AI-based component must be documented both in quantitative and qualitative terms, with a clear description of what is intended to be achieved by using it in the framework of processing.
- The relationship between the goal pursued by the use of the AI component in a given processing operation and the conditions guaranteeing the lawfulness of such processing is documented.
- 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.
- Potential users of the AI-based component are categorized.
- When appropriate, other possible uses and secondary users have been described[32] together with the legal grounds for its use.

Control objective: Definition of the intended context of the AI-based component

In compliance with the obligation of analysing the processing context in which the AI-based component is integrated[33] it is necessary to know the circumstances in which the relevant processing occurs as well as any other factors which may impact on expectations of stakeholders and on the rights of data subjects. The analysis of these circumstances shall help to determine the most appropriate technical and organizational measures for ensuring compliance of regulatory standards.

Controls:

- The legal, social, economic, organizational, technical, scientific or other contexts related to the inclusion of the AI-based component in the processing must be documented.
- The organisational and/or contractual structure is defined between the parties and thus the distribution of tasks and responsibilities.
- The determining factors of the efficacy of said component are described. They include 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.
- The requirements applicable to human operators in charge of supervising and interpreting the operation of the AI-based component must be defined.

---

[31] Article 5.1.b of the GRPD – Principles relating to processing of personal data. Purpose limitation. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e1807-1-1

[32] Please note that this is an especially sensitive question when considering data protection, since using data that have been collected for an intended purpose for an entirely different purpose is a misuse that could go unnoticed. Therefore, any secondary uses should be documented and have legal grounds.

[33] Article 24.1 of the GDPR – Responsibility of the controller. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e3043-1-1

- When appropriate, interaction between the AI-based component with other own or third-party components, systems or applications, and the distribution of responsibilities for maintenance, updating and minimising system privacy issues must be documented.
- The levels or thresholds for interpreting and using the results yielded by the relevant AI-based level must be defined.
- Those contexts for which including the AI-based component in processing for reasons of incompatibility with its purpose or characteristics must be identified and described. It must also identify and describe whether the component has an inadequate level of reliability and/or accuracy with regard to the relevance that the processing could have for the data subject[34].

Control objective: Analysis of proportionality and necessity

When the AI-based component is audited in the context of a processing procedure requiring a data protection impact assessment, the necessity and proportionality [35] of using such AI-based component with regard to its intended purpose must be assessed.

Controls:

- The use of the AI component in processing must be assessed against other possible options from an approach focusing on the rights and freedoms of data subjects.
- In case of new developments, a comparative efficiency analysis and adequateness of results of the AI-based component must have 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.
- When addressing a new issue, the motivations and grounds for addressing this issue by using an AI-based component must be documented.
- When addressing a well-known problem, the grounds for changing the previous operation system that have led to a change in the previous mode of operation must be documented, describing, in its case, the new control objectives intended by using the AI component in the framework of the procedure.
- The risk to the rights and freedoms of data subjects introduced by using an AI-based component in data processing must be analysed and managed.

Control objective: Definition of the potential recipients of data

In compliance with the obligations provided in order to safeguard the rights of data subjects[36] especially those regarding transparency and provided data[37] the potential

[34] It is important to take into account when carrying out this context analysis the impact that the processing will have on the life of the data subject, especially for present and future opportunities.

[35] Article 35.7.b of the GDPR – Data protection impact assessment. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e3546-1-1

[36] Chapter III of the GDPR – Rights of the Data Subject. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e2161-1-1

[37] Articles 13.1.e and 14.1.e of the GDPR: Information to be provided where personal data are collected from the data subject (https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e2254-1-1) and information to be provided where personal data have not been obtained from the data subject, respectively. Available at: (https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e2355-1-1), respectively.

recipients or categories of recipients of personal data processed by the AI-based component must be identified, including those who are in third countries or international organizations.

Controls:

- Information obligation to data subjects with regard to data processing arising from the inclusion of the AI-based component are identified.
- Such obligations are identified both for data directly obtained from data subjects and for data obtained otherwise.
- When determining such obligations, the recipients or categories of recipients to whom the personal data processed by the AI-based component were or are to be communicated, including those who are in third countries or are international organizations, must be identified whether the data are obtained directly from them or from other sources of information.
- When determining such obligations, 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[38] decision on adequacy are identified. Data transfers made in compliance with appropriate guarantees, based in application of binding corporate rules or included in the exceptions referred by article 49, section 1, must include a reference to such guarantees, including how a copy of such guarantees may be obtained, or at least to the fact that they have been provided[39].
- Data recipients, including those from third countries or international organizations, are identified under the activity or activities recorded in the Records of Processing Activities in which the relevant AI-based component is included.

Control objective: Limitation of data storage

In compliance with the principle of storage limitation[40], and with the exceptions provided, data used by the AI-based component, either used for training or generated by it, must not be stored for longer than needed for their intended purposes[41].

Controls:

- 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 must be duly identified, especially when it is related with compatible purposes or included in any of the exceptions provided in the regulations[42] (for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes).
- The life cycle stages of the AI-based component in which the processed personal data are to be stored must be determined and justified.

---

[38] European Commission, How the EU determines if a non-EU country has an adequate level of data protection. Available at: https://ec.europa.eu/info/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en

[39] AEPD. International transfers. Available at: https://www.aepd.es/es/derechos-y-deberes/cumple-tus-deberes/medidas-de-cumplimiento/transferencias-internacionales (only in Spanish)

[40] Article o 5.1.e of the GDPR – Principles relating to processing of personal data. Storage limitation Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e1807-1-1

[41] The processing framework in which the audit is being carried out could be developing a component or, in other cases, the processing could be that in which such component has been incorporated.

