

Accuracy, suitability and quality of data in processing personal data with Artificial Intelligence

EXECUTIVE SUMMARY

This document formulates a series of reflections on the relationship between data quality and the principle of accuracy provided for in the General Data Protection Regulation (GDPR) in the context of the processing of personal data, highlighting that both concepts must be interpreted according to the purpose pursued and the suitability of the processing. Although closely related, the data quality requirements and the accuracy principle of the GDPR are not equivalent concepts. Data quality is a broader concept, applicable to both personal and non-personal data and must meet requirements beyond the GDPR.

The interpretation of the GDPR principle of accuracy should not be limited exclusively to the veracity or up-to-date nature of the data. Depending on the purpose of the processing, the adequacy of the data may also require other properties, such as precision, the level of granularity, frequency of updating or any other characteristic whose absence may compromise the fulfilment of the purpose pursued. The veracity and currentness of the data must be required when they are necessary for this purpose, especially when their absence may affect the rights and freedoms of natural persons, without this implying that they constitute enforceable requirements for any processing.

Likewise, when a processing operation of personal data incorporates non-personal data, these must also meet the quality requirements that are appropriate for the purpose pursued, to the extent that they may influence the suitability of the processing or those results that may have effects on the rights and freedoms of natural persons.

Quality assessment cannot be limited to input data. It is also necessary to establish objective metrics on the quality of the results generated by the processing, since, without them, it is not possible to verify or demonstrate that the aim pursued is effectively achieved and that the processing maintains the required conditions of suitability.

This relationship between quality and suitability acquires special relevance in processing aimed at the development or evolution of artificial intelligence systems and models. Datasets that do not meet the quality requirements for such development cannot be considered necessary for processing. Consequently, access to such datasets could be justified only to the extent strictly necessary to objectively assess their quality and determine their possible fitness for the intended purpose.

Governance mechanisms and data management processes must coordinate both the obligations arising from the processing of personal data and the requirements of the development, acquisition, deployment and evolution of artificial intelligence systems, ensuring the consistency of quality and regulatory compliance criteria throughout the life cycle.

Finally, defining, evaluating, and reviewing data quality requirements requires a multidisciplinary approach. To this end, it is essential to have, at least, the joint participation of specialists in data science and data protection, who are guided by objective criteria and verifiable evidence throughout the life cycle of processing.

Keywords: artificial intelligence, accuracy, minimisation, purpose, suitability, data quality, datasets, governance.

CONTENTS

I.	INTRODUCTION	4
II.	DATA QUALITY AND GDPR-ACCURACY	4
A.	Data quality	4
B.	Accuracy and minimisation in GDPR	5
C.	Data Quality Features in Standards	6
D.	Data quality features beyond the standard	6
E.	Individual data quality vs dataset quality	7
F.	Data quality vs veracity and currentness	9
G.	Alignment of GDPR-accuracy with data quality	10
III.	PROCESSING OF PERSONAL DATA AND AI	11
A.	Types of personal data processing	12
B.	Data quality in personal data processing that includes AI systems	13
C.	Requirement of quality in non-personal data within personal data processing	14
D.	AI development as a processing operation	15
E.	Determining Data Quality Requirements in AI System Development	16
F.	Quality analysis of data sources for AI	16
G.	Inadequate quality requirements	17
IV.	FINAL THOUGHTS	18
V.	ANNEX: EVOLUTION OF THE CONCEPT OF ACCURACY IN SPANISH LEGISLATION	20

I. INTRODUCTION

The concept of data quality is essential to the data economy when products or services depend on access to datasets for their development. In addition, such datasets need to be of the right quality to achieve the intended purpose, and meet different requirements, in a specific context¹. These sets should contain representative data that provides the amount of information needed to ensure the success of data-driven processes, whether personal data or not. A particular case would be the development of AI models and systems² based on machine learning techniques and/or that require datasets for verification, validation and testing.

On the other hand, the processing of personal data, which may include one or more artificial intelligence systems in its operations, must comply with the GDPR,^{3 4} in particular with regard to the principles of accuracy and minimisation in relation to personal data. Likewise, the result of a processing operation of personal data must meet the condition of suitability to fulfil its purpose, and as soon as the result relates to identified or identifiable persons, it must be ensured that their rights and freedoms are protected, in particular in respect of both principles. A specific case of personal data processing is one in which sets of personal (and non-personal) data are necessary for the development of AI models and systems.

This technical note discusses the application of these concepts (accuracy, minimisation, suitability and data quality) in the context of processing to which the GDPR applies. It is aimed at controllers, processors and DPOs, as well as for those involved in the development, testing and maintenance of products or services that require access to personal data sets, such as many types of AI models and systems.

II. DATA QUALITY AND GDPR-ACCURACY

A. DATA QUALITY

Data quality is defined in different standards (e.g. UNE 0079, ISO 25012, ISO 5259, ISO 8000) in relation to any type of data, whether personal data or not, such as the **degree** to which data characteristics must satisfy **stated and implicit needs** when used **under specific conditions**. The ISO 22989 standard⁵, in relation to AI systems, defines the data quality verification process "*in which data is examined for completeness, bias and other factors which affect its usefulness for an AI system*". These definitions serve as technical guidance, and do not replace or modify the legal requirements arising from the GDPR.

Such definitions link the concept of data quality to the adequacy of data to serve a **purpose** ("*satisfy stated and implicit needs*" or "*usefulness for an AI system*") in a given **context**⁶. "*Stated and implied needs*" could also include compliance with obligations under different regulations, compliance with standards, contractual commitments, and an entity's internal policies.

These specific needs are represented through so-called **data quality requirements**.

¹ This is how it is expressed in the AI Act for the models and systems within its competence in recital 12

² The term AI system or model is used in this text with a broader meaning than that used in the AI Act and in the European Commission's guidelines on the definition of an AI system, and may include other techniques based on data science, such as logistic regressions, mathematical optimisations, statistical methods, etc.

