eticas

Guide to Algorithmic Auditing

January 2021

INDEX

FOREWORD

This Guide to Algorithmic Auditing has been developed and reviewed by a research team at Eticas Research and Consulting SL under the commission and supervision of the Spanish Data Protection Agency. The methodology proposed here has been developed based on specialized texts in this field and the experience of the audit team at Eticas Research and Consulting, with the collaboration of Dr. Carlos Castillo at Pompeu Fabra University.

SUMMARY

This Guide to Algorithmic Auditing offers guidelines and methodological principles for auditing products and services within the field of Artificial Intelligence (AI), especially those that make use of algorithms and collect or process personal data at some point during the process.

AI services based on using algorithms are spreading rapidly in both the public and private sectors. However, algorithms are often defined as "black boxes" of computer code and data, with results becoming increasingly unpredictable and uncontrollable. This raises numerous concerns about their ethical, social, legal, and even commercial impact. Respecting the fundamental rights to privacy and personal data protection is part of these concerns. The phrase "algorithm" encompasses various types of systems, depending on the data it handles, the type of internal operations it undertakes, and its performance objectives, among others.

This guide is not intended to provide an exhaustive technical definition of these types of technologies, nor create a specific audit methodology for each of them. Overall, its objective is to set forth a general methodology that acts as a roadmap for auditing diverse algorithmic applications. Thus, this Guide is specifically aimed at data controllers responsible for implementing algorithms, processing data, and conducting audits. At the same time, it is also intended to broaden the general public's knowledge, as they have become increasingly interested in understanding these issues.

KEYWORDS

Algorithmic Auditing, Algorithms, Artificial Intelligence, Machine Learning, Big Data, GDPR, Personal Data Protection, Attribution of Responsibility, Accountability, Legal Compliance, Ethics.

I. INTRODUCTION

The recent and rapid development of new technologies for processing big data, especially those that make use of algorithms and Artificial Intelligence (AI) techniques, have key social, economic, legal, and ethical implications. The rise of these new technologies, however, is taking place in a pre-regulatory framework, which does not help develop and implement them in an explainable, equitable, and ethical way, as would be ideal. If we understand that the efficiency of a new technology also depends on how and to what degree it serves people and social development as a whole, these drawbacks also reduce technological efficiency. In this scenario, the need to regulate the use of AI solutions and algorithms is clear. There are currently initiatives and proposals at the European-wide level and Spanish agencies and governing bodies working to establish guidelines in this regard[1], but it is necessary to strengthen this regulatory framework.

The use of algorithms is steadily increasing in both the public and private sectors, including the political, legislative, technological, financial, telecommunications, healthcare, manufacturing, transportation, energy and education sectors, to name a few. However, algorithms, especially machine learning algorithms, often become opaque sets of computer code and data, making it difficult for other people or entities to understand, predict or control what is going on inside them and what their implications will be. For this reason, the definition of algorithms as "black boxes" has become widespread. This implies that the use of algorithms can have an undesirable impact on individuals, groups of people or society as a whole, giving rise to potential risks often related to possible systemic biases and forms of discrimination, capable of affecting vulnerable individuals or social groups. These types of social impacts will be defined throughout the Guide. Moreover, the opacity of algorithmic systems calls into question the respect for privacy and personal data protection. As will be seen throughout this Guide, this implies analyzing algorithms within the social, economic and cultural context to which they belong, and

[1] For further information, we suggest consulting the following sources: *Artificial Intelligence for Europe*, *A European strategy for data*, *Commission Report on safety and liability implications of AI, the Internet of Things and Robotics*, *White Paper on Artificial Intelligence: a European approach to excellence and trust*, *Ethics Guidelines for Trustworthy AI*.

according to the perspective of the people they affect, directly or indirectly.

Under these circumstances, algorithmic audits are a necessary way to make this technology more explainable, transparent, predictable and controllable by citizens, public institutions and also companies, either before the development of the system, during its development or a posteriori. These audits also contribute to improving the mechanisms of attribution of responsibility and accountability of algorithmic systems. The methodology for auditing algorithms, however, is not yet simple or fully defined, which is a challenge.

In this complex context, Eticas Research and Consulting presents this Guide to Algorithmic Auditing, with three main objectives:

- The first, and most general, is to clarify the link between conducting algorithmic audits and safeguarding fundamental rights to privacy and personal data protection, as set forth in the Charter of Fundamental Rights of the European Union.
- The second is to provide clarity regarding the necessary regulatory framework for algorithmic systems, aiding in the correct interpretation and implementation of the General Data Protection Regulation (GDPR) of the European Parliament and of the Council, and its expansion where necessary.
- The third, which represents the primary concern of this Guide, is to offer guidelines and methodological principles for conducting algorithmic audits, thus allowing these technologies to be examined so that they are designed, developed and implemented in a legally acceptable way, but also in a predictable, suitable, desirable, sustainable and socially just and responsible manner.

This Guide to Algorithmic Auditing, thus falls within the framework suggested by the Spanish Data Protection Agency (AEPD)'s Guide to

Aligning AI Procedures with the GDPR, regarding effective compliance with the principles of personal data protection and how the correct approach and development of an algorithmic audit can contribute to this objective. This is imperative given that an algorithmic audit may in turn have an undesirable impact from a social, legal, political or commercial point of view, if performed inadequately. This is true because an audit may imply a reconfiguration or change in the implementation of an algorithm that is more harmful than the previous one. Algorithmic auditing also requires special attention in the collection and processing of personal and sensitive data involved in analyzing the algorithm.

1.1 WHAT ALGORITHMS SHOULD BE AUDITED

In this Guide, the word algorithm is defined from its simple and current origins in the field of computer science, which is the most widespread. From this perspective, an algorithm basically consists of a set of defined, non-ambiguous, ordered and finite instructions or rules that typically answer a question, make a decision, solve a problem, perform a computation, process data or carry out a task.

There are different types of algorithms, depending both on their operating mode and their objectives. Given this difficulty, this Guide proposes a general, replicable audit methodology, intended to act as a roadmap for others to apply to different specific cases. The main focus of this methodology is to detect, prevent and help correct potential undesirable consequences derived from the use of algorithms.

The audit methodology presented in this manual is expressly designed to analyze algorithms that may have a negative impact on individuals or social groups, especially those in more vulnerable situations. It is deemed especially important to audit algorithms that may affect access to education, work, services or social benefits, and/or that are implemented in judicial, public health or other public areas of social relevance and interest.

Any algorithm must be developed and implemented in such a way that it can be audited. However, algorithms thought to have a social impact pose a greater risk to the personal data protection and privacy and safety of individuals. Section III of this Guide proposes a definition of types of social impact, bias and discrimination that an algorithm may incur and must avoid. For example, while an algorithm used to sort materials on an assembly line is relevant from an operational or economic point of view, it is not of interest from a social impact perspective. In contrast, staff selection algorithms have various implications for workers' rights. As already noted, this can lead to rights violations such as gender discrimination.

On the other hand, for an algorithm of these characteristics to be audited with quality assurance, it must meet a series of minimum requirements detailed in the Methodology section of this manual. This is what we shall call an “auditable algorithm.”

According to current data protection regulations – GDPR and LOPDGDD (the Spanish acronym for the Organic Law on Data Protection and Guarantee of Digital Rights) – any automated processing that significantly impacts a person's life must always be supervised by a person. This implies a clear definition of roles of responsibility relating to the development and application of an algorithm, as well as the obligation to establish risk prevention and mitigation measures. To improve and strengthen compliance with these measures, this Guide recommends that any algorithm used in the public sector that meets the requirements detailed in this Guide should be subject to an algorithmic audit. Likewise, algorithms used in the private sector should progressively move towards undergoing this same process as part of their legal and social responsibilities.

1.2 WHO THIS GUIDE IS INTENDED FOR

This Guide to Algorithmic Auditing is intended primarily for those people responsible for the development and application of algorithms, as well as their auditing. Therefore it focuses primarily on those responsible for products and projects. However, it also aims to provide a structured framework of understanding for the sociological and technical teams involved in these processes, including: data protection officers, cybersecurity managers, ethical and legal compliance officers, technical staff and software development and data science teams. Finally, it also seeks to broaden the general public’s knowledge, as they have become increasingly interested in understanding these issues.

II. ALGORITHMIC AUDITING WITHIN THE CONTEXT OF DATA PROTECTION REGULATIONS

As mentioned in the introductory section, this Guide to Algorithmic Auditing focuses on developing an auditing methodology, particularly for those algorithms whose development or implementation may have a social impact that principally affects individuals' data protection and privacy rights.

This section has two main objectives. The first is to explain how the development and application of **algorithms with social impact can**

affect personal data protection. The second is to explain **the current regulations related to conducting algorithmic audits,** and indicate how these audits can aid effective compliance with these regulations. In the same vein, the mapping of current regulation will allow us to locate concepts used in the algorithmic auditing methodology within the framework of data protection.

Algorithms, especially those incorporating machine learning techniques, can handle and process massive amounts of data, including personal and sensitive data. However, as has been repeatedly pointed out, algorithms are often particularly complex and opaque in their design and behavior, making it difficult to know and control how such data is processed. At the same time, it has been shown that extensive data analysis can reveal information of a sensitive nature, which the data would not show in isolation. In addition, the purpose and usefulness of these systems is not always communicated in a clear and transparent way, even as algorithms are increasingly implemented to replace tasks previously performed by humans, including organization, prediction, recommendation, or decision-making support, among others.

The development of new techniques for the collection and processing of big data in recent decades has led to a strengthening of ethical and legal standards regarding privacy and personal data protection. However, they are still insufficient to fully comply with these rights, especially since they are not developing at the same pace as technological solutions. The rights to privacy and personal data protection are established as fundamental rights in various national and European Union regulations. Specifically, the Charter of Fundamental Rights of the European Union describes them in Articles 7 and 8 as follows:

- Article 7. Respect for private and family life:

"Everyone has the right to respect for his or her private and family life, home and communications."

- Article 8. Protection of personal data:

"Everyone has the right to the protection of personal data concerning him or her. Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified. Compliance with these rules shall be subject to control by an independent authority."

The special importance of rights related to the use of algorithms that process personal data, especially those that do so extensively, highlights the need to establish effective measures of control, correction, responsibility, accountability, and transparency regarding data processing. For this reason, this Guide offers **guidelines and methodological principles for conducting algorithmic audits**, which make it possible to analyze and identify points of tension that may imply a breach of data protection regulations. Such audits enable us to detect possible biases or bad practices in automatic data processing, with a view to correct them and include them as design requirements in the development and application of AI algorithms and solutions. This involves developing mechanisms to examine these technologies and help ensure that they are designed, developed, and implemented in a legally acceptable way, but also in a predictable, suitable, desirable, sustainable and socially just, and responsible manner.

Regarding compliance with legal regulations, which concerns us in this section, it should be noted that, since May 25, 2018, the GDPR is directly applicable to the member states of the European Union and is defined as: **Regulation 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[2] (hereinafter, GDPR). The Spanish transposition of this European Regulation has resulted in the formulation and enactment of Organic Law 3/2018, of December 5, on Data Protection and Guarantee of Digital Rights[3] (hereinafter LOPDGDD).

For its part, the **Organic Law on Data Protection and Guarantee of Digital Rights** (LOPDGDD) complements and specifies the provisions of the Regulation in the Spanish setting, reinforcing the importance of the data controller complying with the principles of data protection and heeding individuals' rights, while including certain provisions for specific processing operations, some of which may rely on solutions that make use of algorithms.

Both the **GDPR** and the **LOPDGDD** set forth guiding principles that any type of processing, including those based on Artificial Intelligence solutions and algorithms, must respect by defining a framework for responsible parties to undergo risk management of the rights and freedoms of data subjects and accountability, or the ability to demonstrate compliance with the requirements established by regulations.

This approach **requires data controllers and processors to address these requirements proactively**, including the case of automatic processing of personal data. This regulatory framework must be taken into account by organizations in order to avoid unnecessary duplication of responsibilities, encouraging contradictory requirements and supporting ambiguity and legal uncertainty in different sectors.

[2] General Data Protection Regulation can be viewed on this webpage: https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32016R0679.

[3] Organic Law 3/2018, of December 5, on Data Protection and Guarantee of Digital Rights can be viewed on this webpage: https://www.boe.es/eli/es/lo/2018/12/05/3.

III. ALGORITHMIC AUDITING METHODOLOGY

The **Algorithmic Auditing** methodology presented in this Guide proposes a framework for auditing algorithms defined as having **social impact**, in the context of the technologies they utilize. The audit focuses on analyzing and identifying aspects of the design, development and implementation of algorithms that may have a disadvantageous impact on disadvantaged groups and be non-compliant with data protection regulations, in order to correct them and include them as design requirements in developing AI solutions. Thus, this audit methodology seeks to ensure that these algorithms are designed, developed and used in a **suitable** manner from a legal point of view, but also that they are more **controllable, desirable, sustainable and socially just and responsible.** This implies that they undergo equal treatment of the social groups involved, are transparent and accessible to the public, and incorporate security mechanisms to prevent, identify and mitigate possible biases. Establishing a general framework for conducting these audits is essential, given that an improperly implemented algorithmic audit can also have undesirable consequences, if it does not propose adequate correction and improvement measures, or pay special attention to measures for the collection and processing of personal and sensitive data involved in the algorithm analysis.

An algorithmic audit is composed of a series of phases united under a single objective: **to identify, anticipate and correct** potential risks arising during the life cycle of the algorithm and the data processed. In turn, this makes it possible to strengthen the mechanisms of **responsibility and accountability and the protection of the rights and freedoms** of the natural persons involved (whether individuals or groups), especially the fundamental rights to privacy and personal data protection.