[42] Article 89.1 of the GDPR - Safeguards and derogations relating to processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e6494-1-1

- Appropriate technical and organizational measures and criteria must have been defined to storage personal data[43].
- The time limits for erasure of stored personal data must have been defined.
- A conservation policy has been defined to keep a sample of training data for the purpose of auditing the AI component, considering the minimum or assumable risks for the data subjects.
- There are procedures to verify that such storage periods, criteria and measures are implemented.
- A procedure for reviewing the analysis of the need and the proportionality of data storage has been defined for those cases where an excessive pattern of data storage has been detected, either in terms of time or quantity.
- A storage policy has been defined for personal data included in the activity records of the AI-based component and privacy strategies (minimisation, hiding, separation or abstraction) must be implemented for operation purposes.

Control objective: Analysis of categories of data subjects

When performing a data protection impact assessment[44] in the framework of a processing procedure including the AI component, the categories of data subjects affected by the processing must be identified and the appropriateness of engaging them (or their representatives) in the assessment process must be evaluated[45].

Controls:

- The categories of data subjects affected by the development of the AI component and its use in the framework of the intended processing are identified.
- The short- and long-term consequences that the implementation of the AI component may have on the categories of data subjects are identified.
- The necessary procedures have been defined to analyse the social context in which the AI component is used, and to collect 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.

C. BASES OF THE AI COMPONENT

Control objective: Identification of the AI-based component development policy

The internal system development policy, particularly the development policy for AI-based components, must be consistent with the organization's data protection policy[46]. It must be

[43] In particular, the minimisation, obfuscation and/or pseudonymisation measures adopted.

[44] Article 35.9 of the GDPR – Data protection impact assessment. Views of data subjects. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e3546-1-1

[45] Recent studies propose to enable response or appeal mechanisms that allow the data subjects, or groups representing them, to intervene and question the logic or outcome of the AI component in the case of automated decision-making processes. See the paper by Kluttz, Kohli, and Mulligan (2018) "Contestability and Professionals: From Explanations to Engagement with Algorithmic Systems. Available at *SSRN 3311894.*

[46] Article 24.1 of the GDPR – Responsibility of the controller. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e3043-1-1

necessary to detail and complement this policy in specific aspects as needed, besides being aligned with the GDPR, and any other sector regulations.

Controls:

- Documents with development policies of products and systems must consider the data protection policy.
- The policies are reviewed, and version controlled.

Control objective: Involvement of the DPO

In compliance with the position and tasks assigned to the Data Protection Officer[47] internal procedures for communication and assignment of responsibilities that allow the relevant DPO to take an active role in providing due advice and to be able to actively participate in the selection, design and/or development of the AI component which supports personal data processing must be defined.

Controls:

- The Data Protection Officer must have the necessary professional qualifications and, particularly, the legal and technical expertise, as well as data protection practice appropriate to the project.
- The Data Protection Officer is assisted and advised by experts on specific matters relating to the AI component being audited.
- Internal procedures must have been defined within the organisation for correct communication between the Data Protection Officers 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.
- The Data Protection Officer has played an active role in the stages being audited, his or her independence of judgment within the organisation and his or her obligations to cooperate with the supervisory agencies have been respected and his opinions and considerations have been taken into account. The opinions and remarks of the Data Protection Officer have been considered.

Control objective: Adjustment of basic theoretical models

For any processing activity to be considered fair[48], it must be eligible for the declared purposes of such processing.

Controls:

- An study and analysis has been carried out regarding the theoretical framework and previous similar experience on which the development of the AI component is based.
- The basic hypotheses and premises considered in order to create and develop the relevant model must have been accurately described, justified and documented[49].

[47] Section 4 of Chapter IV of the GDPR (articles 37,38 and 39) – Data Protection Officer. Available at: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e3782-1-1

[48] Article 5.1.a of the GDPR – Principles relating to processing of personal data. Lawfulness, fairness and transparency Available at: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e1873-1-1

[49] The audit team may consult this information before or after drawing up the Analysis Plan, depending on how they consider that this may condition the objectivity of the audit.

- A procedure has been established for the critical and verified revision of the reasoning arising from acceptance of important hypotheses for the development of the AI-based component (for example, examining which are the arguments for a causal relationship that models an algorithm, such as the selection of variables defining a certain phenomenon).
- A careful analysis has been carried out in order to establish appropriate premises regarding the potential proxy variables intervening in the AI-based components.

Control objective: Appropriateness of the methodological framework

For any processing activity to be considered fair, it must be eligible for the declared purposes of such processing.

Controls:

- 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, must be duly documented.
- Depending on the results of the analysis of the problem to be solved and in a justified way, the development model to be used is determined (for example: supervised, unsupervised or others). In case of supervised models, the model for supervising the learning process of the algorithm, also the degree of supervision and the basis for such supervision are specified.
- The metrics for measuring the behaviour of the AI component have been selected and measured.
- A procedure for recording and monitoring the deviations in the behaviour of the AI component with respect to the defined metrics that allows to carry out a monitoring of the circumstances which may arise in an erroneous or biased behaviour has been implemented.

Control objective: Identification of the basic architecture of the AI-based component

In compliance with the principle of accountability the development of the AI-based component is documented in such a way that allows to understand any aspects related to its implementation, its operating context and the interconnections with other processing elements.

Controls:

- The project analysis phase of the AI-based component must include, as part of the requirement catalogue, a series of specific requirements too guarantee privacy and personal data protection.
- When programming AI-based components, the coding principles[50], codes and coding best practices[51] [52] [53] applied must be followed and documented in order to guarantee that the code is readable, secure, low-maintenance and robust.

[50] What Are The Best Software Engineering Principles?. Luminousmen Blog, 2020. Available at: https://luminousmen.com/post/what-are-the-best-engineering-principles

[51] Clean Code: A Handbook of Agile Software Craftsmanship, Robert C. Martin, Pearson 2008

[52] Clean Code Summary. Samuel Casanova. Available at: https://samuelcasanova.com/2016/09/resumen-clean-code (Spanish)

[53] Clean Architecture: A Craftsman's Guide to Software Structure and Design, Robert C. Martin, Pearson 2017

- The basic architecture of the AI component must be identified and documented, including 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[54].
- A systematic procedure for documenting the component implementation procedure must exist and be implemented in order 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[55] and useful documents such as requirements specifications, functional and organic analyses, guidelines, etc.
- When the AI component code was impossible to access, a reverse-engineering process or other alternative method must be used, such as the use of a zero-knowledge proof (ZKP) which, despite not being able to access the code, 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.