³ They must guarantee and be able to demonstrate that they are compliant, according to Article 24 of the GDPR

⁴ Although only the term GDPR is used, it extends to the competence established in Organic Law 7/2021, of 26 May, on the protection of personal data processed for the purposes of prevention, detection, investigation and prosecution of criminal offences and the execution of criminal sanctions.

⁵ ISO 22989 Artificial intelligence concepts and terminology

⁶ In the AI Act, recital 12 defines the context in which the IA system operates

B. ACCURACY AND MINIMISATION IN GDPR

Although data quality can be considered a classic principle in the field of privacy, the truth is that the GDPR does not refer to the term “*Data quality*”⁷, although it does establish two principles that are particularly related to it, such as that of minimisation and accuracy in article 5(1)(c) and (d):

“1. Personal data shall be:

(c) **adequate, relevant and limited** to what **is necessary in relation to the purposes** for which they are processed (“**data minimisation**”);

(d) **accurate** and, where **necessary, kept up to date**; every reasonable step must be taken to ensure that personal data that are inaccurate, **having regard to the purposes** for which they are processed, **are erased or rectified without delay** (“**accuracy**”).

While data quality requirements are not a principle of the GDPR, their role is instrumental in facilitating the practical application of those principles. Both act in a complementary, coherent and autonomous way to establish the *quality band* that the data must have, demanding that they are correct in relation to the purpose and limiting the set of data to what is necessary. Both for the personal data used as input to the processing and for the data generated as output by the processing.

		OPERATIONS FOR WHICH THE DATA IS REQUIRED		
WHERE DO YOU LIVE?		KNOW THE NATIONAL PHONE CODE	DETERMINE TAX REDUCTIONS	POSTAL DELIVERY
	Spain	Accurate	Inaccurate	Inaccurate
	SPAIN Madrid	Excessive	Accurate	Inaccurate
	SPAIN Madrid Gran Vía 320 1º	Excessive	Excessive	Accurate

Figure 1 The necessary balance between accuracy and minimisation must be found to meet the purposes of the processing

Compliance with the principles of data minimisation and accuracy is **linked to the purpose of the processing**⁸. As far as it is linked to the fulfilment of the purpose, the meaning of the principle of data accuracy in the GDPR (hereinafter “GDPR-accuracy”)⁹ has a different extension, and is not strictly linked to, for example, the veracity and currentness of each of the data. The term “GDPR-accuracy” is linked to the fact that the data is suitable to achieve the purpose of the processing, i.e. to ensure that the processing is suitable. This implies that the input data must be sufficiently accurate (GDPR) for the processing to achieve its purpose and, if the result of the processing is personal data, including decisions and profiling, these results must respect the rights and freedoms of natural persons, in particular in relation to the requirement that the data be truthful.

Compliance with the principle of GDPR-accuracy may involve satisfying different “technical” requirements, such as adequate precision, sufficient granularity, an adequate collection frequency, absence of bias, completeness of attributes, etc. An extensive interpretation of the

⁷ Except in article 47 “Binding corporate rules” in paragraph 2 (d) “The binding corporate rules referred to in paragraph 1 shall specify at least: d) the application of the general data protection principles, in particular purpose limitation, data minimisation, limited storage periods, data quality, data protection by design and by default, ...”.

⁸ In this case, compliance with principles such as fairness, on which data processing could be based to avoid, for example, bias, should also be considered, without prejudice to those established in article 9 of the GDPR and the AI Act.

⁹ The term accuracy in relation to data has a different semantic scope than that used in the standards shown above and in other regulations. This is not the only case. Terms that appear in the GDPR, such as transparency, data subject, processing, consent, controller or risk, among others, are used in other regulatory contexts with a very different meaning. When used in contexts where different regulatory frameworks apply, it is important to add a “surname” to terms in order to accurately determine their meaning.

principle of GDPR-accuracy that goes beyond the fulfilment of the purpose could affect the rights of data subjects, to the extent that it would enable data processing that could go beyond strict needs, contradicting the principle of minimisation-GDPR.

Furthermore, such extensive interpretation could lead to disproportionate burdens for controllers and processors and, in particular, for developers and organisations responsible for the evolution of AI systems, by requiring them to access and manage datasets with levels of, for example, veracity, currentness, precision, etc., that may not be necessary. This could even go against the possibility of obtaining datasets of sufficient quality to develop adequate systems and, in addition, require those controllers to pay attention to rights beyond those established in the GDPR¹⁰. In any case, the possibility of correcting personal data that is inaccurate in relation to the data subject in accordance with the right to rectification (Article 16 of GDPR) will be preserved.

C. DATA QUALITY FEATURES IN STANDARDS

There are criteria for assessing data quality. For example, the ISO 25012 lists a number of characteristics to consider: accuracy, completeness, consistency, credibility, currentness, accessibility, compliance, confidentiality, precision, traceability, understandability, availability, portability, and recoverability. The standard establishes that its evaluation must be carried out according to the **specific context of use**, so it is not a matter of absolute characteristics but relative to that context.

Although the term "accuracy" can be read in the list of characteristics of the standard, it would be a mistake to pair this term with the term GDPR-accuracy. ISO-accuracy is the degree to which data correctly represents the true value of an attribute, concept, or event in a specific usage context. ISO-accuracy could be one of the necessary conditions to achieve GDPR-accuracy, i.e. to comply with the purpose of the processing, but not necessarily. In addition, other features of the standard, such as precision, as well as features that are not listed in the standard, such as the absence of bias, may be required to achieve GDPR-accuracy.

Like any standard, the set of data quality characteristics reflects past experience and allows for the establishment of a common reference framework that facilitates the comparison of the evaluations obtained by different actors. However, compliance with all the characteristics is not required, nor do they exhaust the necessary requirements to guarantee suitability for any purpose. Therefore, the following mistakes should not be made when using these standards:

- Apply "**checkbox quality**", understood as an abstract review of the dataset without taking into account the specific context and purpose, or conducting evidence-based analysis.
- Establish that all features are equally relevant. For a given context and purpose, some of them may be irrelevant and therefore not enforceable.
- Assume that the features to be evaluated are **exhausted in the standard list**. The standard is not exhaustive, but a valuable starting point that must be adapted to the specific circumstances of each processing.