An algorithmic audit **can be internal or external**. However, an audit must always involve the collaboration of the internal staff member or institutional team (or client) implementing the algorithm and the team developing (or who developed) it. The external audit may be more objective, if performed by a reliable entity with certified experience and training, whose members undertake adequate information security measures and follow a consolidated methodology.[4]

3.1 GENERAL OBJECTIVES OF ALGORITHMIC AUDITING

The proposed methodology seeks to provide **quality assurance** for algorithms with social impact developed and implemented by public and private institutions, researchers, entrepreneurs and innovators. Likewise, it seeks to overcome shortcomings in the processes and measures of **responsibility** and **accountability** for actions derived from algorithmic operations. This implies establishing procedures for analyzing these systems involving, on the one hand, the pursuit of critical reflection and **awareness** of their possible impact and, on the other, the implementation of **transparency** mechanisms that make it possible to understand the steps involved in the design and development of the system.

The purpose of an audit is to **identify or anticipate errors, risks or threats** (actual or potential) and **help correct them**. This can occur at any stage of the system's development, both in its design and implementation, as well as in the operational phase and beyond. Therefore, auditing also makes it possible to outline a strategy for improving processes with algorithmic intervention in the future and to respond to flaws once the algorithm has been implemented. However, the importance of implementing auditing methods prior to the deployment and commissioning of these systems should be emphasized. The technology sector, including companies and public institutions, must get into the habit of auditing their algorithms, as a way of ensuring their

[4] To avoid unnecessarily complicating the methodological development of algorithmic auditing, the distinction between internal and external auditing will not be constantly referred to throughout this document.

social responsibility. As we shall see, this shares many of the same principles and results of data protection and privacy assessments.

Depending on **who performs this audit**, the specific objectives may vary. This means that an audit carried out with **research objectives** will generate fundamental and applied knowledge about the behavior of algorithmic systems and their effects, and report these findings to society. In the case of audits developed by **civil society organizations**, the objective may be to investigate systems that could affect the people they work with or advocate for. In the case of **consulting**, the audit may act to recommend improvements in the systems developed by public or private institutions, to prevent them from generating biases and forms of discrimination. As a final example, if an audit is performed by the **same institution that develops or implements the algorithm**, it will act as a self-assessment of risk and impact.

The type of assessment that an audit can carry out will depend on the development and implementation phase of the algorithm, or its **life cycle**. This means that during the early stages of the audit, **analysis of potential risks** can be performed, while in the later stages, measures for the **analysis of real impact** can be implemented.

The algorithmic audit methodology proposed here takes into account the importance of performing both a **technical analysis** – which allows us to evaluate the **effectiveness** of the system itself (in accordance with its set objectives) – and a **qualitative analysis**. This second part of the audit aims to assess the **desirability** and **acceptability** of an algorithm from a broader perspective, bearing in mind how it is **implemented**, how it is **integrated** into its social context, what previous systems it replaces (if any), what new dynamics it introduces, etc.

When an algorithm is audited, the objective is to **gain knowledge about the system itself and about the environment** (general and specific) to which this system belongs and operates within. This implies asking whether its behavior is adequate and relevant, whether it complies with current legislation, whether it is effective, whether it is replicable in similar contexts and whether it is robust. In addition, it implies asking whether it is transparent, whether it is explainable, whether it is useful, whether it is used appropriately and whether it is desirable from an ethical, social and cultural point of view. This should make it possible to know whether the algorithmic model may have been designed on an unbalanced or inappropriate basis, or whether its development or behavior may have harmful consequences on people. In this sense, it is also a matter of making the results more **predictable**, less uncertain and more **controllable** by the citizenry as a whole.

Conducting an algorithmic audit requires prior consideration of these factors in order to establish a working method, as well as follow the phases and steps to complete it in a suitable manner.

3.2 GUIDING PRINCIPLES OF AUDITING

The algorithmic audit methodology presented in this Guide is based on **four pillars or guiding principles**, which are not organized in a hierarchical manner, but are of equal importance and are **complementary**, and must be taken into account throughout the auditing process:

3.2.1 LEGAL AND ETHICAL COMPLIANCE

First of all, every algorithm must comply with **current legal and ethical standards**. In this regard, the auditing of an algorithm must consider the applicable legal framework and the rights and values involved. In the case of personal data protection, as previously explained, this would be the framework established by the General Data Protection Regulation of the European Parliament and of the Council, and Organic Law 3/2018, of December 5, on Data Protection and Guarantee of Digital Rights, as well as all legal texts and sectoral rules related to the specific scope of action applicable to an audited algorithm.

In addition to this, algorithms must comply with related ethical standards and codes, and must be designed, implemented and reviewed from an ethical perspective, respectful of social norms in terms of privacy, data protection, equality, social cohesion, freedom and trust. Finally, algorithms are expected to respect and promote respect for fundamental rights that may be impacted during its design and implementation, beyond the right to privacy and data protection (Arts. 7 and 8 ECHR). This includes rights such as the safety (Art. 3, ECHR) and freedom (Art. 5 ECHR) of people involved.

3.2.2 DESIRABILITY

The second relates to the **desirability of the system**. An algorithm with social impact must always be explainable, accurate, replicable, transparent and fair. For this reason, it is essential to pay attention to the "problem" that the audited system intends to provide a solution for, and examine whether the technology used is indeed the best way to address it. The perspective of political and cultural analysis is essential in order to make an adequate prognosis of the audited system from a technical and sociological point of view. This should help ensure that the solutions provided are as non-invasive as possible, while at the same time meeting the expectations and needs of the parties involved as efficiently as possible.

For an algorithm to be desirable, it implies that it does not discriminate against individuals or groups and especially does not have a detrimental impact on vulnerable individuals or groups in a distinct way from other individuals or groups by somehow reinforcing or influencing factors causing their vulnerability. Equally important is that the system is not biased. In this sense, those responsible for the design and implementation of an algorithm must consider factors that may have a general impact (within the scope of applying the algorithm) and a differential impact (among different population groups) on how these individuals use, understand and interpret the functions, characteristics and objectives of data processing.

3.2.3 ACCEPTABILITY

A third crucial aspect in evaluating an algorithmic system is its **social acceptability**. An audit of an algorithm with social impact must ask whether or not the audited system is acceptable from a social point of view, and in the eyes of society. A system that has an effect on people's lives, either directly or indirectly, must be understandable, controllable, sustainable and, to some extent, beneficial to the parties affected by it. For example, an algorithm that classifies the profiles of people applying for help from social services may be socially rejected or perceived inappropriately by the public. This could happen if their functions, objectives and expected results are not adequately and transparently communicated, or if they are not suitable or necessary in the eyes of the population. In this sense, Article 13 establishes the obligation for the data controller to inform of the existence of automated decisions, including profiling, and to provide the data subject with meaningful information on the processing logic applied, as well as the scope and expected consequences of such processing.

In this regard, the information provided about the algorithm must be clear and sufficient for citizens and customers to understand and assess the benefits and detriments it brings, as well as to participate in a meaningful way in its development and implementation, either directly or through public representatives or specialized professionals. Likewise, the acceptability of a given algorithm depends on whether it is well aligned with the public or private objectives explicitly communicated to users or stakeholders. In this sense, the design of the algorithm must pay special attention to those aspects that may conflict with the predominant values or cultural characteristics of the social environment in which it will be applied. An example would be a facial recognition system based on Caucasian feature data, which is not able to correctly identify people with Asian features. Not taking these elements into account can affect both the efficiency of the automatic system and the reputation of the organization in charge of its design and implementation, by causing possible discriminatory effects.

3.2.4 PROPER DATA PROTECTION AND DATA MANAGEMENT

Fourthly, but no less importantly, it is essential that an algorithmic audit attests to the **responsible and proper management of the data** involved throughout the algorithm's life cycle. This must comply with the aforementioned data processing principles established by the GDPR and the LOPDGDD, such as accuracy, limitation of the storage period, limitation of purpose, data security and confidentiality.

This implies that the data must be of good quality, up-to-date, from reliable sources, appropriate for the objective pursued by the system, and stored and processed using relevant techniques and for a clear and pre-established period of time. In any case, data must be able to be deleted and updated and must meet, if necessary, anonymization criteria adapted to the specificities of the case.

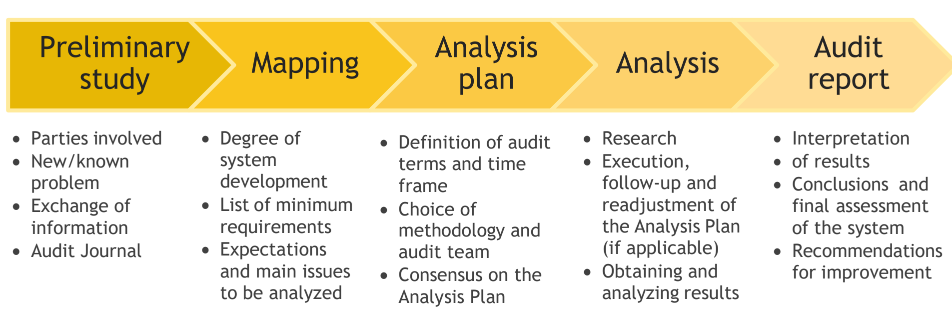
It should be kept in mind that good data quality and data management, including the documentation of all data processes that affect the training of an algorithm, are essential not only regarding its correct behavior but also the transparency for the people it concerns, both specifically and for society as a whole. Poor knowledge of the input or output data of an algorithm can turn it into a black box that is difficult to explain and audit.

3.3 AUDITING STAGES

An algorithmic audit is a **dynamic process**, which is defined in parallel to the development and functioning of the algorithm. For this reason, it should not be seen as an immutable set of steps, reproducible in the same way for each algorithmic audit, but rather **it should be adapted to the specific algorithm and the specific context** of each situation in which it is used.

However, it is possible to identify **a series of general stages that every audit should follow**, with defined objectives. It should be noted that these five stages do not have to be carried out strictly in the order presented below. Given the dynamism of the audit, as previously mentioned, the process is necessarily cyclical in nature, which requires completing the information detailed in each of the stages, so that feedback is generated, while respecting the order outlined here as much as possible. On the other hand, it should be noted that the methodological approach to the auditing process proposed in this Guide combines different techniques of quantitative and qualitative analysis.

This section explains the five stages of the auditing process proposed by the methodology presented in this Guide:



3.3.1 PRELIMINARY STUDY (STARTING POINT): WHO, WHAT AND HOW WAS IT DONE PREVIOUSLY?

The first step in auditing an algorithmic system is to understand **who** commissions, designs, develops, finances and implements it[5] and **what problem** this algorithm aims to solve. At this point, it can already be seen whether the implementation of this algorithm involves the collection or processing of personal data, in which case it would fall within the scope of the GDPR and the LOPDGDD.

In assessing the efficiency and appropriateness of the algorithm, it will be particularly useful to understand whether the algorithm designer/implementer is doing so to address a **"new" problem**[6] or if it is a **"known" problem**,[7] previously handled by a method that did not involve the use of an algorithm. This may imply a change in the way data is collected and/or processed, a variation in the data collected and/or processed, or for the data to be collected and/or processed when previously this was not done. In any case, a **risk analysis** will have to be made for this processing, as it will make use of an algorithmic system.

The answer to these questions poses two different scenarios for system analysis. In the first case (new problem), it will be relevant to understand **when and why** the decision was made or the **need to use an algorithm** was detected. In the second case (known problem), it will be

[5] See the roles of data controller and data processor established by Articles 4, 24, 26 and 28 of the GDPR, mentioned in section II of this Guide.

[6] To speak of a "new" problem does not imply that the problem did not exist or had not been detected before, but rather that the people or organization(s) designing, developing and implementing an algorithm had not addressed it before.

[7] A problem can be considered as known if this particular problem has been dealt with before, or if a significantly similar problem has been dealt with from objective observations. In other words, it is possible that previous algorithms have been used for the same purpose or that involve human protocols that seek to be reproduced by the algorithm.

a matter of verifying **since when and why** this problem is now addressed by the algorithm.

To this end, it is essential to establish a **fluid exchange of information** with the client and the algorithm development team, in order to resolve these doubts and any that arise in subsequent stages. This initial exchange of information can be more or less formal, depending on the circumstances, and it is advisable that it be carried out in compliance with the principles of **responsibility and accountability, through some means of a record,** preferably in writing. This is because this information will be vital for undertaking the rest of the process and will be useful for consulting again in later stages. At this point, it is recommended to sign a **confidentiality agreement** between the auditor and the auditee detailing the objectives of the data exchange, its means and requirements.

With the start of this phase and under the aim of improving the transparency, traceability and quality of the process, it is recommended to start an **Audit Journal** that collects relevant information on interactions and exchanges of information with the client, important decisions taken, problems detected, suggestions for improvement for the present or future, etc. In principle, this journal is conceived as an internal document, which may be updated by the audit team throughout the auditing process in order to collect essential information.

3.3.2 MAPPING THE SITUATION: HOW, WHEN, WHY AND FOR WHAT PURPOSE WAS THE ALGORITHM DEVELOPED AND IMPLEMENTED? DOES IT MEET MINIMUM REQUIREMENTS TO BE AUDITED?

This second stage is dedicated to gathering **basic information** about the algorithm and the context that it belongs to and impacts. It involves two main objectives: the first is to find out whether or not minimum requirements are met to **determine whether or not the algorithm can**

be audited with quality assurance. For this purpose, a list of requirements for conducting the audit is set forth in the following pages. The second is to **clarify the expectations of the analysis and identify the main issues to be analyzed** in the audit. In turn, this will enable the Analysis Plan to be drawn up as part of the audit's next stage (3).

One initial issue to address is the **degree of the algorithm's development**. That is, the algorithm to be audited may be a project that has not yet started or is in an early stage; it may be in design or development; it may already be designed, evaluated or trained; it may be in the operational phase (this may imply that it is interacting with the world); or it may be an algorithm that has already been used. Being clear about the algorithm's degree of development from the start is important, because depending on the algorithm's degree of development, the audit process will vary. Not all the same information is available at all stages, nor is the same type of correction, reworking or bias mitigation measures possible, for example. We will return to this issue later on.

One of the main objectives of the audit's second phase is to obtain basic information about the algorithm, in order to verify whether it is possible to audit. Thus, at this point it will be appropriate to frame the problem of the algorithm within the scope of the Records of Processing Activities (RPA) associated with the case, in compliance with Article 30 of the GDPR. Accordingly, each data controller and, where appropriate, processor of personal data, shall keep a record of the processing activities carried out under its control.