D. DATA MANAGEMENT

Control objective: Data quality assurance

In compliance with principles relating to processing of personal data[56] processed personal data must be accurate and up-to-date with regard to the purposes for which they are processed.

Controls:

- There must be 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.
- There must be supervisory mechanisms for data collection, processing, storage and use processes.
- A previous analysis must have been carried out together with a measurement of the sample used for training the relevant model. It must have been verified that sample size is adequate, as well as whether frequency and distribution of each

[54] A loss function J(θ) measures the level of unsatisfaction with regard to the model predictions with respect to a correct response and using certain θ values. There are several loss functions, such as the mean squared error or cross-entropy. Choosing one or the other depends on several factors, such as the selected algorithm or the desired level of confidence, but mainly depends on the purpose pursued by the model. Since the purpose of training an specific model is to make predictions which are as close as possible to the correct response and minimising erroneous or unsatisfactory outcomes, the goal is to find the optimal values for θ which minimise the results of the loss function. This is known as optimization. As happens with the loss function, there are several optimization methods which directly impact model performance and training time. One of the most used methods is the gradient descent method https://www.iartificial.net/gradiente-descendiente-para-aprendizaje-automatico/ (Spanish)

[55] API: qué es y para qué sirve (API: What it is and what it is for). XATACA, 2019. Available at: https://www.xataka.com/basics/api-que-sirve

[56] Article 5.1 of the GDPR – Principles relating to processing of personal data Principle of accuracy. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e1807-1-1

variable, their intersection or the relevant groups for the study are appropriate regarding defined parameters or to reality.

- Analyses must have been carried out: both at the beginning and in each iteration of the global learning process, and on the sample used to train the model. It has 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.
- It must have been verified that variable distribution is appropriate and that the component is not especially sensitive or ignores any of them.
- There must be procedures to analyse, measure and detect any possible imbalances between the amount of data that the component collects on a certain variable with respect to another and which may lead to behaviour deviations.
- An accurate compensation analysis has 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.
- A sample size analysis has been carried out regarding data storage for audit purposes.

Control objective: Definition of the origin of the data sources

In compliance with principles relating to processing of personal data[57], personal data must be processed according to the principles of lawfulness, fairness and transparency and for specified, explicit and legitimate purposes (principle of purpose limitation) with the express prohibition for processing special categories of personal data[58] except in the circumstances and with the derogations provided for.

Controls:

- The origin and the data sources context used for training and validating the model must be identified.
- The selection process of data sources used to train the relevant AI-based component must be documented and justified.
- There must be legal grounds to used personal data in the different stages of the AI-based component life cycle.
- The collection and use of personal data is justified 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[59].
- If sensitive personal data are processed, the need for their use has been assessed and certain circumstances justify to lift the general prohibition to process such data.

[57] Article 5 of the GDPR – Principles relating to processing of personal data. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e1807-1-1

[58] Article 9 of the GDPR – Processing of special categories of personal data. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e2051-1-1

[59] Additional information linked to data subjects for the purposes of subsequent analysis in case of accuracy problems or bias issues due to characteristics that should not be associated to the decision-making process. For example, discrimination on the grounds of gender, race, social origin, etc.

Control objective: Preprocessing of personal data

In compliance with principles relating to processing of personal data[60], personal data need to be processed in application of the principle of minimisation.

Controls:

- Criteria to carry out previous cleansing of original data sets and any other tasks identified as required throughout the different iterations of the AI-based training process must be well identified and documented.
- Data cleaning techniques and best practices[61] used in the data cleansing process must be well grounded and documented.
- Classifying variables must define clearly distinguishable and identifiable types.
- The structure and properties of the processed data set must be documented including the number of data subjects and the extension of used data.
- Previously, used data must have been 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.
- The relevant variables for the model must have been determined, identifying the variables associated with special data categories and proxy variables, including the necessary information for their interpretation.
- Data minimisation criteria must have been determined and applied to the different stages of the AI component, using strategies such as data hiding, separation, abstraction, anonymisation and pseudonymisation as may apply for the purposes of maximising privacy in the operation of the relevant AI-based component.
- Used databases must have an associated data dictionary that enables their analysis and understanding.
- Segregation and de-identification strategies must have 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 in order to analyse correlations between variables, measure the degree of accuracy of the AI component with regard to certain attributes or ensure that no biases are introduced[62].
- Data selection and assessment must have been carried out with the involvement of an expert in modelling techniques and data science who is in charge of understanding the complex processes from reality that are being modelled, and of analysing and interpreting the data used by the component.
- Data to be used for training and validating the AI-based must have been previously pre-processed and cleaned in order to detect any possible abnormality which may require previous processing (boundary values, incomplete records, etc.) and to convert any heterogeneous data sources to a homogeneous format.

[60] Article 5 of the GDPR – Principles relating to processing of personal data Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e1807-1-1

[61] Data cleansing procedures include a series of techniques which, applied on the dataset registries, enable detection and correction of records in order to reduce duplications and inconsistencies. More information available at: https://elitedatascience.com/data-cleaning

[62] For example, in order to control that a certain AI-based component does not discriminate against on the basis of gender, information on the gender of persons included in the database must be collected for the purposes of verifying whether the component behaves differently depending of the value of the gender variable, regardless of whether the gender variable is used in the AI component training or not.

- The necessary modifications must have been introduced in the format of input data, when it is not appropriate with regard to the functioning of the AI component or because it is not representative of the reality it intends to reflect[63].
- When appropriate, an analysis is carried out regarding the degree of data anonymisation and the possible risk of re-identification.
- When appropriate, if data imputation techniques have been used to complete the information of the data set, the procedures and algorithms used for such imputation have been documented.