In order not to fall into these errors, it is necessary to extract the list of characteristics and their requirements according to the contexts and intended purposes of the processing.

D. DATA QUALITY FEATURES BEYOND THE STANDARD

In relation to the previous section, and in the case of developing AI models and systems, it may be necessary to consider more characteristics in relation to data quality, such as¹¹:

¹⁰ Article 16 of the GDPR also links rectification to the purpose of the processing.

¹¹ They are listed in recital (67) and article 10 of the AI Act in paragraphs 1 to 4.

- The elimination of biases that compromise the purpose of the processing¹².
- The statistical properties appropriate for certain intended purposes.
- The representativeness of the context in relation to the intended purposes.
- That the particular features, characteristics or elements of the specific geographical, contextual, behavioural or functional environment in which the AI system is intended to be used are reflected.
- The frequency of data collection, which may be relevant when processing is being carried out based on geolocation, energy consumption, Internet activity, etc.
- The predictive power of the selected variables. These are those with the ability to accurately anticipate the value of the outcome variables, as they are strongly related to them, allow clear separation between classes and reduce the uncertainty of the outcome.
- The relevance of samples for the learning process. Relevant samples are those that provide meaningful and useful information to adapt the model in each training iteration. *Active-learning* techniques in machine learning are aimed at selecting relevant samples, which will depend on the design of the model, as well as the degree of pre-training it already has.

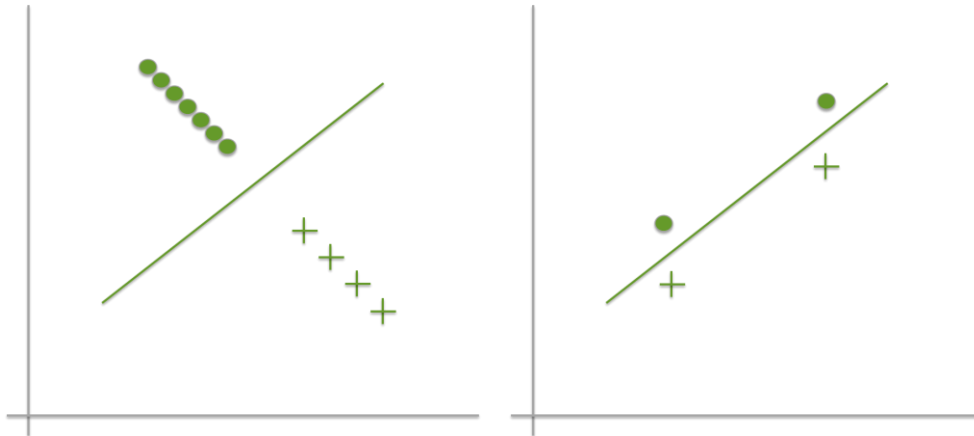


Figure 2 In a simple and obvious example of machine learning classification in which two classes have to be divided linearly. The graph on the left uses a dataset in which most of the samples are not relevant to make a proper classification, while the one on the right is a very relevant dataset to obtain such a linear classification.

Active-learning *techniques*, representativeness, relevance, and predictive power not only have advantages for the development and performance of an AI model or system, but can also be a very useful tool for applying the principle of data minimisation.

E. INDIVIDUAL DATA QUALITY VS DATASET QUALITY

Although it has been addressed indirectly, it is important to make explicit the distinction between the quality of the data, considered individually, and the quality of the data set¹³. The latter concept acquires special relevance in the case of the development of AI systems based on machine learning.

For example, the following table is the case:

¹² Biases are not in all cases negative, but, in some of them, both biases in the datasets used in machine learning may be necessary in cases such as the development of fraud detection systems, and biases in the outcome of a processing operation may be necessary for precision medicine.

¹³ The use of the English term "data quality" can be interpreted both in the singular and in the plural, and can lead to misunderstandings.

DATASET

ID	GENDER	PROVINCE	SALARY	EXP.
ID1	H	MAD	57400	Medium
ID2	H	MAD	67000	Low
ID3	H	MAD	90000	High
ID4	H	VAL	36900	Low
ID5	H	MAD	55000	Medium
ID6	H	MAD	76000	High
ID7	H	MAD	30000	Low
ID8	M	MAD	45000	Low

Figure 3 Dataset with possible compliance with the ISO standard, interpreted at the registry level, but clearly biased in relation to gender.

The above table might obtain a high degree of quality according to the ISO standard, interpreted at the registry level, but the dataset is clearly biased in relation to gender. If the purpose of the processing is to have estimates or make decisions that affect all genders, the data set shown would not be suitable to fulfil the purpose.

The processing of personal data that aims to develop AI systems using machine learning is one of many processing where the achievement of the purpose does not depend so much on the quality of the data taken individually, or on a user's specific record, but on the quality of the dataset.

ISO		AI Act and Others																													
<table><tr><th>Data/Record</th></tr><tr><td>Accuracy</td></tr><tr><td>Completeness</td></tr><tr><td>Credibility</td></tr><tr><td>Currentness</td></tr><tr><td>Accessibility</td></tr><tr><td>Compliance</td></tr><tr><td>Confidentiality</td></tr><tr><td>Efficiency</td></tr><tr><td>Precision</td></tr><tr><td>Traceability</td></tr><tr><td>Understandability</td></tr><tr><td>Availability</td></tr><tr><td>Portability</td></tr><tr><td>Recoverability</td></tr></table>	Data/Record	Accuracy	Completeness	Credibility	Currentness	Accessibility	Compliance	Confidentiality	Efficiency	Precision	Traceability	Understandability	Availability	Portability	Recoverability	<table><tr><th>Dataset Dataset</th></tr><tr><td>Consistency</td></tr></table>	Dataset Dataset	Consistency	<table><tr><th>Dataset</th></tr><tr><td>Relevance</td></tr></table>	Dataset	Relevance	<table><tr><th>Dataset</th></tr><tr><td>Bias</td></tr><tr><td>Statistical properties</td></tr><tr><td>Frequency</td></tr><tr><td>Representativeness</td></tr><tr><td>Geographic coverage</td></tr><tr><td>Contextual coverage</td></tr><tr><td>Functional coverage</td></tr><tr><td>Behavioral coverage</td></tr></table>	Dataset	Bias	Statistical properties	Frequency	Representativeness	Geographic coverage	Contextual coverage	Functional coverage	Behavioral coverage
Data/Record																															
Accuracy																															
Completeness																															
Credibility																															
Currentness																															
Accessibility																															
Compliance																															
Confidentiality																															
Efficiency																															
Precision																															
Traceability																															
Understandability																															
Availability																															
Portability																															
Recoverability																															
Dataset Dataset																															
Consistency																															
Dataset																															
Relevance																															
Dataset																															
Bias																															
Statistical properties																															
Frequency																															
Representativeness																															
Geographic coverage																															
Contextual coverage																															
Functional coverage																															
Behavioral coverage																															