The following is a **list of requirements** that the audit client must meet in order for the algorithm to be audited with quality assurance[8].

---

[8] Like the full methodology presented in this Guide, this is an original list of requirements, created according to the items that must be included in the Records of Processing Activities (see the previous note of this document and Art. 30, GPDR), the auditing experience of the research team at Eticas Research and Consulting and previous academic texts, among which the work of Mitchell, *et al.* (2019) stands out.

3.3.2.1 *List of recommended requirements for an algorithm to be audited with quality assurance:*

- **Identifying and contact information** of the person(s) or institution(s) in charge of and responsible for different aspects related to the system's design, development and implementation and, if applicable, the representative of the responsible person(s) and the data protection officer;

- **Date of algorithm creation** and, if possible, the version of the algorithm[9].

- **Algorithm license.** This record should bear in mind whether the ownership of the algorithm is public or private and the contractual conditions existing between the developer and the person responsible for the algorithm's use. This may be an element that limits access to the algorithmic code.

- Data on the **basic architecture** of the algorithm, including data on the system's way of learning, training and operating.

- **Other reference details and specifications about the algorithm**, not reflected in the previous sections such as: articles or publications containing more information about the algorithm, citation data on the algorithm, or feedback data on its performance.

- **The theoretical framework** by which the model is developed.[10]

---

[9] If this is an algorithm developed from previous versions of the same algorithm, it will be useful to know how this version differs from previous versions.

[10] The audit team may consult this information before or after preparing the Analysis Plan, depending on how it thinks it may influence the audit's objectivity.

- **Methodological framework** and explanation of the **methodology** used to define the model (including underlying assumptions).

- **Access to and information about the algorithmic code:** This must comply with quality standards. This means information about the code, including information and clarifications about it necessary for intelligibility, such as:

    - the programming language(s);
    - explanatory notes;
    - programs, packages and libraries required for viewing it, etc.

- **Access to information about the algorithm's API** (application programming interface), if developed.

- **Access to information on the database(s)** used for developing the algorithm, and the databases used for its training (training database) and its testing or evaluation (testing database). In this regard, the client must provide information on at least the source(s) of the data collected in the databases and the motivations for why this data has been chosen, as well as the categories of data used (non-personal, personal, sensitive...).

Just like the code, the databases feeding the algorithm must respect quality standards that make them readable, understandable and usable. For this reason:

- databases must have an orderly and coherent structure among themselves;
- as much as possible, the data should be quality, accurate and up-to-date, i.e. contain as few invalid records as possible;

- the variables and the amount of data associated with them must be clearly identifiable and manageable;
- it must indicate whether anonymization or pseudonymization operations have been performed on the data;
- it is recommended that the databases be accompanied by a glossary for better understanding.

- Definition of the categories of involved parties affected by the implementation of the system and/or whose data is processed by it, including **groups involved** in the algorithm and the description of their identifying variables, especially those considered **vulnerable groups**, either by the developer and implementer of the algorithm or by the audit team. Where appropriate, to further expand this issue, it will also be recommended to identify **organizations of a social nature**, whose work focuses on improving the living conditions of these people or vulnerable groups.

- Information on **model training and evaluation**, including:
  - frequency and distribution of data and variables in the database(s);
  - information on the pre-processing of the data, its processing during model development and its post-processing;
  - parameters and criteria applied to achieve the impartiality of the model, or that act as the internal evaluation of its effectiveness;

- where possible, a general description of the technical and organizational security measures implemented in the algorithm.

- **Purposes or intended uses** of the algorithm: initial notions about who, how it is used and what it is used for, including:

  - main uses and purposes of using the algorithm;
  - primary users of the algorithm;
  - potential uses and secondary uses[11];
  - categories of recipients who were or will be given information about personal data processed by the algorithm, including recipients in third countries or international organizations;
  - where applicable, transfers of personal data to a third country or an international organization, including the identification of such, and in the case of transfers, documentation of their proper safeguards, as referred to in Article 49.1, second paragraph;
  - where possible, the periods foreseen for the deletion of different categories of data;

- **Objectives** of the algorithm's use: what are the aims of the algorithm's use in quantitative and qualitative terms. In the case of addressing a *new problem*, the client will be required to explain the motivations and arguments for addressing this problem. In the case of addressing a *known problem*, the client

[11] Note that this is a particularly sensitive issue when it comes to the protection of personal data, since using data for a different purpose than it was collected for reveals a malpractice that could go unnoticed. Thus, these uses should be communicated and founded on a legitimate basis.

should provide information on whether the objectives are the same as those pursued by the previous operating mode, or whether they have changed. Assessing the algorithm not only in terms of its use, but also in terms of its objectives, will allow a more accurate evaluation of the system.

- Information about the **dynamics, activities and processes** the system is integrated with. This includes details about the team working with the system, the organizational processes integrated with it, and the internal activities and dynamics it is a part of, or that are modified by its application. Like the objectives, it will be important to know whether these issues remain more or less unchanged from the previous operating mode (known problem).

- Information on the **responsibilities of the parties involved** regarding the model's behavior. This includes delving into the hierarchy of responsibilities of the system developers and the responsibilities of the system controllers regarding its application. This outline of roles is especially related to the concept of responsibility regarding data processing and the distribution of responsibilities of each data processor, as developed in Chapter 4 of the GDPR. Who is held responsible and accountable will depend on who designed, developed, commissioned and implemented the system. For example, the developers may be in charge of processing if the system is not developed within the same organization that implements it. Therefore, it is imperative to delineate responsibilities and roles within the algorithm's development and overall solution.

- Information on **determining factors** of the system's effectiveness, such as: the socioeconomic/environmental context, the instruments used to capture model input data, available resources, applicable policies and regulations, or other factors that may modify the system's performance. If the problem to be addressed by the algorithm is known, it will be valuable to know whether the circumstances accompanying the resolution of the problem are the same as before the implementation of the algorithm.

It should be noted that this is a non-exhaustive list of requirements with the minimum level of information that should be available to the audit team in order to evaluate an algorithm. It should also be considered that, depending on the degree of the algorithm's development or classification, it may be impossible to provide some of this information. In this case, it should be provided later and the client must commit to doing so when it becomes available.

If the **client is unable or unwilling to provide any of this information**, it will reduce the quality of the audit, jeopardizing both its completion and quality assurance. However, if this does not occur on a recurring basis for several requirements, or for any particularly important one, it will not necessarily mean that the audit cannot be conducted. It will be up to the audit team, based on their knowledge and experience, to assess the impact of the lack of these requirements on the quality of the audit and determine whether the audit should go ahead or not.

3.3.2.2 "Auditable" algorithm:

As indicated in previous sections of this Guide, every algorithm should be auditable. However, this guide focuses on those algorithms that may have a social impact, especially linked to a breach of the fundamental rights to personal data protection and privacy.

From the perspective of this Guide to Algorithmic Auditing, **an algorithm should be audited whenever** it collects or processes personal or sensitive data, may affect the lives of individuals and/or relevant social groups or vulnerable groups (especially if they impact issues such as access to education, work, services or social benefits, or operate in areas such as the law or public health), may engage with any types of social impact listed, or may involve some form of discrimination or bias at some stage in its life cycle.

In addition, an algorithm **may be audited provided that it complies with the minimum requirements** set forth in the list of requirements included in this section, according to the criteria of the audit team. In this sense, the reasons why an algorithm is or is not auditable have to do with a variety of issues, related to the algorithm itself, the context it belongs to, the people responsible for it, and administrative and legal issues.

At this point, it is also recommended to have **initial contact with all parties involved in the algorithm** development and implementation process and also with the parties affected by it (people and groups of interest, including social organizations as mentioned above). For this purpose, the dynamics of information exchange defined in the first step of the audit can be continued, or else interviews, focus groups or short surveys can be conducted.

3.3.3 ANALYSIS PLAN: HOW, WHEN AND FOR WHAT PURPOSE IS THE AUDIT CONDUCTED?

Once it has been verified that the system meets the minimum requirements to be audited, the next step is to define the audit analysis plan and get approval from the client. This mainly consists of identifying, defining and agreeing with the client on the audit's **object of study**, its **specific objectives, hypotheses and research questions,** the **methodology and techniques of analysis,** the **parameters of interpretation** of the results and **the tentative time frame** for the audit.

Similarly, based on the information available up to this point, at this stage of the analysis, **the formation of a suitable audit team should be defined** for the specific case. This team will be conditioned by factors such as the type of system used or the sector the model falls within. In the case of both internal and external audits, this team may include staff from both the auditing and auditee entities, though it is recommended that those working on the model's analysis be independent to ensure greater objectivity. The audit team should include **technical professionals** such as analysts capable of performing the technical part of the audit, especially data scientists; and social profiles, such as sociologists or legal experts, capable of bringing to light the deeper socio-economic, legal and ethical implications of the systems.

On the other hand, in order to correctly define the Analysis Plan, it is advisable to **review the theoretical and methodological framework** the client has used to developed the system, as well as outline the first notions about the **activities and processes** of the organization they belong to. It is also important to carry out a **study on the specific legal, social and economic context** in which the system is implemented, in order to better understand the acceptability and desirability of it as a whole. This information will be collected during the audit process and included in the Audit Report.

As part of preparing the Analysis Plan, it is up to the audit team to **define the planned methodology** for the audit, which as mentioned above, will vary depending on each specific case. This includes specifying, at least, the following aspects in agreement with the client:

- The **parts of the system** to be audited.
- The main **variables** or "mother variables" the scope of the audit analysis will consider.
- The **intersections** between variables to be studied (if applicable).
- The **groups** to be monitored and their defining variables.

- The **respective methods, metrics and quantitative analysis techniques** (statistics, surveys...). Examples are provided in the next section.

- The **methods and techniques of qualitative analysis** for the system (focus groups, interviews, participant/non-participant observation, ethnographic analysis, etc.) and the people or groups that will be approached to participate in the study.

- The **parameters of interpretation** for the results, which should be established in agreement with the client. These include:

  - sample percentages of different variables considered representative within the databases;

  - significant percentages and cut-off figures for the interpretation of the measurements made;

  - the minimum or maximum measures of system accuracy, desirability, and acceptability, if applicable.

- The **steps to be followed** in the audit.

- Estimated time frame for conducting the Analysis Plan.

- The **tentative schedule of follow-up meetings.**

Once the Analysis Plan has been defined by the audit team, which also details the deliverables and the work schedule, it must be shared with the client and agreed upon by both parties before proceeding with its implementation. In case of discrepancies, the terms of the analysis may be readjusted so that both parties are in agreement.

At this point, the audit team will be able to make a series of **preliminary recommendations** for the improvement of the system, according to what has been observed so far. It should be kept in mind that these recommendations will have a different scope depending on the degree of the algorithm's development. From the start, it may contribute to redesigning systems with an advanced degree of development, or simply suggest general considerations for implementation of systems in early stages. This is an important point for the audit, which as we have explained, is a cyclical process. If at this point, substantial problems or important issues to be addressed have already been detected, the audit may need to show how the client responds to these requirements in order to continue the process.

The audit team should also be in a position to **identify difficulties or obstacles** of continuing with the audit. In the event that the team believes it is not possible to continue, the audit may be temporarily stopped, as they wait to resolve any matters (with the consequent postponement of the following steps). Or the possibility of extending the study may be dismissed through a well-reasoned and argued report of the motives behind this decision and the results obtained so far. Regarding the observations made at this point, it may be necessary to rework or readjust the Analysis Plan, which will once again be agreed upon by the parties involved.

3.3.4 ANALYSIS: IMPLEMENTING THE ANALYSIS PLAN

This phase consists of **carrying out the Analysis Plan** defined and agreed upon with the client. It should be noted that, during the analysis process, it may be necessary to readjust aspects related to the audit's methodology, time frame and objectives, which will be agreed upon with the client. The study of the algorithm, as mentioned above, is composed of **two more or less distinct parts**, corresponding to the analysis of the system from **quantitative and qualitative** perspectives.

First, before proceeding with the planned analyses, the audit team should **conduct a status review** regarding the aspects detailed in the Analysis Plan to properly analyze the results obtained. In other words, at this point the audit team will conduct a review of the **theories**

underlying the creation of the model, the **reasoning behind important assumptions** in its development (for example, examining the arguments behind a causal relationship that an algorithm models, such as the selection of variables defining a phenomenon), and the **methodologies** used.

Likewise, the audit team should also conduct a study of **aspects related to the context** in which the algorithm is designed, developed and implemented, whether social, economic, organizational, environmental, technical, scientific or any other kind. This basically consists of knowing **the reality that the system interacts with as much as possible**, in order to analyze the potential implications in its real-world context.

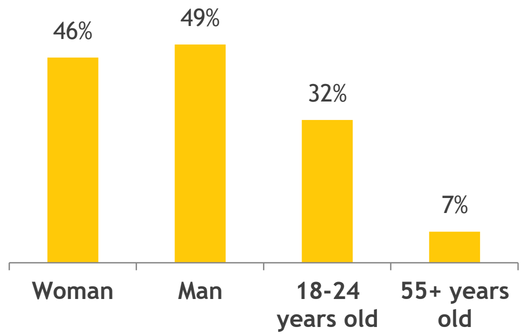
3.3.4.1 Technical auditing

Starting with the guidelines for the algorithm's **quantitative audit**, a **description of the database(s)** used by the client to develop, train and evaluate the system should be made. Moreover, the validity of the samples relating to the variables and groups relevant to the study will also be examined. To this end, an initial description of work will be carried out, defining the **identification, quantification and analysis of the frequency and distribution of the variables and intersections between variables and groups** relevant to studying the database (including protected groups). This will take into account the information provided by the audit client about which variables are considered most relevant for the model's development.