Control objective: Bias control

In compliance with principles relating to processing of personal data[64] processed personal data must be accurate and updated with regard to the purposes for which they are processed.

Controls:

- Appropriate procedures must have been defined in order to identify and remove, or at least limit, any bias in the data used to train the relevant model.
- It must have been verified that in training data used as input for the relevant model there are no previous historical biases and, if they are, a different and bias-free data source has been chosen or training data have undergone appropriate cleaning and scrubbing so that are suitable for use.
- Appropriate measures must have been adopted in order to assess the need to have additional data for improving precision or removing any possible bias.
- Human supervision mechanisms must have been implemented in order to control and ensure that results are bias-free.
- The implemented mechanisms must enable data subjects to request human intervention, provide feedback or challenge the results obtained by means of automated decision-making algorithms.

E. VERIFICATION AND VALIDATION

Control objective: Adapting the verification and validation process of the AI-based component

In compliance with principles relating to processing of personal data[65], and more specifically the principle of accountability[66], it must be possible to prove that the methods used to include the relevant AI-based components in the processing activities or their development are compliant with the principles relating to processing of personal data and any other obligations provided in any data protection regulations.

Controls:

[63] For example, considering an AI-based component intended for processing natural language, if its functioning does not have the capacity to adjust to changes in the words that make up input texts, it is possible that it does not behave as intended. In this case, it is recommended to either select a more orderly and mappable input or adapt the functioning of the relevant component to the existing format of the input data.

[64] Article 5.1.d of the GDPR – Principles relating to processing of personal data. Principle of accuracy. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e1807-1-1

[65] Article 5.1.b the GDPR – Principles relating to processing of personal data. Purpose limitation Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e1807-1-1

[66] Article 5.2 of the GDPR – Principles relating to processing of personal data. Principle of accountability https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e1807-1-1

- The verification and validation process, the techniques used, the verification and test assembly carried out, the results obtained, and the proposed actions must be duly documented.
- Guidelines, standards or regulations must have been established or followed 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.
- The necessary control and supervision mechanisms must be implemented in order to ensure that the AI-based component efficiently complies with its intended goals and purposes.
- Metrics and criteria, on which verifications within the verification and validation process shall be carried out, must be defined and justified.
- A testing strategy must be defined, and, related with this strategy, there is a complete testing plan to assess the correction of the AI component both from structural and functional terms.
- Personnel involved in AI-component verification and validation tasks must be qualified to carry out the necessary checks in order to ensure that the component has been correctly built and behaves as expected.

Control objective: Verification and validation of the AI-based component

In compliance with principles relating to processing of personal data[67], it must be possible to prove that the AI component processes personal data appropriately and in compliance with the accuracy principle.

Controls:

- The testing plan must 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.
- White-box testing of the network design or the AI component must be considered as part of the testing plan[68].
- White-box testing[69] at code and implementation levels must be considered as part of the testing plan.
- Black-box testing[70] as required must be considered as part of the testing plan in order to ensure that functionality of AI-based component is guaranteed, that such component behaves as expected and that information integrity is preserved.

[67] Articles 5.1.a and 5.1.b of the GDPR – Principles relating to processing of personal data. Lawfulness and purpose limitation https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e1807-1-1

[68] For example, for neural networks, these tests include analysis of neuron coverage, threshold coverage, sign change activation, layer coverage, etc. References to this can be found in the AI systems certification scheme of the Korean Software Testing Qualification Board: https://imbus.cn/upFile/Uploadfiles/AI%20Testing_Testing%20AI-Based%20System%20Syllabus%20v1.3.pdf

[69] This type of test is strongly linked to the source code, and, from a quality and security perspective, is key for early detection of failures and vulnerabilities presented by the component in its development and effective implementation. Particularly, it is necessary to carry out an static analysis intended to establish whether the generated code is correct, it may have been tampered with, there are dead, unreachable or redundant code areas, as well as to verify codification of back door in libraries or other elements used in the implementation of the relevant component which may entail a modification of defined functional and non-functional specifications. Usually, they follow a bottom-up approach, inspecting and verifying he behaviour of each component individually before proceeding to integrate such components into their intended system, so that, once said components have been validated individually and the integration has been completed, the global behaviour of the system can be tested.

[70] Black-box testing may include equivalence partitioning, boundary value analysis, decision table testing, state transition testing and use case testing.

- As part of the test plan, the necessary tests must be provided to check security, 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[71].
- The validation test plan must include verification of boundary values and extreme test cases which may make the component to function in an unexpected manner.
- There must be a documented cleaning procedure to correct any errors, shortcomings or inconsistencies detected during the verification and validation process.

Control objective: Performance[72]

In compliance with principles relating to processing of personal data[73], AI component must process personal data appropriately and in compliance with the accuracy principle.

Controls:

- Metrics or sets of aggregated metrics used to determine the precision, accuracy, sensitivity[74] [75] [76]or other performance parameters of the relevant component in consideration of the principle of data accuracy must be appropriately established.
- The rate values of false positives and false negatives yielded by the AI component must be known and must have been analysed and interpreted in order to establish their accuracy, specificity and sensitivity of the component behaviour.
- The level and definition of the performance parameters required for the AI-based component in the framework of the processing supported by such component must have been assessed.
- The performance values between different options of AI components have been compared in the context of a process of selection of the most appropriate component for a specific processing procedure.
- Output variables must be defined and determined with special consideration to those that constitute special data categories.
- Appropriate measures have been adopted in order to ensure that data used are exhaustive and updated.
- Both the relevant parameters and their cut-off values must be determined so that the model considers certain variables in order to obtain significant results.

[71] See Mayoral Vilches, V., Olalde Mendia, G., Perez Baskaran, X., Hernández Cordero, A., Usategui San Juan, L., Gil-Uriarte, E., Olalde Saez de Urabain, O. and Alzola Kirschgens, L., 2018. Aztarna, a footprinting tool for robots. arXiv, pp.arXiv-1812.