Figure 4 Specific characteristics of the data or record, and those that depend on the data set, and not on the samples evaluated individually.

Other cases could be trend analysis, correlation analysis, optimisations, resource allocation, model development, decision-making in public policies or risk analysis in areas such as logistics, education, finance, energy, mobility, health, science, etc.

In these cases, to the extent that the data set contains personal data, the principle of GDPR-accuracy must take into account such important characteristics as that the data, taken as a whole, do not contain biases that compromise the purpose of the processing.

Thus, in artificial intelligence systems, the quality of the data must also be evaluated at the level of the dataset, as it is essential to guarantee the correct functioning of the system and, therefore, to comply with the purpose of the processing involved in its development. This is without prejudice to the quality of the individual data, insofar as it produces effects on individuals.

The controller, under the principle of accuracy and accountability, must guarantee and document the traceability of the set of data used, including the origin of the data, the transformations applied and the selection and exclusion criteria.

F. DATA QUALITY VS VERACITY AND CURRENTNESS

The quality of the data set for a specific purpose, in terms of personal data, may be related to the **veracity and currentness** of the data. However, the GDPR is even explicit that the currentness of the personal data is a requirement of accuracy – GDPR only when necessary to the purposes for which the personal data are processed.

The application of techniques for the elimination of biases, anonymisation techniques (not only to anonymise, but also to minimise data), differential privacy techniques or the generation of synthetic data may imply that some values do not directly reflect the reality of an individual or do not meet certain technical criteria of ISO-accuracy. However, the resulting dataset may be of sufficient quality to be used as input in all or some of the development phases of an ML-based AI system. In other words, the set of data can be considered to comply with the principle of GDPR-accuracy to the extent that it is appropriate for the purpose of the processing, without this implying that the requirements of veracity are dispensed with when these are relevant.

Moreover, there are techniques to increase the quality of the dataset that may involve the use of values that do not directly reflect the individual reality. Thus, for example, it can be considered a dataset made up of records with multiple fields in which some of them are empty or incomplete (completeness problems). One technique to increase the quality of the dataset is to estimate the possible value of empty fields by extrapolating it from statistical measures obtained from the rest of the records. This is a proper technique, but it introduces formally false and inaccurate values, sacrificing ISO-accuracy for ISO-completeness.

ORIGINAL DATASET				IMPUTED USING THE OVERALL MEAN			IMPUTED USING THE PROVINCE MEAN		
ID	PROVINCE	SALARY	EXP.	PROVINCE	SALARY	EXP.	PROVINCE	SALARY	EXP.
ID1	MAD	57400	Medium	MAD	57400	Medium	MAD	57400	Medium
ID2	VAL	-	Low	VAL	65150	Low	VAL	22250	Low
ID3	GR	90000	High	GR	90000	High	GR	90000	High
ID4	VAL	36900	Low	VAL	36900	Low	VAL	36900	Low
ID5	BCN	-	Medium	BCN	65150	Medium	BCN	45000	Medium
ID6	VAL	7600	High	VAL	7600	High	VAL	7600	High
ID7	VAL	-	Low	VAL	65150	Low	VAL	22250	Low
ID8	BCN	45000	Low	BCN	45000	Low	BCN	45000	Low

Figure 5 In the table on the left, there are fields that were not originally completed. The data table can be populated with different strategies. In the table in the center, we have chosen to enter the average of the salaries of the entire table. On the right, the average salaries of the same province have been used.

Personal data sets that have "invented" fields can maintain an adequate level of quality from a technical perspective. The ISO-accuracy characteristic would not be met for certain individual records, but that would not be relevant for a specific processing. In addition, the principle of GDPR-accuracy would not be violated if the set of data is adequate to achieve the purpose of the processing.

Similarly, a bias correction technique consists of applying corrections to normalise by group, calculating group bias and eliminating systematic difference:

ORIGINAL TABLE					BIAS-MITIGATED TABLE			
ID	Age	Income (€)	Post Code	Credit History	Age	Income (€)	Post Code	Credit History
1	45	60000	28001	0,80	45	60000	28001	0,80
2	25	28000	28003	0,70	25	28000	28003	0,90
3	50	72000	28002	0,85	50	72000	28002	0,85
4	38	58000	28001	0,90	38	58000	28001	0,90
5	29	30000	28003	0,60	29	30000	28003	0,80
6	41	65000	28002	0,85	41	65000	28002	0,85
7	33	32000	28003	0,65	33	32000	28003	0,85
8	41	65000	28002	0,95	41	65000	28002	0,95
9	36	70000	28001	0,75	36	70000	28001	0,75

Figure 6 In the original table, there is a bias in the assessment of credit history in relation to people living in the 28003 zip code. The average for the other postal codes is 0.85, while the average for people with ZIP code 28003 is 0.65. The bias takes a value of 0.2, which can be corrected by normalizing the values.