It should also be studied whether the system works with proxy variables, especially if these proxy variables are relevant to the algorithm. Proxy variables are those variables that are not of great interest when isolated, but may reveal important (or sensitive) information (through inferences) when analyzed together with other variables. For example: if the algorithm is aimed at predicting a potential rights violation, and a person is defined as being at risk of suffering a rights violation from the analysis of variables X, Y and Z, then "risk of a rights violation" would not be a variable explicitly collected in the database, but rather would be derived from the analysis of these three proxy variables (X, Y and Z). Analyzing the robustness of these relationships is essential to the audit, since these variables can modify the model in a decisive way.

Below are some basic graphs as examples, which represent the fictitious results of a frequency analysis of variables and the strongest positive and negative correlations between variables:

Ex. Frequency graph of sociodemographic variables[12]*:*



[12 and 20] These two simplified graphs represent the fictitious results of a descriptive analysis of the database used to train a resource allocation algorithm:

The first one shows the number of people in the database who are female, male, between 18 and 24 years old or over 55 years old. We see that the frequency of the 55+ age group is notably lower than the rest of the age groups.

The second graph shows that this model is more likely to assign a woman resource 8 and less likely to assign resource 3; more likely to assign an unemployed person resource 1 and less likely to assign resource 8; more likely to assign resource 4 to a person over 55 and less likely to assign resource 3; and more likely to assign resource 1 to an immigrant and less likely to assign resource 4. The graph also shows that the model does not show significant correlations for the allocation of certain resources to certain groups.

Ex. Graph of correlations between sociodemographic variables and the allocation of resources/support:[13]

Analysis of the results of these initial measurements will allow **initial conclusions** to be drawn about the audited algorithm, **preliminary recommendations** to be made based on the data analysis (if appropriate) and **research questions or educated hypotheses** to be raised about the system's performance.

To give an example: let's say one observes a very unequal distribution between variables, bogus correlations between them, or an inappropriate use of proxy variables. In this case, a research question can be posed or a hypothesis put forward, according to the relevant theory, that the model may have a disadvantageous behavior for a vulnerable group implicated in these variables. These hypotheses should be questioned throughout the study and be reflected in the Audit Report.

In this context, it will be assessed whether **the sample size of the analyzed variables is sufficient** in the database. To determine which variables and/or groups can be analyzed as robustly modeled by the system and which are not, it is recommended that they have at least more than 5% representativeness in the sample. Below this percentage,

[13] This graph shows that this model is more likely to assign a woman resource 8 and less likely to assign resource 3; more likely to assign an unemployed person resource 1 and less likely to assign resource 8; more likely to assign resource 4 to a person over 55 and less likely to assign resource 3; and more likely to assign resource 1 to an immigrant and less likely to assign resource 4. The graph also shows that the model does not show significant correlations for the allocation of certain resources to certain groups

the variable and/or group may have too little representation in the database. This should be pointed out in the study regardless, since it could be affecting the accuracy of the model. However, as indicated in the previous section, these percentages of interpretation should otherwise be agreed upon with the client during the development of the Analysis Plan.

In the event that any of the variables do not reach the minimum sample size agreed in the Analysis Plan, recommendations may be made in this regard, such as requesting a review of how information is collected related to these variables or the amounts of respective data, or the list of variables and/or groups to be studied may vary regarding what was agreed in the Analysis Plan (this difficulty and any corrective measures will be reflected in the Audit Report). At the same time, the **representativeness of a database** – for example, regarding the socio-demographic composition of the individuals or groups present in it – may be questioned in relation to the results after analyzing the frequency and distribution of variables: whether regarding a given population as a whole or within the database, or regarding a specific group, or another reference point. This point will also help verify that the system's data sources are reliable and sufficient, and that the data is being adequately managed at the quantitative level.

On the other hand, it should be studied **whether the distribution of variables is adequate**, or whether the system pays too much or too little attention to any of them. If we consider an algorithm used for resource allocation, for example, it will be noteworthy to study how these resources are allocated (how many, which ones, to whom, how?). Moreover, as part of this initial mapping, the logic behind the strongest correlations between variables will be assessed.

Once the descriptive analysis has been carried out, we will proceed to detect and study the algorithmic bias, implementing the **measurements agreed upon in the Analysis Plan.**

It is important to note that this bias analysis methodology not only studies the impact of the system on key protected groups (these are noted in section III of this Guide), but also focuses on **dynamically detecting and analyzing those vulnerable groups that may be discriminated against by the system**, given the case and the context to which the system belongs. It should be noted that a vulnerable group can be defined by intersections between variables (such as: retired, non-white and low-income women, or other combinations of personal and temporal variables that impact relevant communities). Therefore, the definition of vulnerable groups potentially affected by an algorithm must be made in accordance with the reality it belongs to. To do this, the patterns of vulnerability and exclusion in a given case must be identified by analyzing the underlying theories and assumptions, as well as the variables, variable intersections, proxy variable combinations and functions used by the system. This case illustrates why an algorithmic audit should not only consist of a quantitative analysis, but also a qualitative analysis, capable of understanding the system within its implementation framework.

From a quantitative perspective, the **methodology of bias analysis** of an algorithmic system is divided into **four main steps**[14]:

i. ***Assignment of data to groups***

The first step is to define the **assignment of data to specific groups**, based on the mapping work of the previously developed algorithm. This means that data relating to particular features or attributes are classified into groups, which may be overlapping ("soft" mapping) or non-overlapping ("hard" mapping). Overlapping refers to the convergence of more than one protected characteristic, such as

[14] This framework is derived from the methodology applied by Carlos Castillo, researcher of the Department of Information and Communication Technologies at Pompeu Fabra University, in previous works carried out in collaboration with Eticas Research and Consulting.

"low-income woman." In most cases, groups will be made according to unique characteristics. Any characteristic assigned to multiple individuals can be used to create such groups, but special attention is paid to protected attributes. These groupings are created to evaluate the extent to which an algorithm may treat or impact one group differently from another.

ii. Identification of vulnerable groups

The second step is to determine **which of the groups that have been defined are considered vulnerable or protected groups within the specific context of the audit**. This means that they should not be disadvantaged by the algorithm's use and, therefore, its impact on them will be specially monitored. A narrow definition of a protected group could be based on the purpose of a technology and thus the appropriateness of the algorithm. For example, if the intent of a certain algorithm is to increase the protection of children of a certain age who suffer domestic abuse, then children of that age constitute a protected group.

iii. Definition of analysis criteria and metrics

The third step **determines the set of metrics** to be used for the analysis of these protected groups. The objective is to analyze whether the **algorithm behaves appropriately regarding the different groups** identified, based on specific criteria of "algorithmic equity." There are multiple definitions of algorithmic equity.[15] Among the most commonly accepted is a definition linked to **group equity**,[16] which means that an algorithm **should not produce disadvantageous results for specific or vulnerable groups**.

[15] For more information, see the work of Binns *et al.* (2018), Castillo (2019), Chouldechova (2017), Dwork and Ilvento (2018), Dwork *et al.* (2012), Holstein (2019), Kim *et al*. (2018), Kleinberg *et al.* (2017), Kyung Lee (2018) and Nayanan (2018), listed in the References section of this Guide.
[16] For more information, see the work of Barocas and Hardt (2017), listed in the References section of this Guide.

In general, group equity is often understood to exist if one or more of the following conditions are met:

- the probability that an algorithm generates a result is not determined by the attribute that defines a specific group (independence);
- this is true even if real data accompany the assignment of a result to a certain group (separation);
- and the measurement performed by an algorithm is not combined with protected attributes to obtain a result (sufficiency).

However, these conditions cannot be met in certain cases, which makes it necessary to link the results to the presence of explicitly protected attributes in order to fulfill the desired objectives. On the other hand, though these definitions of algorithmic equity focus on groups and do not guarantee that an algorithm behaves fairly with different individuals, academic literature on the subject indicates that it is complex to develop consistent mechanisms to measure **unequal treatment at the individual level**. This is a form of measurement, which some authors believe could undermine measures of group equity by ignoring **broader contextual factors**.[17]

For this reason, among others, the **contextual framework in which an algorithm operates must be analyzed**, both from a quantitative and qualitative point of view, and used to interpret its results in terms of algorithmic equity. This is especially important in cases where an algorithm is used to sort items such as people, groups of people or similar categories. In this case, it is recommended: on the one hand, that there is a **sufficient presence of defining elements of the protected group**, to be able to monitor that the algorithm does not sustain forms of

[17] For more information, see the work of Heidari *et al.* (2018) and Speicher *et al*. (2018), listed in the References section of this Guide.

discrimination and differential treatment at the group level; and on the other hand, that the elements related to groups **are treated consistently**, to avoid forms of individual discrimination, i.e. for potential differences in the treatment of individuals to be determined solely by their non-protected attributes (Castillo, 2019).

As indicated above, there are multiple definitions of possible metrics for assessing algorithmic bias, and their choice will depend on issues related to the way the algorithm works, its objectives, the type of information it handles, among other things. However, a certain degree of consistency must be maintained. The metrics outlined here are based on the assumption that the algorithm **can have a positive or favorable result, or a negative or unfavorable one**, or that it is possible to **order these results on a scale from the most positive to the most negative**, or vice versa (e.g. an algorithm ranking job applicants).

To assess whether a system effectively treats different affected groups equitably, it is advisable as a general process – to be applied to different cases – to study whether the system analysis reports **differential impact or treatment rates** and whether there are significant differences between **false positive rates (FPR)** and **false negative rates (FNR)** among different groups. This consists first of quantifying the extent to which an algorithm has a different impact on different individuals or groups and the extent to which it treats individuals or groups of individuals differently. Secondly, the objective is to identify whether there are unfavorable differences for the protected group between the rates of false positives, false negatives, true positives or true negatives assigned to this group compared to another group. In other words, it aims to examine whether a system overestimates or underestimates a certain group, in a relevant way with respect to another group and to the system's objectives. In general, these metrics quantify the extent to which an algorithm treats people differently (disparate treatment, DT) and the extent to which an algorithm has a different impact on different people (disparate impact, DI).

For this bias assessment, it is recommended to use standard tools such as Aequitas Bias and Fairness Audit Toolkit[18], AI Fairness 360 Open Source Toolkit[19] or Algorithmic Equity Toolkit[20], among others.

iv. *Application of metrics to group analysis*

The fourth step consists of **applying the chosen metrics relevant to the specific case and analyzing their results for the selected groups.** If data is processed in several stages in a system (such as data collection and data analysis), the analysis of these metrics is carried out for each stage or step separately. As an example, some possible analysis metrics and dummy values are listed below:

- **Impact ratio**, this ratio is calculated as the percentage of the protected group with positive prediction/outcome divided by the percentage of the unprotected group with positive prediction/outcome. Typically, values below 80% are considered problematic and should be further checked to see if such disparity is due to a case of algorithmic discrimination. Values close to 100% are considered more equitable.

- **False positive and false negative rates.** A false positive is a positive prediction in reality that turns out to be negative in the algorithmic results. Conversely, a false negative is a prediction classified as negative that turns out to be positive in the actual case. A group is considered to have underestimated risk by the algorithm if the false negative rate is greater than the false positive rate (the latter being greater than 0). On the other hand, disparity between groups is usually

[18] For more information, see Aequitas' webpage: http://www.datasciencepublicpolicy.org/projects/aequitas/.
[19] For more information, see AI Fairness 360's webpage: https://aif360.mybluemix.net/.
[20] For more information, see Algorithmic Equity Toolkit's webpage: https://aekit.pubpub.org/.

considered to exist if the false negative rates assigned to compared groups have a substantial difference.

Let's say we are analyzing an algorithm that predicts the risk of a population suffering poverty in order to more efficiently prioritize social resources and allocate them to those at high risk. We analyze differential treatment by groups and have found that the algorithm yields these results:

Group	Population	Prediction of high risk	Prediction percentage
A	80	48	= 48/80 = 60%
B	40	12	= 12/40 = 30%

In this table, we can see that the algorithm assigns a high risk more frequently to group A. If we calculate the impact ratio as 30%/60%, we see that its value is 50%, which is lower than the reference value of 80%. This implies that if group A is the least unprotected of the groups compared, the disparity observed in the treatment would not imply discrimination in theory.

Group	False negative rate	False positive rate
A	0.55	0.14
B	0.72	0.12

Here we see that group B is more likely to have a substantially higher false negative rate than group A (0.55 / 0.72 = 76%). This means that someone in group B is more likely to be misclassified as low risk. These two values give us indications of the negative differential impact on group B. If group B is the most unprotected, then these results should be analyzed in their social and operational context to determine if there is any bias or discrimination.

Once again, it is important to remember that these interpretation parameters will have to be agreed upon with the client during the Analysis Plan development stage. **The interpretation of the results of these measures will always depend on the specific case.** For example: it is not the same for a vulnerable group to have 30% more FNR than for a privileged group. In the first case, the model would be generating a disadvantage that could be discriminatory towards the protected group. In the second case, it can be considered a form of positive and even necessary discrimination.

On the other hand, there may be cases where a high difference in the false positive and negative rates, which discriminates against a protected group, can be justified by the functioning of a specific system. For example, this happens when the cut-off value for the risk assignment of vulnerable groups with a high sample size in relation to a phenomenon is intentionally set high in the model design. This could be done to reduce the presence of this group in the risk allocation. A case example could be an algorithm aimed at predicting the risk of recidivism for the prison population in some states of the United States, where the at-risk population is predominantly African-American. However, an ethical and socially desirable analysis should be carried out of those systems as defined by these terms.

In addition, it will be possible to evaluate how the **system responds to new/different input data** (some data may be exchanged with others) and **commands** imposed by the audit team. For example, it is possible to assess the accuracy of the estimated result of an algorithm for an individual or a group, based on the analysis of the profiles of the people who make up the training databases, and by considering how it has behaved with other individuals or groups of comparable characteristics.

This would indicate to the client the extent to which the algorithm can be trusted when applied to a particular case.