[72] The term "performance" is used in the sense of being effective with regard to data protection, rather than in the sense of being efficient or effective with regard to other aspects. Depending on the relevant processing or component, the confusion matrix, accuracy, sensitivity, specificity, or the receiver operative characteristic (ROC) curve and the area under the curve (AUC) based on the rate of true positives, false positives, true negatives and false negatives More information can be found, among others, on: https://sitiobigdata.com/2019/01/19/machine-learning-metrica-clasificacion-parte-3/, https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/, https://imbus.cn/upFile/Uploadfiles/AI%20Testing_Testing%20AI-Based%20System%20Syllabus%20v1.3.pdf

[73] Article 5.1.d of the GDPR - Principles relating to processing of personal data – Principle of accuracy.. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e1807-1-1

[74] Precision represents how close a result is from the true value; accuracy represents the rate of correct predictions (that is, correct predictions of a true statement) and specificity represents the fraction of true negatives (that is, incorrect predictions of a false statement).

[75] Falsos positivos, o la importancia de comprender la información. (False positives, or the importance of understanding information.) Cuaderno de cultura científica, 2015. Available at: https://culturacientifica.com/2015/10/07/falsos-positivos-o-la-importancia-de-comprender-la-informacion/

[76] Machine Learning: Selección de métricas de clasificación (Machine Learning: Selection metrics for classification.) SitioBigData.com Available at: https://sitiobigdata.com/2019/01/19/machine-learning-metrica-clasificacion-parte-3/#

- There must be 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.
- A dimension reduction[77] must have been carried out in order to achieve a balance between complexity and generalization.

Control objective: Consistency

In compliance with principles relating to processing of personal data[78], AI component must process personal data appropriately and in compliance with the accuracy principle.

Controls:

- A procedure must have been set in order to verify whether the obtained results presents significant changes with respect to the results expected, and to act accordingly.
- A threshold must have been established in order to determine when an obtained result deviates from the expected result based on identical or similar data inputs (significant deviations).
- It must have been determined 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.
- The effect of changes in low prevalence variables within the training dataset in output results of the AI-based component must have been assessed.
- Appropriate measures must have been adopted to ensure component independence[79].
- It must has been verified that there is no correlation between the results and the additional variables associated to data subjects[80] that are not a part of the process variables and which may evidence the existence of biases.

Control objective: Stability and robustness

In compliance with the principle of accountability[81] the AI-based must be subject to continuous monitoring processes in order to adjust to the modifications occurred in the

[77] Model dimensionality is related to the relationship between variables fed to the model for learning purposes and the number of samples used for training. If the number of parameters is very high, the model will be too well adapted to the characteristic of training data capturing all relevant information but also all existing noise; this is called *overfitting*. Therefore, in the testing and validation phases, when making predictions regarding new data, prevision shall be notably significantly lower and the model will not be capable of generalizing rules to predict results with regard to data that it has not seen. This emphasizes the importance of dividing the input data set used in the training phase in two disjointed sets (training and validation, usually in an 80-20 rate) in order to find out the behaviour of the model when fed new data. http://www.revistaindice.com/numero68/p22.pdf

[78] Article 5.1.d of the GDPR - Principles relating to processing of personal data – Principle of accuracy.. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e1807-1-1

[79] An AI-based component is deemed to be independent if the possibility of generating results is not determined by the attribute that defines an specific group.

[80] Discrimination arising from decision-making based on characteristic correlated to data such as race or gender, which are linked to protected categories, is known as proxy discrimination. More information available at: https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3572098_code499486.pdf?abstractid=3347959&mirid=1

[81] Article 5.2 of the GDPR – Principles relating to processing of personal data. Accountability Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e1807-1-1

environment and detect any readjustment requirement[82] as a consequence of changes in the context either internal or external to processing[83].

Controls:

- Within the possible or actual context of function of the relevant component, the factors whose variation may impact the properties of the AI component and may establish the need to manage its readjustment must be identified.
- The behaviour of the AI component in unexpected environments or use cases must be assessed.
- A time estimation in which a 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 must have been carried out.
- It must be documented whether the AI component has been built with an static approach, a dynamic approach or a continuous learning approach by design[84].
- As for continuous learning AI component, the degree of adaptability to new input data or types of data must have been assessed, and the monitoring procedures and mechanisms must have 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[85].

Control objective: Traceability

In compliance with principles relating to processing of personal data[86] and, most specifically, the right of data subjects to not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her[87], behaviour of the AI-based component must be capable of being supervised through traceability mechanisms, including human means.

Controls:

- There must be 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.
- There is 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.
- Appropriate monitoring and supervision mechanisms must have been implemented for the AI-based component, such as log files and results records, which enable to assess the behaviour of the component in interaction with environment, to measure that the relevant outputs are adjusted to the responses

[82] Such readjustment may entail re-training, a new weighing assessment or a restructuration, depending on the relevant type of AI used.
[83] Such as data deviations, new risks for the rights and freedoms of data subjects or changes in decision criteria that support the model.
[84] An AI component can be built using an static approach, that is, the behaviour defined in the models building and training phases remain unchanged according to data used in the learning process. Alternatively, the component, once in production, uses input data not only to generate an output but also to adjust and improve the model.
[85] In this type of systems, it is necessary to avoid the phenomenon known as “catastrophic forgetting", in which, as the systems learns new data different from those used in the training phase, and modifies the relevant parameters in order to adjust to new input data, it deems invalid and overwrites previous knowledge acquired in the training phases.
[86] Article 5 of the GDPR - Principles relating to processing of personal data Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e1807-1-1
[87] Article 22 of the GDPR – Automated individual decision-making, including profiling Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e2838-1-1

of real-life processes that they model and to any potential inconsistency between expected behaviour and the automated one.

- There must be a record of incidences and previous abnormal behaviours detected and remedied.
- Monitoring mechanisms must be available for human operators for monitoring and verification purposes.
- A procedure 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, must be implemented and documented.
- Appropriate mechanisms must be adopted within the framework of the processing so that the results and decisions taken may be entirely the responsibility of human operators.