Thus, quality understood as suitability for a purpose is not automatically identified with truthfulness in the absolute sense. The relevance of the veracity of the data varies depending on the type of processing. Thus, in aggregate or statistical processing, higher levels of abstraction or transformation of the data may be admitted, while in processing that produces effects on specific persons, the requirement of veracity and timeliness, at least in the result of the same, must be stricter. This is without prejudice to the fact that it must be guaranteed that there is no incorrect or distorted representation of the interested person, or that adverse effects are generated on their rights and interests. In short, the use of data that does not reflect reality must be analysed in terms of its impact on decision-making, profiling, or any other significant effect on the natural person, and must be justified and documented.

G. ALIGNMENT OF GDPR-ACCURACY WITH DATA QUALITY

The concept of GDPR-accuracy aligns with the data quality definitions of the data economy standards insofar as both **are linked to the principle of finality**. However, the two terms are not interchangeable. Data quality applies to both personal and non-personal data within the industry framework, while GDPR-accuracy applies to personal data within the scope of the GDPR¹⁴.

On the other hand, the level of data quality may be determined by technical or functional requirements that do not derive from data protection regulations, for example, compliance with obligations under other regulations, compliance with standards, contractual commitments and internal policies of an entity. Consequently, those requirements do not necessarily have to coincide with the requirements of the GDPR principle, which must be assessed in terms of the purpose of the processing and its effects on individuals.

In both cases, the terms data quality and GDPR-accuracy are not configured in absolute terms, but in relation to the purpose of the processing. Therefore, **data that are adequate or accurate, and therefore meet the quality for one purpose, may not have the necessary quality for a different purpose**. Similarly, **untruthful or non-current data, to the extent that they allow the purpose of a specific processing to be fulfilled, would be sufficiently accurate-GDPR or would have the necessary quality** for such processing.

¹⁴ As noted above, the reference to the GDPR used in the text includes L.O. 7/2021.

III. PROCESSING OF PERSONAL DATA AND AI

A process is a set of mutually related, or interacting, activities that transform input elements into outputs¹⁵. Every process has a purpose, and its implementation will be effective when it is capable of fulfilling that purpose. Likewise, to ensure that these results meet minimum **requirements of effectiveness, evidence** must be obtained through the **definition of specific metrics** for each type of process that allow verifying whether these levels are reached.

To the extent that the process involves personal data, we are dealing with GDPR processing, which in order to comply with the principle of purpose must be able to achieve its purpose, that is, to be **suitable**¹⁶.

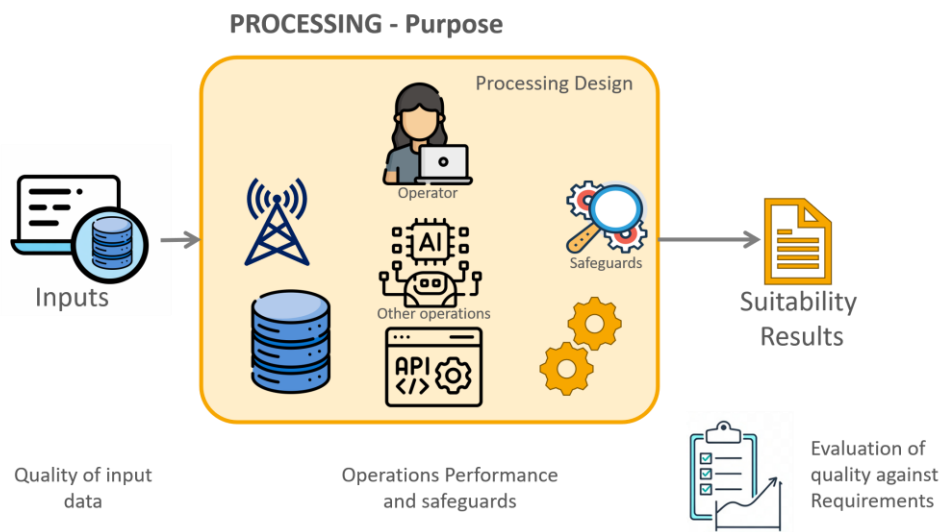


Figure 7 The processing output must be fit for its intended purpose, which will depend on its design, the performance of the operations that form it, the safeguards deployed and the data entering the processing.

Compliance with the suitability requirements will depend, fundamentally, on the following factors:

- The design of the processing as a set of operations of a different nature¹⁷ and the relationship between them.
- The quality of the inputs for the processing, including the quality of non-personal data.
- The quality of the data that are the results of each of the operations that make up the processing, which generate intermediate data.
- The safeguards deployed in the processing (e.g. pre-processing of the data, human supervision, supervision with algorithms, convergence assessment, filtering according to policies, alerts, management of complaints and incidents, ...), which could be applied to correct quality problems of the input data, quality problems in the intermediate data generated by each of the systems implementing the processing operations and, also, the safeguards that correct the output of the processing.

These factors are interdependent, and must be adequately balanced to achieve the desired degree of suitability in the processing outputs.

¹⁵ ISO 9001:2015

¹⁶ STC 14/2003, of 28 January, suitability judgment: to assess whether the measure is likely to achieve the proposed objective.

¹⁷ Operations executed by people, by digital systems, by AI systems, by the controller, by a processor, in the cloud, automated, etc.

A. TYPES OF PERSONAL DATA PROCESSING

The result of the data processing, both personal and non-personal, can be an action, a decision, including a decision about a natural person, a profile about that person or personal data in general. It can also result (even simultaneously) in non-personal data, such as a strategic decision, a statistical value, the definition of a policy, or an AI model that does not contain personal data.