3.3.4.2 Qualitative auditing

In parallel, the **qualitative part** of the analysis, **equally necessary for algorithmic validation**, will be developed. Since, as already indicated, the audit is a cyclical process, this part of the qualitative analysis is fed back and provides essential information for the quantitative analysis. This consists of holistically collecting, analyzing and integrating all the necessary information into the analysis and interpretation of the results. This information can be collected and analyzed through **a review of academic literature** on the subject, or other documents of interest. Moreover, this can be done through the **exchange of information** with the different parties involved in the design, development and implementation of the algorithm, and the parties directly and indirectly affected by it, or by conducting and interpreting the results of **interviews**, in-depth interviews, **surveys, focus groups,** participant or non-participant **observation, ethnographic studies, expert panels,** etc.

The qualitative analysis of an algorithmic system focuses primarily on examining whether the **principles of ethical and legal compliance, acceptability, desirability and protection of personal data are met** in the specific context of the system. To this end, it studies the **objectives and uses** of the algorithm, the **protection or lack thereof** for individuals and groups affected by it, as well as **compliance with applicable political, social, legal and ethical standards,** and its **integration into broader dynamics**. As explained above, depending on the type of system, this involves (re)analyzing the sociodemographic composition of the algorithm's target group in its social framework, examining (or re-examining) the theoretical literature on the phenomenon or variable to be measured, and studying the composition of the sample used to train the algorithm. For example, in the case of an algorithm designed to predict the risk of homelessness in a particular city, it will be necessary to collect and analyze real data and theoretical literature on those variables that best reflect the likelihood of people living on the street in this city, the number of people who are homeless and their relevant membership groups, among others.

It should be noted that an essential aspect of the qualitative evaluation of a given algorithm is to understand **how and who is affected by its creation and use**. For this reason, it is especially advisable to gather information from individuals, groups or organizations affected by it, and understand their levels of satisfaction and attitudes regarding the use of this technique in relation to a given problem. This study will allow the audit team to propose improvements to the system, based on a more complete understanding of its social impact.

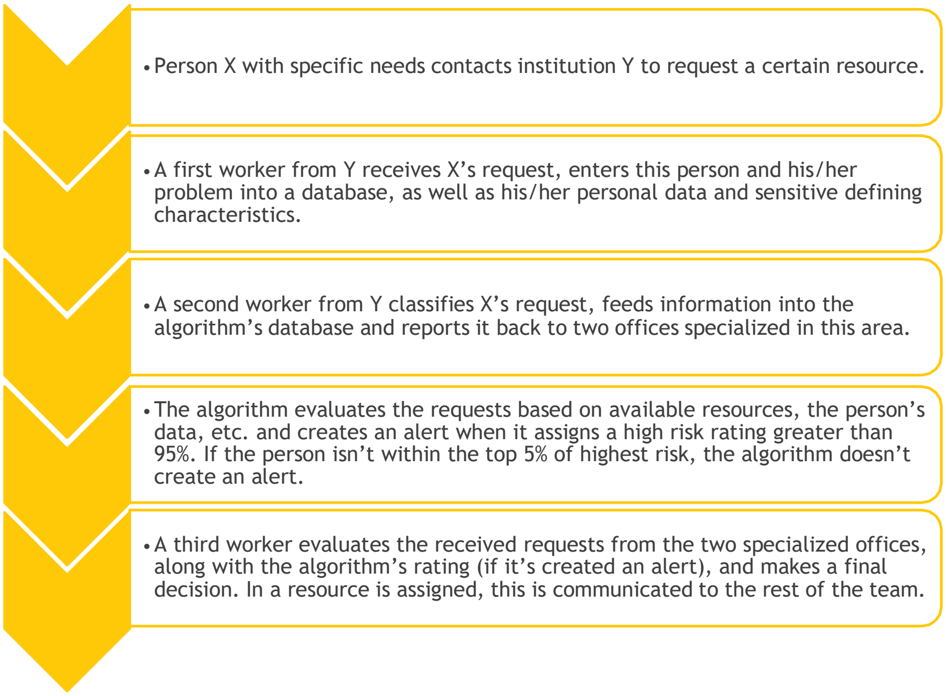
Likewise, the qualitative analysis of an algorithmic system includes examining **what role it will play in the processes** it is a part of, and **also analyzing the profile, training or satisfaction of the team** that interacts with it. This implies answering some relevant questions, including: What are the processes the algorithm forms part of? Does the algorithm have an adequate role in these processes? Have the routines and dynamics of the organization(s) that use the algorithm changed, or are they maintained regarding the situation before the algorithm? What are these routines and dynamics today? Who are the teams and the professionals that interact with the algorithm? Are they sufficiently educated and trained to use the algorithm in an appropriate manner? To this end, information will be gathered on the **roles and professional profiles** of the team members, their **responsibilities** in relation to the operation of the system, the **training** provided to this team, as well as other aspects, including: whether their level of **confidence** in the system is high,[21]

[21] In accordance with Article 22 of the GDPR, the audit must reflect whether, in the case of an algorithm integrated into a decision-making process, it allows human intervention and continues to give precedence to the professional judgment of those specialized in the area,

whether all workers interacting with the algorithm use it in a **unified** manner, or if on the contrary they apply results in a disparate manner, or have data on internal/external **satisfaction** with the system. It should also be asked whether the **privacy conditions, data protection principles and security measures** are being complied with in a proper way and in accordance with the legal regulations and ethical codes in force.

An example diagram reflecting the basic steps of using a fictitious resource allocation algorithm is shown below:

Ex.: Diagram of the process of using a resource allocation algorithm[22]*:*



over the result provided by the algorithm. It is also recommended to analyze the weight of the algorithm's result in the final decision.

[22] In this example case, we see that the applicant will be subject to a partially automated decision by the resource allocation algorithm, thus he/she has the right to know (among other aspects) the process and the reasons why the decision is made to allocate a particular resource or not and, if necessary, to appeal to have his/her case fully evaluated by a human.

The audit also aims to combat **algorithmic opacity** by suggesting **transparency** measures that help explain the algorithm's features, weaknesses and strengths, and results. For example, this includes making public which variables, intersections or proxy variables are the most determinant to the system, or whose variation would have most affected its results. **Responsibilities and accountability measures** linked to the results of the system's design, development and implementation should also be adequately communicated.[23]

The results of the qualitative and quantitative analyses will be reflected and interpreted in the Audit Report, in accordance with the parameters defined in the Analysis Plan.

3.3.5 REPORT: EXPLANATION, INTERPRETATION OF RESULTS, RECOMMENDATIONS AND CONCLUSIONS AUDIT OF THE AUDIT.

After conducting the analysis, an **Audit Report** will be prepared, which will provide a record of the process performed, as well as **the legal and ethical compliance, accuracy, acceptability and desirability** of the model based on the interpretation of the results. As indicated in previous sections, the parameters of interpretation for the results will be agreed upon with the client during the development stage of the Analysis Plan. This interpretation can be made at three levels:

- regarding the database used for its development, training and implementation,

- regarding the objectives for the creation and use of the algorithm,

[23] The recommendations section will present practices and measures recommended for the mitigation of biases, the redesigning or improvement of the system, and also for the improvement of its qualitative aspects, such as the correct use of the system, or the implementation of accountability and transparency measures.

- regarding the real context to which it belongs.

Likewise, the **adherence of the results to the initial Analysis Plan** and the interpretation of the analysis results **according to the established objectives, hypotheses and research questions** will be assessed. Final **recommendations** and possible **error mitigation measures** for the improvement of the algorithm development or its implementation, or for future system redesigns will also be provided.

The result of the audit should show in a clear and easily understandable way **the system's level of risk**, preferably in relation to each of the variables or groups predominantly observed in the analysis. An example of a risk assessment table is reproduced in the Appendix of this Guide (Appendix 3).

This assessment must be clearly documented regarding the results of the metrics applied in the quantitative and qualitative analyses performed.

The Audit Report should have a length appropriate to the complexity, timing and contents of the audit analysis, and should at least include information on:

- the title of the project and the name of the audited system;
- the date of the audit report[24] and the name of the authors of the report/study, if applicable;
- the responsibility of the audit team in relation to the quality of the system;
- the explanation and contextualization of the specific case study, including all relevant information about the audited algorithm, collected as part of the list of initial requirements, but also on the social,

[24] The date of the report should be when the audit procedures necessary to form an opinion on the system's level of risk have been conducted.

economic, organizational, legal, ethical or technological framework to which the system belongs;

- the methodology and steps of the algorithm analysis process, including information on the terms and time frames of the audit, agreed upon with the client in the Analysis Plan;

- the results of the qualitative and quantitative analysis performed during the audit, organized and represented in a visual and orderly manner;

- a reasoned and argued explanation of the interpretation of the results, including the assessment of the system (by parts);

- the general and specific conclusions drawn from the interpretation of results, including positive and negative aspects of the audited algorithm;

- a list of recommended practices and measures for system improvement, created in relation to the algorithm's specific case, which are operational, clear and implementable;

- a list of references used in the preparation of the report;

- an appendix section (if applicable).

The Appendix section of this Guide (Appendix 2) includes, by way of example, a more developed sample table for the preparation of an audit report. This example will give the reader a better idea of the contents that should be included in an audit report.

IV. RECOMMENDATIONS FOR SYSTEM IMPROVEMENT AFTER AN AUDIT HAS BEEN CONDUCTED

When using an algorithm that handles personal or sensitive data, or that may have an impact on the life of a person or a group of people, it is advisable to conduct an algorithmic audit.

An algorithmic audit should point out the positive and negative aspects of the audited system and, especially in the case of the negative aspects, **provide recommendations that will enable the client or organization to improve the algorithm or its implementation.** Like the rest of the audit process, specific recommendations for the improvement of an algorithm **will depend on the specific case and the exact results** after analyzing the accuracy, desirability or acceptability of the system.

In addition to **identifying possible non-compliance with regulations that need to be rectified,** this type of audit makes it possible to identify aspects that can be improved and optimized to make the algorithm **more explainable, transparent, predictable and controllable.** Its practice is recommended for those responsible for algorithms with social impact, whether they are public bodies or private entities, in which case it will also contribute to promoting corporate social responsibility.

In this section, some **examples of specific recommendations** that might be presented after conducting a system improvement audit are put forward to help the reader of this Guide understand what this issue entails.[25] It is key to bear in mind that the recommendations made will always be determined by the system's **degree of development**. Here are several recommendations that could correspond to different phases, **divided into sections** that they define below. These include specific advice aimed at ensuring that the data processing performed by the algorithm complies with **data protection laws and principles**. They also underline the **importance of implementing and reinforcing transparency mechanisms** through the supervision of the algorithm's

[25] Once again, these recommendations are based on prior experience of the audit team at Eticas Research and Consulting and Pompeu Fabra University.

operations, ensuring compliance with certain obligations **by the data controller** and guaranteeing data subjects their rights.

4.1 RECOMMENDATIONS FOR DATA MANAGEMENT AND ALGORITHM ACCURACY

4.1.1.1 Regarding the theoretical/methodological basis of the system

- In the event that **inaccuracies** are detected in the basic assumptions underlying an algorithm, it will be recommended to revise them based on relevant theory and data available.

- Likewise, it will be recommended to strengthen **the review of academic literature** on those aspects, variables and contexts affected by the system, if these are considered insufficient or inadequate.

- The same recommendation applies for the **methodological basis** of algorithm creation, in case these are not considered suitable, such as how to collect system training data.

4.1.1.2 Regarding the database

- Review the veracity, reliability and updating of the original **source** of the data.

- Examine the **representativeness** of the sample of a variable, intersection or a group of variables that define an analysis group, regarding given parameters or its reality.

- **Minimize** the collection of data generally and specifically that is not necessary for the algorithm's purpose or whose collection may stigmatize specific individuals or groups.

- **Imbalances** between the amount of data that the system collects on a given variable compared to another could lead to deviations in the system. The recommendation to minimize or expand the amount of data should incorporate an accurate trade-off analysis, establishing the relationship between the amount and type of data to be collected/discarded and those necessary to guarantee the effectiveness and efficiency of the system in question.

- If categories of data or **variables that are necessary for the correct modeling** of the algorithm have not been collected in the system's training or testing database, it may be recommended to include them. In some cases, failure to collect certain variables during the system training process may mean that the algorithm does not identify or "learn" them and cannot use them in the future.

- The case referred to in the previous point may occur in systems that need to collect information on sensitive attributes to **perform their function**, or **assess that the system is accurate concerning those attributes** (for example: to control that an algorithm does not discriminate on the basis of gender, there must be information on the gender of the people in the database).

- Modify the **format of the input data**, if it is not satisfactory because it does not represent the reality it reflects, or how the system works (for example: if we consider the case of a natural language processing algorithm and the way the algorithm works, it does not have the capacity to adapt to changes in the words that make up the input texts. It is likely that the algorithm will not behave in the desired way if the input texts are not schematically organized. In this case, a more orderly form of data input may be chosen, or the system behavior may be adapted to the format of the input data).

- Change **the way data is collected**. It is possible that, in some case the way in which the algorithm collects data is not adequate, and it is advisable to apply filters in the collection, expansion or restriction of the data collected.

- **Clean or restructure the database** and the classification of variables into clearly distinguishable and identifiable types.

- **Clean or restructure the database** glossary, if it is not readable or effective in understanding the database.

- Review whether the **distribution or frequency of variables** collected in the database is insufficient, since this may lead to system imbalances.

4.1.1.3 Regarding the management of data and variables

- In the event that the database contains information on identifying attributes of vulnerable groups, it may be recommended that this information **not be included in system modeling** (but only to assess its accuracy), or that its behavior regarding these variables be monitored over time.

- Study the cases of **variables with very low rates in the sample, which are not considered robustly modeled** and raise the alarm when the system detects them. This refers to the case of variables with no or few cases collected in the database. It also refers to those variables that would give rise to a different model if their presence in the database experienced a small upward or downward variation.

- In the event that an algorithm is **not accurate or discriminates against specific social groups** because of their association with a given attribute, it will be recommended that this behavior be reviewed or the system be redesigned to correct it. In this sense, it is possible to recommend integrating one or more variables not initially considered into the model, but detected as discriminatory during the analysis through proxies or other methods. The purpose of this is to have greater control over the variable and for the algorithm to correctly identify it as a measurement value.