Control objective: Security

In compliance of principles relating to processing of personal data.[88] and of the obligation of ensuring data protection by design and by default[89] and security of processing[90], the AI-based component must process personal data by applying the principles of data protection in an efficient and effective manner, and by integrating the necessary technical and organizational measures to ensure a level of security appropriate to the risk, and, especially, with regard to confidentiality, integrity, availability, and resilience of processing.

Controls:

- A risk analysis must have been carried out with regard to the risks for rights and freedoms of persons, and the results of this risk analysis must allow to determine the security and privacy requirements of the AI-based component in the framework of this proceeding.
- Those requirements related with data protection and security must have been 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.
- Available standards and best practices for secure configuration and development of the relevant component have been applied.
- The necessary measures to ensure protection of the processed data must have been 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.
- Measures must have been implemented to guarantee component resilience and its capacity to withstand an attack[91].

---

[88] Article 5.1.f of the GDPR – Principles relating to processing of personal data. Integrity and confidentiality Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e1807-1-1

[89] Article 25 of GDPR - Data protection by design and by default. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e3063-1-1

[90] Article 32 of the GDPR: Security of processing. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e3383-1-1

[91] Artificial intelligence attacks consists on unauthorised third parties manipulating AI-based component in order to change its behaviour to accomplish a malicious purpose. Attacks can target design components, libraries, software, hardware and any other implementation element. Since these types of components are increasingly integrated in components critical to society, such attacks represent a potential risk which may have significant effects on national security. Thus, the need to identify its potential vulnerabilities, such as limitations in algorithms or scarce controls in data used, and any sources of potential threats to implement security measures and best practices requirements to guarantee resilience. Comiter M. Attacking Artificial Intelligence - AI's Security Vulnerability and What

- Appropriate procedures must have been implemented in order to monitor the functioning of the component and early detect any potential data leak, unauthorised access or other security breaches.
- Component users and operators must have sufficient information and must be aware of their security duties and responsibilities regarding data protection and safeguarding data subjects' rights and freedoms.

Policymakers Can Do About It. Belfer Center for Science and International Affairs, 2019. Available at: https://www.belfercenter.org/sites/default/files/2019-08/AttackingAI/AttackingAI.pdf

IV. CONCLUSIONS

As stated in the introduction, one of the tools “to ensure and to be able to demonstrate” compliance with the GDPR is to perform audits on processing, and, at a later stage, supervision processes on conduct codes as set forth by article 40 of the GDPR and certifications, as per article 42 of said regulation. Both instruments require to have objective criteria for assessing regulatory compliance. Although general criteria may be common to all processing activities, some others may include specific features due, among others, to technological elements on which the processing is based. One of such technological elements to be considered is the incorporation of artificial intelligence-based components to processing of personal data, either for the development of the component itself, for operation purposes, or for any other possible scenario.

This document intends to be a first approach to determining a set of control objectives and controls, designed from a data protection perspective, in order to be included in the audit for a processing procedure which incorporates artificial intelligence components. In order to perform an audit, having a list of controls and control objectives constitutes a reference to establish and verify the fitness for purpose of the processing being audited, compare it to other processing activities and assess its evolution. And, as in all audits, the controls to be considered and their respective verification methods must be chosen and adapted by the auditor. Choosing controls which are appropriate for auditing an specific data processing procedure shall depend on several factors: type of processing, customer requirements, the specific audit, the purpose and scope of such audit, and the results of a processing risk analysis and the audit process itself.

Although auditing methods are well known, when focusing on the AI components they may present certain particularities that need to be considered, as stated in this document.

Considering the evolution undergone by this technological environment, this report can only be an orientation, evolving document, whose future versions shall have to contain feedback on its implementation.

V. ANNEX I: DEFINITIONS

For the purposes of this document, it is useful to previously define a series of relevant terms in order to enable understanding of developed concepts which are part of an AI-based component audit with regard to data protection.

Anonymisation

Following the specifications provided by the Regulation (Whereas 26[92] of the GDPR) this document considers that such *"information which does not relate to an identified or identifiable natural person"*. Therefore, anonymisation is understood as the process intended to convert data into anonymous data and break its link to the person they refer to, so that the person is no longer identifiable by such data.

AI-based component learning

There are four approaches to the development of AI-based components.

- **Supervised learning:** an operator acts as the "instructor" of the component, feeding training data to the system, and this training data include input data and also the correct output data for such input data; that is, they feed labelled data. The component must reproduce the same "pattern" on subsequent iterations in order to generate new output data under the same logic.
- **Unsupervised learning:** Unlike supervised learning, there is no operator feedback. Components are designed to be able to detect underlying patterns and rules in data, and to summarise and group information units included in data.
- **Semi-supervised learning:** This is a compromise between the two previous approaches. They include some labelled input data, although most of them are not, and automatic procedures are used as a complement.
- **Reinforcement learning:** In this case, the component shall be designed so as to observe the interaction between the system and its environment and leverage it to improve its function. During the training process, the system analyses and values different possible behaviours, with the goal to automatically determine the optimal choice in a specific context. The reinforcement signal consists in simple feedback understood by the system as a "prize" and enables to determine how "appropriate" is a given behaviour.

Audit

Audit is a systematic, independent and documented procedure designed to obtain objective evidence (records, statements of fact or any other information) and assess them

[92] Whereas 26 "The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes."

in an objective manner so as to establish the degree to which audit criteria are complied with (set of policies, procedures or requirements used as a reference)[93].

Data protection audit of AI-based components

This refers to the part of a data protection audit performed in a data processing procedure the purposes of which is limited to the AI-based components of the relevant processing.

AI-based components

The first edition of ISO/IEC TR 29119-11 "*Software and systems engineering -Software testing - Part 11: Testing of AI-based systems*"[94], and "*White Paper On Artificial Intelligence - A European approach to excellence and trust*"[95] or the European Parliament Paper "*Artificial Intelligence and Civil Liability*"[96] uses the term "AI-based system" for those systems which include one or more AI-based components[97].