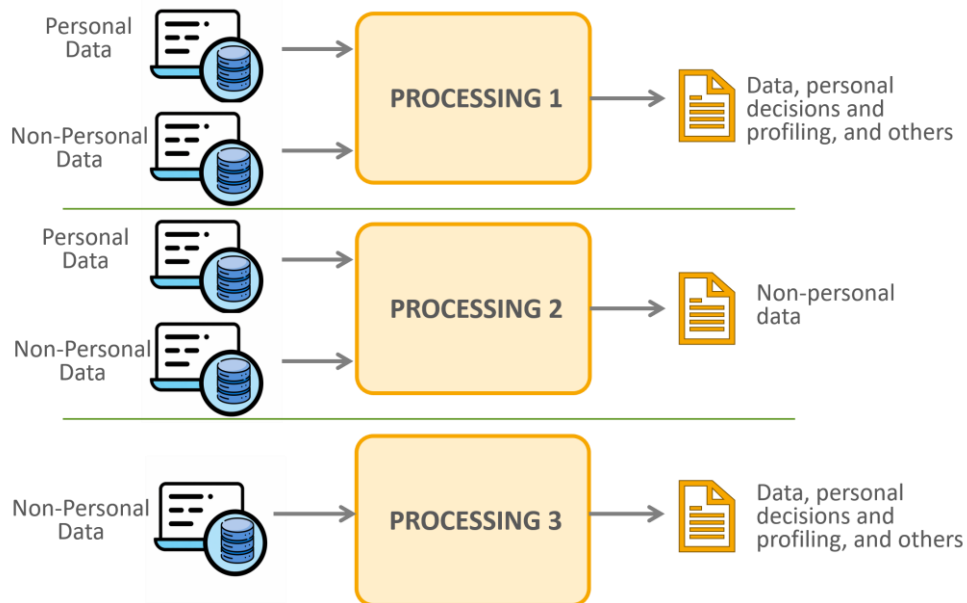


Figure 8 Three cases of processing of personal data classified according to the criterion of whether the data entering the processing is personal or not, and whether the output is personal data or not.

A processing operation (case 1) may, on the basis of personal and non-personal data (the latter could be, for example, scales, statistics, budgets or others) obtain an output that could determine whether a citizen is going to receive a grant or financial assistance, assess their financial reliability, determine their eligibility for employment, determine their productivity, give them access to a service or premises, etc. That is, to make a decision, profile it or infer other data. This output, which involves data linked to a natural person, must be sufficiently accurate-GDPR (and compliance with the data minimisation principle-GDPR) for the purpose of the processing to be fulfilled in a suitable manner. In turn, the personal data entering the processing will have to be accurate under the GDPR (and compliance with the data minimisation principle-GDPR). Also, the set of all the data entering the processing, including non-personal data, must have sufficient data quality so that the purpose can be achieved as the processing is designed and implemented.

A processing operation (case 2) may use personal (and non-personal) data to define a public policy, such as the distribution of educational resources according to the profile and performance of students, analysing an economic situation in a region, etc. In this case, the output of this processing would not be personal data. However, the output must be of sufficient quality so that the purpose can be fulfilled and the processing of the personal data used is fair and legitimate. In turn, the data entering the processing must meet the conditions described in the previous case.

A processing operation (case 3) that, without receiving any personal data from the outset, makes a decision about a natural person, profiles or infers data that is linked to a natural person is more difficult to consider. One example would be a lottery draw, the generation of numbers in a lottery, or any other case in which a rule is applied that is independent of the individual but that is ultimately associated with that individual or affects them significantly. To the extent that the result is personal data, it will have to meet the same conditions as in the previous cases.

B. DATA QUALITY IN PERSONAL DATA PROCESSING THAT INCLUDES AI SYSTEMS

The different operations of a particular processing can be implemented in different ways: with people and/or with systems and/or in particular with one or more AI systems. If we focus on the AI system, it must produce results of sufficient quality to fulfill the purpose of the processing.

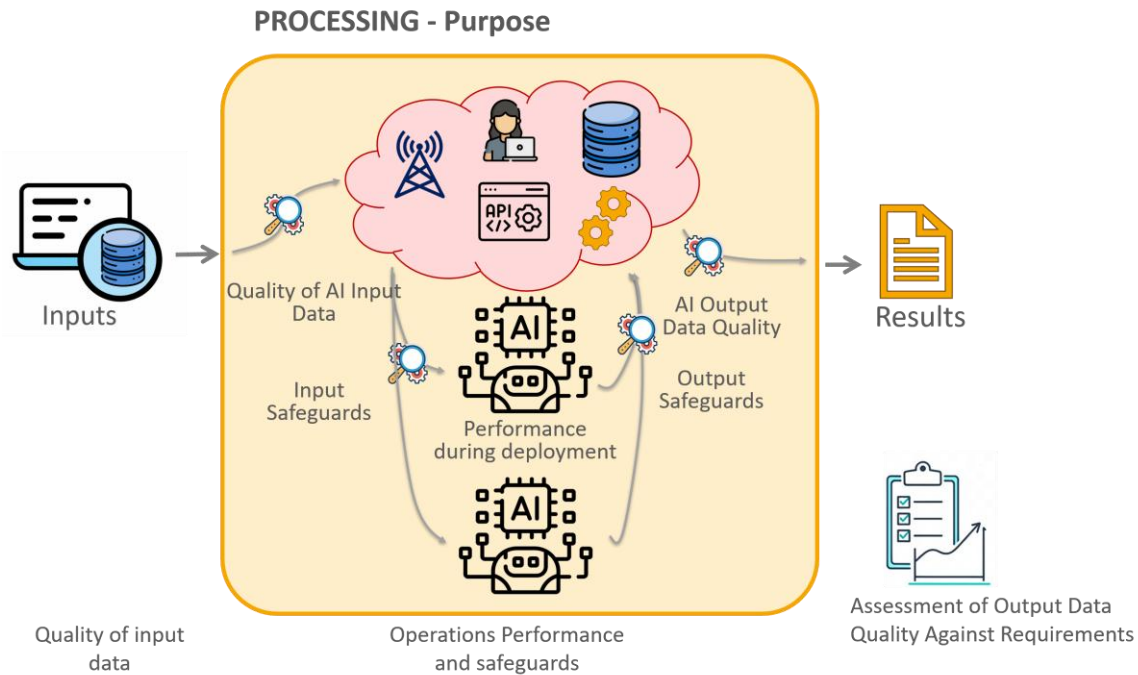


Figure 9 Data quality aspects to consider when deploying one or more AI systems/models in a processing operation. Two have been deliberately represented in the figure, to show that there is no direct correspondence between the AI system and the processing

The AI system could be as simple as an AI model with limited interfaces or, conversely, a complex generative AI SaaS service, or agentic, that already incorporates numerous operations and safeguards¹⁸ both at the input and output of the system.