- It is also recommended to conduct an effectiveness analysis of these matters by comparing **subgroups affected by the issue.**

- This may also occur regarding certain **intersections** between variables, in which case the review and possible reconfiguration of the system's behavior towards them will also be recommended.

- It is recommended to pay special attention to the **amount of information** collected in the database about those attributes/variables that the system under/overestimates, or the **rules** by which this situation may occur (in certain cases it may be intentional).

4.1.1.4 Regarding algorithm performance

- A relevant question may be to review the model's **level of statism or variability** regarding the type of input data it handles, the data collection structure, the environment it interacts with, the way the system learns, etc. This includes assessing whether or not the system can and should adapt to new data or new types of input data, whether it can draw valid conclusions from the information format it handles, whether it can learn new relationships between input and output data, etc.

- In the event that the evaluation indicated in the previous point is negative, it may be recommended to **change the way the system learns.** In other words, move from a more supervised to a less supervised mode, or vice versa.

- An important issue, though complicated to predict, is the **future behavior of an algorithm**. This will depend to a large extent on the data that algorithm interacts with, the feedback ecosystem generated by this data, and other factors relating to the context, which are changeable. In this case, it is recommended to **monitor the system's behavior over time towards factors whose variation may affect its future behavior.** For example, this variation may be related to those variables whose representation in the system's training databases is too small or too large, without corresponding to the social reality, or understanding that this social reality may change.

- It is also recommended to monitor how changes in the **social, economic, organizational, environmental, etc. context** over time may affect the system's development and performance. These changes may have impacts on the target variables of the model algorithm, altering its efficiency. For example, let's say the female homeless population increases abruptly in a given population evaluated by a risk allocation algorithm. If the algorithm is not able to capture or learn this social transformation, it could duly underestimate the risk of women and limit public resources allocated to them.

- Finally, **periodic audits** are recommended, which are not restricted to a particular moment in the development and implementation of the algorithm, but allow its evolution over time to be evaluated. Annual evaluation of a system is usually sufficient.

4.2 RECOMMENDATIONS FOR ETHICAL AND LEGAL COMPLIANCE

- Generally speaking, **compliance with the fundamental rights to privacy and personal data protection** must be respected and, as far as possible, promoted, both in the processes of designing, developing and implementing an algorithm as well as during the audit process. This should also be true for all rights that may be affected by the specific case of an algorithm.

- The development and implementation of an algorithm, and any algorithmic audit that is conducted must pay special attention to those **aspects of an algorithm that may not comply with the provisions of the GDPR, the LOPDGDD** or other sectoral norms or national or international regulations.

- In particular, respect for the **principles of data processing**, contained in both the GDPR and the LOPDGDDD, must be promoted.

- It is recommended that the development and implementation of algorithms be carried out in accordance with the provisions of the **ethical and deontological codes** of the sector it belongs to.

- It is also recommended to apply the notions contained in the **guides and manuals of good practices** issued by competent authorities.

- As part of the development and implementation of an algorithm, **measures** must be established **to facilitate the exercise of people's rights.**

- Any method of data processing, whether conventional or through the use of an algorithm, must be reflected as an activity in the **Records of Processing Activities**. It is thus recommended that algorithmic audits verify that the processing carried out by the algorithm is adequately recorded in the RPA and collects all the information set forth in Article 30 of the GDPR.

- The collection and processing of personal data and sensitive data should be especially analyzed regarding these issues, especially when it concerns **vulnerable groups.**

4.3 RECOMMENDATIONS FOR GREATER ACCEPTABILITY AND DESIRABILITY

4.3.1.1 Regarding the system's use

- During the development of the algorithm, and prior to its implementation, it is recommended that a **review and case study** be conducted by the teams that will use the algorithm, or apply its results, so that they can **provide feedback and suggest changes or improvements based on their experience.**

- It is recommended to ensure the necessary **training and experience** of workers who interact with the model (directly and indirectly). This will help to ensure that the level of trust in the human teams is adequate, i.e. not too much, not too little. In this way, it will be easier to ensure an appropriate balance between **professional judgment and the results of an algorithm.**

- It is also advisable to carry out **continuous training** that allows workers to replace their previous practices and protocols with the algorithm's new interaction dynamics, and to internalize important aspects of the technical performance of the system, its scope and limitations.

- In addition to training activities, it is recommended that **satisfaction data** be collected, both from **workers** who interact directly with the algorithm as well as **those who are affected by its results**, if possible.

- It is especially advisable to collect data on the **satisfaction and attitudes of stakeholders**, since the development and use of the algorithm has an impact on them, especially if these are vulnerable individuals or groups, and to count on their collaboration during the system's auditing process, especially when recommending improvements.

- In the event that the algorithm affects **individuals or vulnerable groups**, it is also advisable to collect satisfaction data from **social organizations** or other types of institutions that work with these people. Also, as in the previous case, seek their collaboration during the system's auditing process, especially when recommending improvements.

- In those cases where the situation permits, it will be of particular relevance to **compare satisfaction data with the system used prior to the algorithm** to provide an answer to the same or a similar problem.

This would make it possible to assess crucial aspects of the desirability and acceptability of the system used.

- It is also recommended to evaluate the **points of weakness and strength, threats and opportunities** of the algorithmic model, regarding the problem solving process prior to the use of the algorithmic model.

- Once an algorithm is in place, it is important to know **how and under what circumstances it is used by the people and human teams that interact with it.**

- A key question in this sense is whether the **results are applied in a unified manner**, or whether their interpretation or use differs depending on the person interacting with it.

- Thus, it is recommended to **collect information on a human scale** used to interpret and apply the results provided by the algorithm. The definition of this scale must be clear, since it determines the weight that the human validation of the algorithm's result may have in making a decision or other relevant process.

- In the case of algorithms used to assist in decision making, such as classification, precision or recommendation algorithms, it is also advisable to **collect data on the algorithm's result and the final decision made by the person interacting with it**, in order to consider possible adjustments to the model in the future. This collection of information should be considered from the design of the algorithm's dynamics and should also be carried out in accordance with the data processing principles specified in this Guide.

- It is also recommended that a process be established whereby it is specified what should be done and how those responsible for algorithm development and implementation should be informed **in the event that human validation indicates that one should proceed in a manner contrary to or significantly different from that indicated by the algorithm**.

- Regarding the **dynamics and processes** to which the system belongs, it is recommended to explain how they have changed compared to before the existence of the system and how the institutions implementing it have adapted to it.

- The same is true for the **specific objectives** pursued by the algorithm's use.

4.3.1.2 Regarding transparency measures and accountability and responsibility mechanisms

- It is recommended that clear information be **made explicit** to those who interact with the model or may be affected by it, as well as to the general public, at least regarding the **objectives of the algorithm, its features, the type of data it processes, how it is used, how the results of the algorithm are used, and with whom the data is shared.**

- It should be kept in mind that users are not, in many cases, in a position to implement or understand this information, so it should be presented in a **concise, simple and, if possible, visual way.**

- In cases where **proxy variables** are used, it is recommended to describe as precisely as possible which proxy variables the system is capturing, how it is combining them, and for what reason.

- Regarding the **precision data of the model**, it should be indicated which are the parameters and cut-off values for the model to consider certain variables when providing results that are significant. Explain this clearly to those affected. This also applies to conducting audits, in which case the process should be as transparent as possible, both for the client, stakeholders and, if applicable, the general public (if the audit is made public by agreement of all parties).

- In the development and implementation of an algorithm, the **distribution of responsibilities** should be explicitly stated, so that it is clear who decides what and who assumes responsibility for the results of the development, implementation and use of an algorithm, especially when these may be negative.

- In certain cases, the accuracy of the system is predetermined by **evaluation measures** performed by the people or organizations developing or implementing the system. It is recommended that these be **clearly explained** to the public. In the event that these measures are inadequate, other **complementary measures** will be provided and a recommendation will be made to **change or supplement** this way of measuring the system's accuracy.

v. APPENDIX

5.1 APPENDIX 1: GLOSSARY

For the purposes of this guide, it is useful to first define a series of concepts relevant to the understanding of the algorithmic auditing methodology.

5.1.1.1 Algorithm

As mentioned in the Introduction section of this guide, the word "algorithm" is used in align with its origins in the computer science field. From this perspective, an algorithm consists of a set of defined, non-ambiguous, ordered and finite instructions or rules that typically answer a question, make a decision, solve a problem, perform a computation, process data or carry out some task. These computational procedures take one or more input values and generate one or more output values. Therefore they are instruments that produce a result, instead of attempting to establish a causal link between a specific variable and its effect.

Algorithms are often implemented in decision-making processes,[26] for classifying items or predicting events. At present, the word "algorithm" is often used in reference to automated computational processes, called Machine Learning Algorithms, which are the most widely implemented during the last two decades. This glossary explains the main characteristics of Machine Learning Algorithms according to their modes of learning.

[26] It is important to emphasize that the use of algorithms in the decision-making process should play a complementary role, and not a substitute for human decision-making, especially in decisions that may significantly affect the lives of individuals, in compliance with Article 22 of the GDPR.

5.1.1.2 Algorithm with social impact

In general, this Guide considers that the use or implementation of an algorithm is particularly likely to have a social impact when it **handles personal data** (or data whose linked identity is deducible), makes **decisions or influences decisions that may have significant effects** on the social workings or lives of individuals. These effects may be **positive or negative**. However, in general, when we speak of social impact, we refer to those **effects that are considered negative**. In the case of algorithms, these negative effects are usually **linked to forms of bias or discrimination.** In this sense, algorithms can **reproduce or reinforce existing inequalities or generate new ones**, thus harming vulnerable individuals or groups.

Similarly, it is important to bear in mind that an algorithm designed for a particular service or product, under reasonable and prudent measures to fulfill a given function, **can have a detrimental effect from an ethical, social and even legal point of view.** This has to do with the high levels of unpredictability in these systems.

5.1.1.3 Algorithmic bias

Algorithmic bias occurs in those cases where a given data-driven algorithmic model repeatedly produces results that are undesired by the people developing, creating and training the system. Often, but not always, this is due to biased collection and use of training data (pre-algorithmic bias). At other times, it is due to problems with the interaction between an algorithm and other processes, once the algorithm is applied in a particular context (post-algorithmic bias).

In cases where these undesirable outcomes result in a form of systematic discrimination, which produces disadvantageous outcomes involving one or more of the so-called protected or vulnerable groups, a **discriminatory algorithmic bias** or algorithmic discrimination is considered to be observed.

5.1.1.4 *Algorithmic discrimination*

Algorithmic discrimination refers to the unequal treatment of an algorithm towards a person X, with respect to another person Y, because of an attribute of X, especially if this is a protected attribute (see definition above). This circumstance does not necessarily imply that the discrimination is negative or disadvantageous, but may also be positive or advantageous. This will depend on how the results are interpreted from an ethical and social point of view, in an overarching context. An example of this would be a form of discrimination that positively affects a protected or vulnerable group (e.g. disabled people) by providing them with significantly more resources than a privileged group (e.g. non-disabled people).[27]

5.1.1.5 *Anonymized data, anonymization*

Following the specifications provided by the GDPR (provisions in 26), this Guide considers that anonymized data can be defined as "information that cannot be linked to an identified or identifiable natural person." Therefore, anonymization means the process of rendering data anonymous, so that a person is not identifiable through it.

5.1.1.6 *Group discrimination*

This form of algorithmic discrimination refers to that which affects a person because of his or her membership in a socially identifiable or protected group. In other words, a group mainly relevant in the social and economic fabric.

---

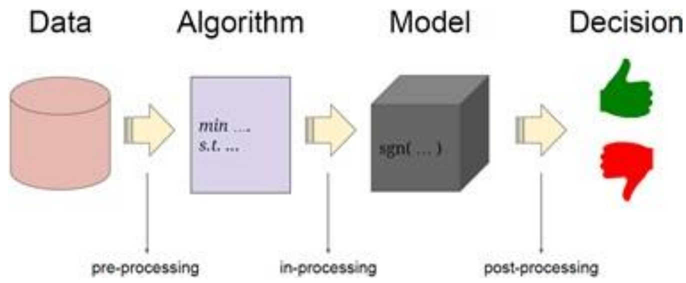
[27] The definitions of discrimination and bias presented in this guide are based primarily on work by Barocas and Selbst (2016), Baeza-Yates (2018), Castillo (2018), Cowgill (2019), Hajian, S., Bonchi, F., and Castillo, C. (2016), Lippert-Rasmussen (2013), Pedreschi et al. (2008). They also follow the interpretation of prior work published by Eticas Research and Consulting.

5.1.1.7 Labelled data

Labelled data is that fed into an algorithm and linked to certain output information. Labels in the data allow the system to know the content of this data. An example of this would be the identification of topics or attributes contained in a text fragment, for an algorithm dedicated to this function. For example: for a certain text in a resource allocation algorithm, the labeled data would indicate that the text refers to a female person, with a problem of food shortage.

5.1.1.8 Lifecycle of an algorithm

The development and implementation of an algorithm has different phases, represented in the graph below. First, a database is collected, which will be used for training and testing the system. Secondly, the algorithm code is programmed and then trained to generate the algorithmic model. This is tested prior to its final implementation.



Source: Hajian, Bonchi and Castillo (2016)[28].

At each stage of an algorithm's development, the audit functions may vary. It is therefore important to determine the algorithm's degree of development and, based on this, establish what analyses can be carried out. The auditing of an algorithm can be regarded from three approaches, depending on these stages: in the **pre-processing phase**,

[28] The graph is adapted from the article written by Hajian, Bonchi and Castillo (2016), often used by one of its authors, Carlos Castillo.

issues related to the input database can be identified and corrected; in the **processing phase**, limitations of the algorithm design can be detected and measures to avoid discrimination can be proposed; in the **post-processing phase**, improvements to modify the results of the developed models can be suggested (Hajian, Bonchi and Castillo, 2016).

5.1.1.9 Input data

Input data is that fed into the algorithm in order to be processed by it.

5.1.1.10 Output data

The output data is that resulting from the algorithmic processing of the input data.