An AI-based component is the implementation of an element that encapsulates the functions related to an artificial intelligence process and which may include algorithms, datasets and other elements that allow for the execution of said component. The relevant component includes both AI-specific aspects and those arising from its implementation in software and/or hardware, which may have an impact on the behaviour of said component.

The GDPR, as established in its article 2, applies to personal data processing. In this case, the AI-based component shall implement one stage of the processing.

Input data, output data and labelled data

Input data are those fed to the AI component to be processed.

Output data are data yielded by the algorithmic processing of input data.

Finally, labelled data are those data fed to an AI component while linked to certain output values. Labelling data allows the system to know certain contents of these data[98].

Personal data

This document uses the definition of personal data[99] provided by article 4.1. of the GDPR: *"any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who may be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online*

---

[93] UNE-EN ISO 19011:2018 - Guidelines for auditing management systems. Available at: https://www.iso.org/obp/ui#iso:std:iso:19011:ed-3:v1:en

[94] ISO/IEC TR 29119-11:2020 Software and systems engineering — Software testing — Part 11: Guidelines on the testing of AI-based systems. Available at: https://www.iso.org/standard/79016.html

[95] White Paper On Artificial Intelligence - A European approach to excellence and trust [online] COM(2020) 65 end, 30 [Last consulted: 29 October. Available at: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

[96] https://www.europarl.europa.eu/RegData/etudes/STUD/2020/621926/IPOL_STU(2020)621926_EN.pdf

[97] A similar statement is made by CENELEC in "CEN-CENELEC response to the EC White Paper on AI". Available at: https://www.cencenelec.eu/news/policy_opinions/PolicyOpinions/CEN-CLC%20Response%20to%20EC%20White%20Paper%20on%20AI.pdf

[98] In Artificial Intelligence systems, the role of the labeller is key to certify data validity and subsequently be able to help the machine to learn so that, in the middle term, artificial intelligence components do not need supervision.

[99] REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL, of 27 April 2016, on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, 4th May 2016. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L:2016:119:FULL&from=DE

identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person."

From the perspective of the AI-based component life cycle (please see the definition of life cycle below) we can find the following processing of personal data.

- When personal data are used in the development stage of the AI-based component.
- When personal data are used in the verification or validation stage of the AI-based component.
- When the IA-based component is included in a processing of personal data (operation stage) as could be a security control processing (including facial recognition) or citizen services (including chatbots).
- In any other stage of the life cycle in which personal data are involved.

Such processing may use datasets from which new personal data may be inferred. All such data either direct or indirect, original or derived, are personal data inasmuch they refer to an identified or identifiable person, and therefore subject to protection as per article 1 of the GDPR.

Personal data are classified in identifiers, quasi-identifiers and special categories of personal data.

The first category refers to such data which are by themselves univocally associated with an specific data subject, such as their ID document number, their full name, their passport number, their Social Security number or any other identifier that fulfils the same goal.

Quasi-identifiers, also called pseudo-identifiers or indirect identifies, are such data that, although they do not directly identify a data subject, conveniently grouped and related to other datasets or information sources, may enable identification of a person or linking or inference with sensitive data. Some data included in this category are date of birth, place of residence, postal code or gender, since they are shared by wide scope of datasets, many of them public, in which an specific data subject may be included.

Finally, special categories of personal data are those types of data to which special protection is given as per article 9 of the GDPR. More specifically, such data are those data *"revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation."*

Other data that, albeit not categorized as special, require special protection due to their nature, are personal data referred to criminal convictions (article 10 of the GDPR); its processing is limited and special guarantees are set forth such as carrying out a data protection impact assessment (article 35.3 b) of the GDPR).

Life cycle of an AI component

The life cycle of an AI-based component is the ensemble of stages in which its evolution is structured, from design to disposal.

The basic stages of the life cycle of an AI-based component could be, in the case of machine learning:

- A pre-processing stage based on the databases used for training and testing the system, and which may include incomplete, unstructured data in different formats,

and therefore the first task must be to prepare these data to be used. Generally, such data are divided into two sets: data used to generate the learning model and data used for its validation.

- A stage for preparation of the component code, which is later trained, in order to generate the algorithmic model. The chosen learning technique shall depend on the nature of the problem.
- A validation stage on the disjoint set of training data reserved to this purpose.
- If the model behaves in a reliable manner, it can proceed to the successive stages of implementation in a commercial model, inclusion in a procession, transfer to production, updating or maintenance and, finally, disposal.

It is quite usual that this development model includes an iterative verification process - re-learning of the component behaviour using real data so as to enable continuous adjustment and improvement.

Algorithmic discrimination

Algorithmic discrimination refers to the fact that an AI-based component may treat a certain individual X differently from other individual Y due to an specific attribute of X. This fact does not necessarily involve a negative or disadvantageous discrimination[100][101].

Group discrimination

Group discrimination refers to that discrimination affecting a person because he or she belongs to a socially identified or protected group.

Statistical discrimination

Statistical discrimination refers to group discrimination based on a statistically relevant fact[102].

Weak AI

Depending on the scope and field of application of artificial intelligence, two categories of AI may be distinguished: strong AI, superintelligence and weak AI. Strong AI could solve any intellectual task which can be solved by a human being. It should be noted that a subtype of AI type could be defined based on strong AI, super-intelligence, which is when AI would go beyond human capabilities. Weak AI, which is the intelligence currently implemented, is characterised by its capacity of developing solutions capable for solving an specific, well-defined problem. This document is focused on weak AI.

---

[100] The definition of discrimination and bias presented in this guide are mainly based on the work done by Barocas and Selbst (2016), Baeza-Yates (2018), Castillo (2018), Cowgill (2019), Hajian, S., Bonchi, F., and Castillo, C. (2016), Lippert-Rasmussen (2013), Pedreschi et al. (2008) Also in their interpretation for previous works published by Eticas Research and Consulting.