The outputs of an AI system (whether they are called infer¹⁹, predictions, content, recommendations, or decisions, among others) that is part of a processing operation may be an input for subsequent operations of the same processing. Among them, and it would be most advisable, operations that implement additional safeguards and human supervision. Therefore, the minimum level of data quality and/or GDPR-accuracy required for the outputs of the AI system will depend on several factors:

- The data quality and/or GDPR-accuracy requirements defined to ensure compliance with the purpose of the processing in a suitable manner.
- Respect for the rights and freedoms of identified or identifiable natural persons when such results relate to them.
- The set of operations subsequent to the AI system within the processing that modify or correct the results generated by that AI system (an AI system could generate outputs for another AI system within the same processing, as in multi-agent systems).

¹⁸ Firewalls, interaction limiters, load balancing, circuit breakers, personal information detectors, text and content classifiers, sanitisers, embedding models, rerankers, memories, RAGs and their management, prompt generator, model selectors, scaling, self-recovery, reformatting, hallucination detector, copyright, reference and citation manager, toxicity filters, service policy enforcement, tool calls, APIs, MCP clients, etc.

¹⁹ The AI Act, in recital (12), use "capability to infer" as the ability to obtain the outputs listed below.

- In particular, the internal safeguards included in the processing design aimed at correcting known errors in the AI system.

The ability of the AI system to provide the required quality once deployed in the processing will not depend solely on the processing, but on several factors in the design of the processing and its context, for example:

- The quality of the input data to the AI system, which may have been the result of other operations carried out within the framework of the processing, such as data collection from sensors, databases or interaction with a human.
- Of the safeguards deployed during the design of the processing that are applied to said input data, such as those that prevent a degradation in the dynamics of the operation on the deployed AI system (feedback loops²⁰).
- Of the adequacy of the AI system, designed for more or less specific purposes, in relation to the purpose of the processing where it has been deployed.

The quality requirements of the data entering the processing may be defined when the conditions of suitability of the processing are known. These, in turn, may be determined taking into account, among others, the requirements of GDPR-accuracy of the result and the impact on the rights of natural persons. The design of the processing, including the safeguards that are planned, may qualify these quality requirements of the data input.

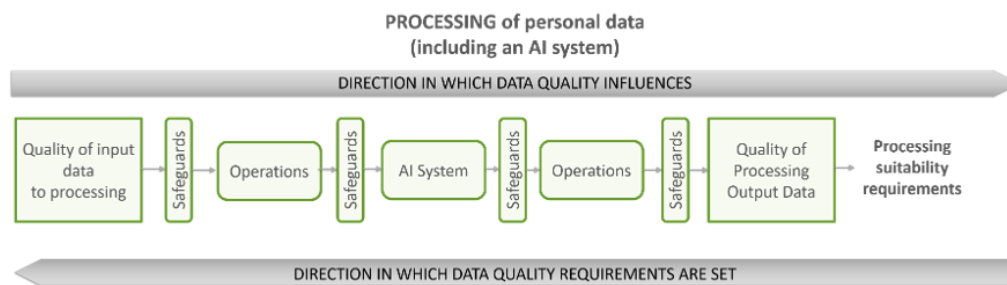


Figure 10 Defining the quality requirements of input data is not possible if the suitability requirements of the likely processing are not known.

The facts show that the perfect system is an illusion, whether it is an AI or non-AI system. The suitability of an AI system will depend on how the design of the processing, where it is deployed, takes advantage of its benefits and adequately manages its defects within the framework of a specific context and purpose. An AI system or model can have very different performance depending on the deployment conditions in a given environment. Therefore, **a perfect AI system would not be required, but an AI system appropriate to the design of the processing,** taking into account the effects of the system on the decisions or evaluations of the people affected. It should be noted that the aggregate performance of the AI system can hide errors or inaccuracies in individual inferences, especially when evaluated using global metrics. This requires the adoption of additional control mechanisms when the results of the system may affect specific persons.

C. REQUIREMENT OF QUALITY IN NON-PERSONAL DATA WITHIN PERSONAL DATA PROCESSING

In many cases, non-personal data will also be involved in the processing of personal data, both as part of the processing and as part of its design²¹. The quality of non-personal data may also be

²⁰ Recital 67 of the AI Act. In the case of deployment without retraining, because RAGs, prompts or memory are fed dynamically.

²¹ They can take the form of databases, files, tables and certain configuration parameters of the systems used in the implementation of the processing.

critical to achieving the adequacy of processing and/or in relation to the impact on the rights and freedoms of identified or identifiable natural persons. To this extent, the quality requirements for non-personal data and appropriate measures must be established to ensure compliance with the processing.

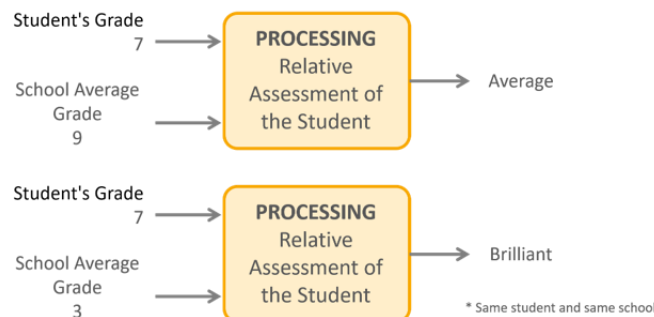


Figure 11 Example of how a non-personal input data item may influence the outcome of the processing and affect an individual. An error (i.e. a data quality issue) in the non-personal input data may result in an output that is not accurate under the GDPR, even if the personal input data are accurate under the GDPR.