5.1.1.11 Personal data

This Guide uses the definition of "personal data" provided by the General Data Protection Regulation (Article 4. 1). That is: "**any information relating to an identified or identifiable natural person**" ("the data subject").

"Identifiable natural person" means "any person who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person."

5.1.1.12 Protected and/or vulnerable groups[29]

The concept of protected groups is of particular relevance to the algorithmic auditing methodology of this Guide, which is based on a

[29] It should be noted that, although the terms protected group and vulnerable group are used interchangeably throughout this text and are very similar, they do have some differences. While vulnerable groups, already mentioned in this document and covered by various regulations, refer to a series of groups in a situation of lesser power or autonomy, the idea of protected group implies the

definition of vulnerable groups or key protected groups, defined by the membership of individuals who share one or more of the following protected attributes[30]:

- Children and the elderly (**age**).
- Having a **physical or mental disability** or illness.
- Gender (**female**) or gender **reassignment**.
- **Sexual** orientation (LGTBIQ+).
- Ethnic or racial origin, skin color, ancestry, national or immigrant status or other data concerning the person's origin (**racial** status).
- **Pregnant** women.
- Political, religious or philosophical **beliefs** or opinions.
- **Union** membership.
- **Genetic, biometric** or **health-related** information.
- Property or material resources, socioeconomic status and social class (**socioeconomic** status).
- Information on **criminal convictions and offenses**.

This is not an exhaustive classification and should be adapted or modified according to each context. Protected groups will be defined dynamically during the auditing process. This issue will be taken up again in the methodology section.

active and special consideration of this group in the context of algorithmic analysis or other types of social evaluation.

[30] This classification has been drawn up mainly according to Articles 6, 9 and 10 of the GDPR, relevant provisions and the European Charter of Fundamental Rights and other relevant texts. Disadvantaged groups can be defined in relation to the attributes mentioned in Article 21 (non-discrimination) of the European Charter of Fundamental Rights: "sex (and gender), race, color, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation."

5.1.1.13 Reinforcement learning algorithm

An algorithm designed to observe the interaction of the system with its environment, and take advantage of this to improve the algorithm's performance. In the learning process, the system analyzes and evaluates different possible actions, with the objective of automatically determining the most suitable one within a specific context. The reinforcement signal consists of simple feedback that the system takes as a "reward" and allows it to determine how "suitable" a certain behavior is. This may involve either maximizing the virtues of the model or minimizing its risks, biases or undesirable effects.

5.1.1.14 Responsibility and accountability

Algorithms are not **autonomous entities**, but they **lack intentionality and free will.** They cannot be granted responsibilities with respect to social, ethical or legal standards.

Algorithmic responsibility is, therefore, given to the person(s) or groups of people or organizations that directly determine the ends and means used for the design, development and implementation of the algorithm, which performs actions with specific intentions and significant consequences, especially when these consequences have negative effects on the life of another. Algorithmic responsibility defines the relationship between the party responsible for the algorithmic system and the party affected by it.

Accountability refers to a person, group or organization assuming this algorithmic responsibility. It refers to the obligation of acknowledging and accepting the consequences of an algorithm’s use, as well as to make amends with and satisfy the people affected by it. It also refers to the responsibility to prevent and avoid possible undesirable consequences in the future. Thus, accountability can be **retroactive** (relating to past actions) or **prospective** (relating to future actions). It establishes a link between the agents and recipients of the consequences of an algorithm and organizes social relations around the procedures necessary for its design and implementation.

5.1.1.15 Semi-supervised learning algorithm

An algorithm that is midway between supervised and unsupervised. It contains some labeled input data but generally most of them are not labeled. Thus, the unlabeled data represent an important source of information for system modeling, but are supplemented by automatic procedures. These algorithms are considered more suitable for model building, since they rely on patterns generated and entered by people, even as they modify them, thus augmenting human expert knowledge.

5.1.1.16 Sensitive data

As in the previous case, the definition of data or "sensitive attributes" used in this guide is determined by those types of personal data that the GDPR confers special protection (Article 9 and provisions) due to their nature and because they are particularly sensitive regarding fundamental rights and freedoms. It is understood that, by default, sensitive data is everything that belongs to so-called **"special categories of personal data"** by the GDPR. Namely, "personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data and biometric data processed for the purpose of uniquely identifying a natural person, data concerning health, data concerning a natural person's sex life or sexual orientation of that natural person."

Other data that by nature requires special protection includes personal data relating to **criminal convictions and offenses**. The GDPR limits this processing (Art. 10), and establishes special safeguards such as carrying out an impact assessment (Art. 35).

5.1.1.17 Social impact taxonomy

This social impact refers to the disadvantageous discriminatory effects or forms of discriminatory bias produced by an algorithm on people's lives, especially if these are caused by reason of their **belonging to one of the vulnerable groups** mentioned above. In this sense, the types of social impact of an algorithm can be classified as forms of discrimination, according to the following taxonomy:

- **racial**,
- gender-based,
- **sexual**,
- relating to **socioeconomic** level,
- relating to **socio-demographic** conditions (such as age),
- relating to **religious, political** or **philosophical** beliefs,
- relating to a **disability** or mental or physical illness

Likewise, this social impact may refer to the negative or discriminatory effects produced by an algorithm, insofar as it contributes to:

- conveying or reinforcing an existing social inequality (**reproduction of inequality**);
- misinforming, generating political disaffection or polarization, hindering access to different or opposing ideas and thus undermining democratic quality (**impact on democratic processes**);
- or violating compliance with individuals' fundamental rights to privacy and data protection **(privacy impact).**[31]

[31] This is an adaptation of the taxonomy developed by the Eticas Foundation team in its Observatory of Algorithms with Social Impact (OASI): https://eticasfoundation.org/algorithms/es/.

5.1.1.18 Statistical discrimination

Statistical discrimination refers to group discrimination based on a fact that is **statistically relevant.** This can occur, for example, in the case of an algorithm dedicated to prediction, which uses data on probabilities that come from the real world (and are statistically relevant), but whose use leads to disadvantageous treatment towards a certain vulnerable social group or collective. A real-world example of this is the case of an algorithm dedicated to recidivism prediction, which was shown to be discriminatory in its use of information on recidivism among black people.[32]

5.1.1.19 Supervised learning algorithm

An algorithm where humans act as "instructors" of the algorithm. In other words, they feed training data into the system, which contains the input data and also the "correct" output data for that input data. This "correct" output data is labeled data. The algorithm must reproduce this "pattern" on future occasions to produce new output data, following the same logic. The objective of this type of algorithm is precisely to "model" the "behavior" of the system.

5.1.1.20 Unsupervised learning algorithm

An algorithm where humans do not act as "instructors" of the algorithm because: the algorithm works with unlabeled data. Humans do not train the algorithm, as in the case of supervised learning. These types of algorithms are designed to be able to detect latent patterns and rules in the data and to summarize and cluster the units of information that make up the data. Therefore, they are especially useful in cases where

---

[32] For more information on this example, concerning the case of the COMPAS algorithm, please refer to the following website: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. It should also be noted that in a similar case in Europe or processing European data, in accordance with Article 10 of the GDPR, the use of such data relating to criminal convictions or offenses should be adequately reported to the competent authorities and have a basis of legitimacy.

the person (developer or manager of an organization) has not defined what he/she is looking for in the data.

5.1.1.21 Variable

This Guide uses the concept of variable as a statistical variable. A statistical variable is the set of values that contain a certain characteristic of the population about which a study (statistic) is carried out and measured.

5.2 APPENDIX 2: TEMPLATE FOR AN ALGORITHMIC AUDIT REPORT

The following is a template for the contents that should be included in the final algorithmic audit report:

1. **Title page**

   1.1 Title of project.

   1.2 Name of audited algorithm.

   1.3 Information about the auditing company (such as name or logo).

2. **Secondary title page**

   2.1 Title of project and name of system audited.

   2.2 Date of report.

   2.3 Name and organization that members of the audit team belong to.

3. **Index of figures and tables**

4. **Introduction**

   4.1 Scope of the audit and the main points agreed upon in the Analysis Plan.

   4.2 Responsibility of the audit team.

   4.3 Audited entity and the report's table of contents.

   4.4 Definition of the algorithmic problem.

4.4.1 Algorithmic design, development and model.

4.4.2 How the algorithm is used: processes, dynamics and equipment interacting with it.

5. Audit objectives and methodology

5.1 General objective of the audit.

5.2 Specific objectives of the audit.

5.3 Terms, time frames and analysis principles agreed upon in the Analysis Plan.

6. Algorithmic discrimination, equity and guiding principles of auditing

6.1 Algorithmic discrimination.

6.2 Algorithmic equity.

6.3 Guiding principles of auditing: ethical and legal compliance (applicable laws), acceptability, desirability, and the protection and proper management of personal data.

7. Theoretical analysis and status review on the subject analyzed by the algorithm

7.1 Status review on the specific case.

7.2 Status review of the specific problem.

8. Hypotheses/research questions on model accuracy

8.1 Hypothesis on the internal validation of the model.

8.2 Hypothesis on algorithmic discrimination.

8.3 Hypothesis on internal validation for groups.

8.4 Hypothesis on the model's acceptability and desirability.

9. Analysis of the composition of training datasets and groups

9.1 Dataset composition, model training and overall risk allocation accuracy.

9.2 Protected groups within the dataset.

9.3 Intersectional structure of training data.

9.4 Differential treatment and impact by group.

9.5 False Negative Rate (FNR) and False Positive Rate (FPR) by group.

10. Desirability analysis

10.1 Relevant information on the social, economic, technical and organizational context in which the model is embedded, how it has been designed and how it is used, how data integrated in the system is managed, compliance with applicable legal and ethical standards.

10.2 Conducting interviews, focus groups, or other methods used to obtain information.

11. Results: interpretation and evaluation

11.1 Quantitative: overall accuracy identified in risk allocation; bias and possible discrimination within the system.

11.2 Qualitative: general assessment of the model and adequacy to the guiding principles.

12. Conclusions

13. Recommendations and possible course of action

13.1 General accuracy.

13.2 Algorithmic discrimination.

13.3 Future redesigns.

13.4 Issues corrected during the auditing process.

13.5 Issues not corrected.

14. References

15. Appendix

15.1 Confidentiality agreement.

5.3 APPENDIX 3: SAMPLE RISK ASSESSMENT TABLES

The following is an example of a risk assessment table of a fictitious resource allocation algorithm related to two affected vulnerable groups: people of foreign origin (immigration) and people over 65 years of age (age). This table gathers information on relevant factors (such as the representation of the group in the training database), the results of relevant measurements conducted as part of the quantitative analysis and observations derived from the qualitative study of the case, which help to validate or refute the hypotheses raised, and complement the quantitative analysis. The result of the assessment made regarding these issues is shown in the last column, "Risk".

Analysis of the accuracy and desirability of the resource allocation model as a function of the 'gender' variable. [Analyzes whether the resource allocation system disadvantages women.]	
Relevant factors	Representation of the groups in the algorithm training database. *"The variable 'gender' is not explicitly collected in the training database, but is inferred through other proxy variables."*
Measurements taken	Disparity between false negative rates (FNRs). *"The rate of FNRs for the female gender group is higher than the rate of FNRs for the male gender group by 51%."* *"That is, there are many more false negatives for resource allocation to females than to males."*
Observations regarding the hypothesis	Noteworthy observations: *"The system most frequently underprotects the vulnerable group (women)."* *"The proxy variables used to determine the female and male group should be made more explicit."* *"It is recommended that the variable 'gender' be included in the modeling of the algorithm for proper evaluation."*
Risk	**HIGH**

Analysis of the system's differential treatment by age groups [Analyzes whether the system treats the 65+ age group significantly differently.]	
Relevant factors	Correlation of the representation in the training database with reality. *"The 65+ age group has a 20% representation in the national census. However, the training database collects only 6% of cases."*
Measurements taken	Differential impact and treatment rates between age groups (DI/DT). *"The system tends to assign a lower risk to people over 65 years of age than to younger groups. The most notable differences hover around 10%."*
Observations regarding the hypothesis	Noteworthy observations: *"The representation of the group in the database is low (only 1% higher than the recommended 5%). Because of this low prevalence, it cannot be claimed that the group is robustly modeled by the system. This may be skewing the accuracy of the model and explain the slight disparity."* *"The representation in the database is also too low with respect to its representation in the census (20% / 3%). It is recommended to revise it to improve the accuracy of the model."*
Risk	**MEDIUM**

5.4 APPENDIX 4: RELEVANT ASPECTS OF THE GDPR AND THE LOPDGDD FOR ALGORITHMIC AUDITING

This appendix of the Guide highlights the **most relevant aspects** of the data protection regulations established by the GDPR and the LOPDGDD, which form the legal basis of legitimacy for the different stages of a solution making use of an algorithm.

These texts point out potential issues to be taken into account or answered when developing or using an algorithm that collects or processes personal data, which any algorithmic audit should pay special attention to.

Firstly, in compliance with the provisions of the GDPR, any algorithm, understood as a processing tool, must comply with the **principles of data processing**; an aspect that must be evaluated when carrying out an algorithmic audit. These principles (Art. 5) refer to lawfulness, fairness, transparency, purpose limitation, minimization, accuracy, limitation of storage period, security and confidentiality in the collection and processing of data and the proactive responsibility of the data controller.

In addition, and in accordance with Article 6 of the GDPR, **for data processing to be lawful it must be based on one of the following legitimate grounds**: the **data subject** (i.e. an identified or identifiable natural person, Art. 4.1.) has given his or her consent to it; this processing is necessary for the performance of a contract of which this person is party; it is necessary for compliance with a legal obligation or to protect vital interests; it is required for the performance of a task carried out in the public interest or in the exercise of official authority vested in the **data controller** (according to Art. 4.7, the natural or legal person, public authority, service or other entity which alone or jointly with others determines the purposes and means of the processing); or it is necessary to satisfy the legitimate interests of the data controller.