[101] An instance of discrimination which may positively impact on a vulnerable or protected group would be that a component developed to model resource allocation provides significantly more support to persons with disabilities that to persons without disabilities.

[102] This may be the case, for example, with a predictive component which uses probability data from the real world (and therefore, by reflecting the outcome of previous decisions, are statistically relevant), but whose use results in a disadvantageous treatment of a vulnerable community or social group. A real-life example of this is the case of an AI-based component designed to predict recidivism, which was proved discriminatory due to the use it made of information regarding repeat offenders from Black communities. More information on this case, concerning the case of the COMPAS algorithm, can be found on the following website: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm . It must be also considered, that, had a similar case happened in Europe or with European data, pursuant article 10 of the GDPR, use of such data regarding criminal convictions should had been duly reported to competent authorities and based on legal grounds.

Audit methodology

Regardless of the field of knowledge in which it is applied, methodology refers to the set of rational procedures, methods and techniques which are systematically implemented during an study or research process to pursue a specific goal.

In the specific case of the audit, the goal pursued is to be able to determine the degree of compliance of the audited processing with regard to the required auditing criteria or requirements, which may arise form regulatory provisions, internal policies and rules of the data controller, other plans defined within the organization (such as the quality assurance plan or the social responsibility plan) or other requirements specified by the stakeholders.

In a process of research and analysis many methodologies may be deployed, although they can be grouped in two broad categories: qualitative and quantitative research.

Quantitative research enables accessing information by means of collecting data regarding certain variables, and reaching certain conclusions by comparing statistics. Qualitative research described phenomena under research, forgoing its quantification, and obtaining information by means of interviews or non-numerical techniques, analysing the relationship between variables obtained form observation and considering, above all, the context and situations around the issue under research.

Control objectives and controls

Pursuant to the relevant ISO[103], control objectives constitute a statement of the goals pursued by the implementation of the different controls, understood as those measures by which risk is modified. Controls include processes, policies, devices or practices, among other actions, intended to modify risk. They may be preventive, detective or corrective depending on how they interact with the threat that leads to the risk.

Profiling

Profiling (as per Article 4.4. of the GDPR) is any form of automated processing of personal data which allows to infer additional information regarding a natural person, in particular to analyse or predict personal information concerning that natural person.

Any processing involving profiling is characterised by three elements[104]:

- It must consist in automated processing methods, including those processing activities which are partially performed by human beings.
- It must be performed in reference to personal data;
- And the purpose of such profiling must be to assess personal aspects of a natural person.

[103] ISO/IEC 27000 Information technology — Security techniques — Information security management systems — Overview and vocabulary. Available at: https://standards.iso.org/ittf/PubliclyAvailableStandards/c073906_ISO_IEC_27000_2018_E.zip

[104] Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679. Article 29 Working Party, 2017. Available at: https://www.aepd.es/sites/default/files/2019-12/wp251rev01-en.pdf

Profiling and decision making concerning natural persons are considered processing according to the Regulation (Whereas 24[105] and 72[106] of the GDPR), and, consequently, shall be subject to the provisions set forth therein.

Risk of re-identification

Re-identification risk analysis is a process of analysing data to find properties that may increase the risk of data subjects being identified. Risk analysis methods can be used before de-identification to help determine an effective de-identification strategy or after de-identification to monitor any changes or outliers.

Algorithmic bias

Algorithmic bias[107] occurs in such cases in which an specific AI-based component generates different results for different data subjects depending on the fact that they belong to an specific community or category (explicitly or on an ad-hoc basis), evidencing an underlying prejudice towards such collective.

This behaviour can be derived from several sources: a bias in training data, a bias in training method (for example, a bias in supervision), an underfitting (overly simplistic) model, an application of the AI-based component in an inappropriate processing or context, etc.

Proxy variables

Proxy variables are variables used instead of the intended variable when such intended variable cannot be measured directly[108]. Although proxy variables are not a direct measure of the intended variable, a good proxy variable is strongly related to the values of the intended variable[109].

In sum, proxy variables[110] are those which can be measured directly and present a sufficiently close correlation to the interest variable, which cannot be measured directly, to allow an substitutive assessment.

---

[105] Whereas 24 of the GDPR: " (...) potential subsequent use of personal data processing techniques which consist of profiling a natural person, particularly in order to take decisions concerning her or him or for analysing or predicting her or his personal preferences, behaviours and attitudes."

[106] Whereas 72 of the GDPR: "Profiling is subject to the rules of this Regulation governing the processing of personal data, such as the legal grounds for processing or data protection principles. The European Data Protection Board established by this Regulation (the 'Board') should be able to issue guidance in that context."

[107] The definitions of discrimination and bias presented in this guide are mainly based on the work done by Barocas & Selbst (2016, "Big data's disparate impact." California Law Review 104: 671.), Baeza-Yates (2018. Bias on the web. Communications of the ACM, 61(6), pp.54-61.), Salgado & Castillo (2018, "Differential status evaluations and racial bias in the Chilean segregated school system." Sociological Forum, 33, 2, pp. 354-377).), Cowgill et al. (Cowgill, B., Dell'Acqua, F., Deng, S., Hsu, D., Verma, N. and Chaintreau, A., 2020, "Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics." In Proceedings of the 21st ACM Conference on Economics and Computation, pp. 679-681), Sweeney (2013, "Discrimination in online ad delivery." arXiv preprint arXiv:1301.6822.) .

[108] Proxy variable. Oxford Reference. Available at https://www.oxfordreference.com/view/10.1093/oi/authority.20110803100351624

[109] Proxy variable. The SAGE Encyclopaedia of Social Science Research Methods, 2004. Available at: https://methods.sagepub.com/reference/the-sage-encyclopedia-of-social-science-research-methods/n768.xml

[110] Everything is a proxy. Machine Learning: Algorithms in the Real World Coursera [last consulted: November 2020]. Available at https://www.coursera.org/lecture/machine-learning-applied/everything-is-a-proxy-AFO5D