Articles 7 and 8 expand on the issue of the **data subject's consent** and the conditions that must be met for it to be considered valid. It should be recalled that the GDPR establishes a **special category of personal data**, particularly sensitive data, whose processing is prohibited except in the cases set forth in Article 9. In the case of personal data relating to **criminal convictions and offenses**, data processing may only be carried out under the supervision of the competent public authorities (Art. 10).

The GDPR establishes a series of **rights of the data subject**, which must be complied with when developing and implementing a system that uses personal data. These relate to: the **transparency** of the information provided to the data subject, adequate **communication** with the data subject and the different ways in which the data subject can exercise his or her rights (Art. 12); the information to be provided when personal data is obtained from the data subject (Art. 13) and when it has not been obtained from the data subject (Art. 14); the data subject's **access** to personal data concerning him/her and to information on its **processing; rectification, erasure, restriction** of processing, **portability** and objection to data processing (Arts. 15-22); and automated processing of data, used in decision-making (Art. 22). According to the latter Article 22 of the GDPR, everyone has the **right not to be subject to a decision based solely on automated data processing**, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her. This Article is therefore of particular interest in the development of algorithmic audits.

Chapter IV of the Regulation refers to the **general obligations of the data controller** (Art. 24), including the responsibility to establish **data protection measures by design and by default (Art. 25).** In addition, this chapter defines the **roles that must be established**, such as the joint controllers (Art. 26), the representatives of controllers or processors not established in the European Union (Art. 27) or the data processor (Art. 28). All of them shall **cooperate with the supervisory authority** upon request (Art. 31). The supervisory authority is the independent public authority established by a Member State of the European Union (Art. 4.21 and 51), with competences and powers (Art. 57 and 58) in the field of data protection. For its part, the appointment, position and functions of the **data protection officer**, as a figure responsible for assisting the data controller, advising him/her and supervising compliance with the requirements imposed by the regulations, are set forth in Articles 37, 38, and 39 of the GDPR.

This chapter also establishes that each organization must prepare **Records of Processing Activities** (RPA) (Art. 30) and detail how this is to be done. However, it is up to each organization to decide at what level of segregation or aggregation it wishes to record the processing of personal data required by its activity.[33]

This chapter also determines that data **must be processed securely,** so as to prevent unauthorized or unlawful processing, loss, destruction or accidental alteration of such data (Art. 32). This implies that the data controller, based on a risk analysis, must: establish technical and organizational measures for pseudonymization and encryption of data; guarantee the confidentiality, security, availability and resilience of data processing systems and services; restore availability and access to data in the event of incidents; and establish processes for regular verification, evaluation and assessment of the technical and organizational measures that ensure the security of the processing. Specific and general risks presented by data processing must also be taken into account, and measures must be taken to ensure that any authorized person who has access to the data can only do so through the instructions of the data controller. In the event of **data security**

[33] For more information on the RPA, we recommend accessing the following AEPD web pages: https://www.aepd.es/es/derechos-y-deberes/cumple-tus-deberes/medidas-de-cumplimiento/actividades-tratamiento and https://www.aepd.es/es/prensa-y-comunicacion/blog/elaborar-el-registro-de-actividades-de-tratamiento. It is also recommended to view the Facilita 2.0 tool that the AEPD has made available for data controllers in the private sector for processing low-risk data: https://servicios.aepd.es/AEPD/view/form/MDAwMDAwMDAwMDAwMDI2MjQ5NzUxNTg3NjUyNzE0MTU4?updated=true.

breaches, notification to the supervisory authority must occur within a maximum of 72 hours (Art. 33), as well as to the data subject, when the breach is likely to put his or her rights and freedoms at high risk (Art. 34).

Article 35 establishes the rules on conducting **impact assessments related to data protection.** These evaluations must be carried out by the person responsible for the processing of personal data, in those cases where the processing of the data may involve a high risk for the rights and freedoms of people, in particular if it uses new technologies, by its nature, scope, context or purpose. This therefore includes various **forms of data processing based on the use of algorithms**, particularly those that process large amounts of personal or sensitive data.

This impact assessment generates a need to establish different forms of **proactive accountability.** This implies that the data controller must actively take control and decide what to do at any given moment, anticipating events. In other words, this responsibility implies active intervention, be it **retroactive**, involving various forms of accountability, or **prospective**, i.e. mechanisms and measures to anticipate risk. Such need requires that the party or parties responsible for the development and application of algorithms that use personal data analyze what data they process, for what purposes they do so and what type of processing they carry out in order to determine what measures are appropriate to comply with the provisions of the GDPR. This is a **particularly relevant** Article, since it is directly related to conducting algorithmic audits, since one of its main objectives, as mentioned above, is to analyze and identify stress points that may involve a breach of data protection regulations, in order to help correct them and include them as design requirements in the development of algorithms. When this **impact assessment** reveals a high level of risk, the data controller shall **consult the supervisory authority** before processing the data (Art. 36).

For its part, the **Organic Law on the Protection of Personal Data and Guarantee of Digital Rights** (LOPDGDD) complements and specifies the provisions of the Regulation for the Spanish situation, reinforcing the importance of complying with the principles of data protection and attention to the exercise of rights by the data controller, while including certain provisions applicable to specific processing operations, some of which may rely on developing solutions that make use of algorithms.

Consequently, both the **GDPR** and the **LOPDGDD** have come to establish the guiding principles that any type of processing, including those based on Artificial Intelligence solutions and that make use of algorithms, must respect by defining a well-developed framework for the data controller's actions based on the risk management of data subjects' rights and freedoms, as well as accountability or the ability to demonstrate compliance with the obligations defined by regulations.

VI. REFERENCES

Agencia Española de Protección de Datos (2018). Guía práctica de Análisis de riesgos en los tratamientos de datos personales sujetos al RGPD. Retrieved from: https://www.aepd.es/sites/default/files/2019-09/guia-analisis-de-riesgos-rgpd.pdf.

Agencia Española de Protección de Datos (2018). Guía práctica para las Evaluaciones de Impacto en la Protección de los datos sujetas al RGPD. Retrieved from: https://www.aepd.es/sites/default/files/2019-09/guia-analisis-de-riesgos-rgpd.pdf.

Agencia Española de Protección de Datos (2019). Guía de Privacidad desde el Diseño. Retrieved from: https://www.aepd.es/sites/default/files/2019-11/guia-privacidad-desde-diseno.pdf.

Agencia Española de Protección de Datos (2020). Guía de Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Retrieved from: https://www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf.

Ada Lovelace Institute (2020). *Examining the black box. Tools for assessing algorithmic system*. London: Ada Lovelace Institute.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016, May 23). Machine Bias. *ProPublica*. Retrieved from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Barocas, S. and Hardt, M. (2017). Fairness in Machine Learning. *NIPS*. Retrieved from: https://mrtz.org/nips17/

Barocas, S. and Selbst, A. (2016). Big Data's Disparate Impact", *California Law Review*, 671. Retrieved from: http://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf.

Barocas, S.; Hardt, M. and Narayanan, A. (2019). Fairness in Machine Learning. Limitations and Opportunities. Retrieved from: https://fairmlbook.org/.

Binns, R. (2018). Algorithmic Accountability and Public Reason. *Philos. Technol.*, 31 (4), 543-556.

Castillo, C. (2018). "Algorithmic Discrimination. Assessing the impact of machine intelligence on human behaviour: an interdisciplinary endeavour. *Proceedings of HUMAINT Workshop*. Retrieved from: https://arxiv.org/pdf/1806.03192.pdf

Castillo, C. (2019). "Fairness and Transparency in Ranking", *ACM SIGIR Forum, 52 (1), 64- 71. ACM*.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, *arXiv*:1610.07524. Retrieved from: https://arxiv.org/abs/1610.07524.

Chouldechova, A.; D. Benavides-Prado, O. Fialko, and R. Vaithianathan, (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency 81, 134–148*.

Centre for Information Policy Leadership (CIPL) (2020). *Artificial Intelligence and Data Protection. How the GDPR Regulates AI*. Washington: Centre for Information Policy Leadership.

Comisión Europea (2018). Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. Artificial Intelligence for Europe

{SWD(2018)137final}. Retrieved from: https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-237-F1-EN-MAIN-PART-1.PDF.

Comisión Europea (2020). *Commission Report on safety and liability implications of AI, the Internet of Things and Robotics*. Retrieved from: https://ec.europa.eu/info/publications/commission-report-safety-and-liability-implications-ai-internet-things-and-robotics-0_en.

Comisión Europea (2020). *Communication: A European strategy for data*. Available at: https://ec.europa.eu/info/publications/communication-european-strategy-data_en.

Comisión Europea (2020). *White Paper on Artificial Intelligence: a European approach to excellence and trust*. Available at: https://ec.europa.eu/info/files/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

Comisión Europea (2020). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. A European strategy for data. Retrieved from: https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf.

Comisión Europea (2020). Commission Report on safety and liability implications of AI, the Internet of Things and Robotics. Retrieved from: https://ec.europa.eu/info/publications/commission-report-safety-and-liability-implications-ai-internet-things-and-robotics-0_en.

Comisión Europea (2020). *White Paper on Artificial Intelligence: a European approach to excellence and trust*. Retrieved from: https://ec.europa.eu/info/files/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

Comisión Europea (2020). *Ethics Guidelines for Trustworthy AI*. Retrieved from: https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top.

Comisión Europea-Grupo de expertos de alto nivel sobre inteligencia artificial. (2018). Ethics Guidelines for Trustworthy AI. Burselas: Comisión Europea. Retrieved from: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

Diakopoulos, N. (2015). Algorithmic Accountability. *Digital Journalism*, 3(3): 400-403.

Diakopoulos, N., and Friedler, S. (2016). How to Hold Algorithms Accountable, MIT Technology Review, November 2016. Retrieved from: https://www.technologyreview.com/s/602933/how-to-holdalgorithmsaccountable/?utm_content=buffer19bc5andutm_medium=socialandutm_source=twitter.c%E2%80%A6.

Dwork, C. and Ilvento, C. (2018). Individual fairness under composition, *FATML*.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., et al.; Zemel, R. (2012). Fairness through awareness, *Proceedings of the 3rd innovations in theoretical computer science conference, 214-226. ACM.*

Eubanks, V. (2018). *Automating Inequality*. New York: St. Martin's Press.

Fumo, D. (2017). Types of Machine Learning Algorithms You Should Know. *Towards data science*. Retrieved from: https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Wortman, J.; Hanna Wallach, V. Daumé III, H.y Crawford, K (2020). Datasheets for Datasets. *arXiv:1803.09010*. Retrieved from: https://arxiv.org/abs/1803.09010.

Hajian, S., Bonchi, F. and Castillo, C. (2016). Algorithmic bias: From discrimination on discovery to fairness-aware data mining, *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*: 2125-2126.

Heidari, H., Ferrari, C., Gummadi, K., and Krause, A. (2018). Fairness behind a veil of ignorance: a welfare analysis for automated decision making. En S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.). *Advances in Neural Information Processing Systems* (pp. 1265–1276). Montreal QC: Curran Associates, Inc.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., et al.; Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (p. 600). ACM.*

ICO (2015). *Auditing data protection: a guide to ICO data protection audits*. London: ICO. Retrieved from: https://ico.org.uk/media/1533/auditing_data_protection.pdf.

ICO (2019). A Guide to ICO audits. Retrieved from: https://ico.org.uk/media/for-organisations/documents/2787/guide-to-data-protection-audits.pdf.

ICO (2020). *Guidance on the AI auditing framework* [Draft guidance for consultation]. London: ICO. Retrieved from: https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-and-the-turing-consultation-on-explaining-ai-decisions-guidance/.

ICO (2020). *Explaining decisions made with AI*. London: ICO. Retrieved from: https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. *In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* ’20). Association for Computing Machinery, New York, NY, USA, 33–44.*

Katell, M. Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and P. M. Krafft (2020). Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* ’20). Association for Computing Machinery, New York, NY, USA, 45–55.*

Kroll, J.; Huey, J.; Barocas, S.; Felten, E.; Reidenberg, J.; Robinson, D. and Yu, H. (2017). Accountable Algorithms, *University of Pennsylvania Law Review* (165). Retrieved from: https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3.

Lippert-Rasmussen, K. (2013). *Born Free and Equal? A Philosophical Inquiry Into the Nature of Discrimination*. Oxford: Oxford University Press.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D. and Gebru, T. (2019). Model Cards for Model Reporting, *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*: 220-220.

Narayanan, A. (23 of February 2018). Tutorial: 21 definitions of fairness and their politics [Abstract and video] *Conference on Fairness, Accountability, and Transparency*, NYC.

Nissenbaum, H. (2001). How computer systems embody values, Computer. *Computer*, 34 (3): 120–119.

Eticas Foundation. *Observatory of Algorithms with Social Impact (OASI)*. Retrieved from: https://eticasfoundation.org/algorithms/.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Sánchez-Monedero, J. and Dencik, L. (2018). *How to (partially) evaluate automated decision systems. Technical Report*. Cardiff University. Retrieved from: https://pdfs.semanticscholar.org/2a2d/ecaa5181d18911c3cc3c0e69e3ebdb7649dd.pdf.

Solans, D.; Biggio, B.; Castillo, C. (2020). Poisoning Attacks on Algorithmic Fairness. Retrieved from: https://arxiv.org/abs/2004.07401

Speicher, T.; Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar (2018). A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual &Group Unfairness via Inequality Indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18). Association for Computing Machinery, New York, NY, USA, 2239–2248.*

Striphas, T. (2012, Feb 1). What is an Algorithm? *Culture digitally*. Retrieved from: http://culturedigitally.org/2012/02/what-is-an-algorithm/.

Vedder, A.; Naudts, L. (2017). Accountability for the Use of Algorithms in a Big Data Environment. *International Review of Law, Computers and Technology*, 31(2): 206 – 224

Wachter, S., Mittelstadt, B. and Russell, C. (2020), Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI. *SSRN*. Retrieved from: https://ssrn.com/abstract=3547922 or http://dx.doi.org/10.2139/ssrn.3547922.

Wieringa, M. (2020). What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 1–18.*

GUIDE TO ALGORITHMIC AUDITING

www.eticasconsulting.com – info@eticasconsulting.com - +34 936 005 400