D. AI DEVELOPMENT AS A PROCESSING OPERATION

A specific case of personal data processing (view case 2 defined in section A of this chapter) involves the development or evolution of an AI model or system with data-driven machine learning. Such development will be a different processing from the processing in which the AI system is used in the operation or deployment phase to process personal data.

The purpose of development processing could be to create a system that performs a specific task, and already within the framework of the processing decide that the nature of that system will be an AI. Or the purpose could be the development of an AI itself, for example, in the case of projects whose funding or grant specifically requires the design of an AI model.

The life cycle of such processing could be represented using the development and evolution phases defined in the ISO 5338 standard.

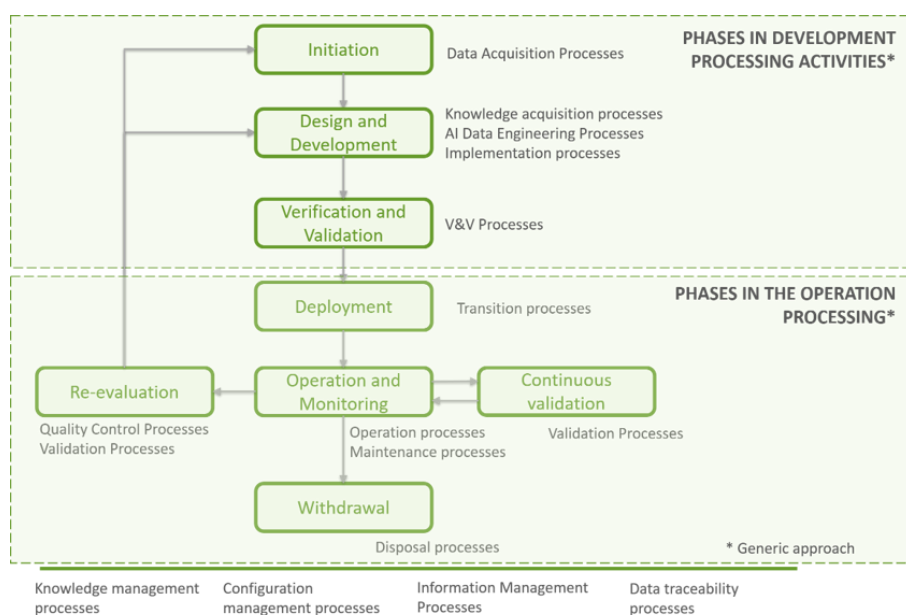


Figure 12 Phases of the AI systems development life cycle according to the ISO 5338 standard, highlighting data processes and a general classification of the phases of the life cycle that would be part of operations of a development processing and those that would be part of an exploitation processing when it involves the processing of personal data.

The performance of the resulting AI system or model, in relation to the quality of the results it offers, will depend, among others, on:

- The quality of development **datasets** (training, testing, validation, evolution).
- Changes in data quality throughout the data processes of the life cycle.
- The development methodology used.
- Design decisions.
- The contexts and intended purposes.
- Implementation factors in deployment (e.g. model quantisation).
- The evolution of the context over time (e.g. cases of context-drift and data-drift)
- In the case of evolution of the AI model, for example, biases generated by feedback loops.

E. DETERMINING DATA QUALITY REQUIREMENTS IN AI SYSTEM DEVELOPMENT

The data quality management applicable to a processing operation intended for the development of an AI system, as indicated in section B of this chapter, may only be defined once the conditions of suitability of the processes or processing in which the system is planned to be deployed are known, since these conditions determine the quality requirements for the data.

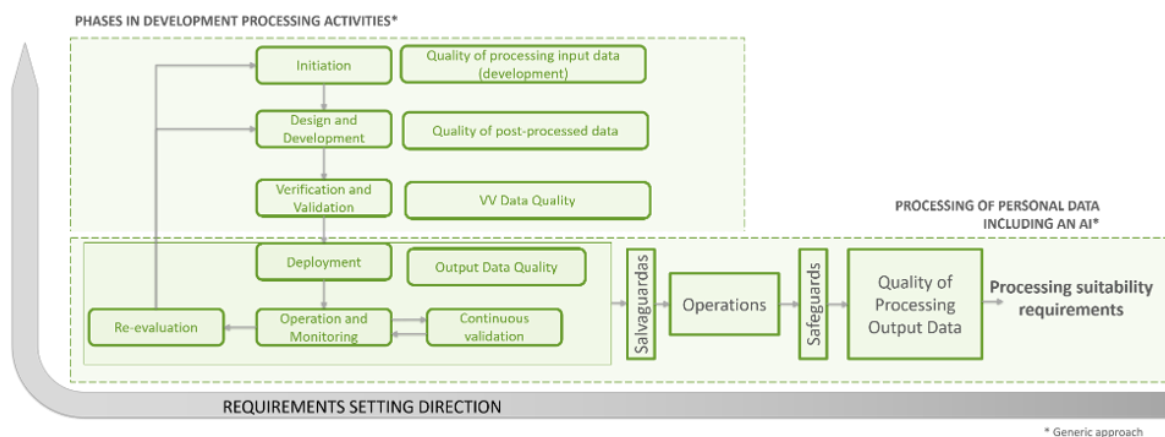


Figure 13 The quality of the AI outcome data will be a requirement that will be derived from the necessary quality of the results in the intended processing.

It might also be necessary to know in advance or establish as a requirement for deployment of the AI system, the safeguards to be included in the processing and performance of other systems, AI or non-AI, that will interact with the former.

Data quality is not a static requirement set in the design phase. It must be understood as a continuous management process throughout the life cycle of the AI model or system and, also, the life cycle of personal data processing. This requires monitoring, reviewing and, where appropriate, updating datasets and the systems that use them, especially when changes may occur in the context, in the input data or in the behaviour of the system (data governance).

F. QUALITY ANALYSIS OF DATA SOURCES FOR AI

For the development of AI systems, access to datasets of sufficient quality is necessary. "Access", as defined by the DGA in its article 2(13) is "data **use**, in accordance with **specific technical, legal or organisational requirements, without necessarily implying the transmission or downloading of data**".

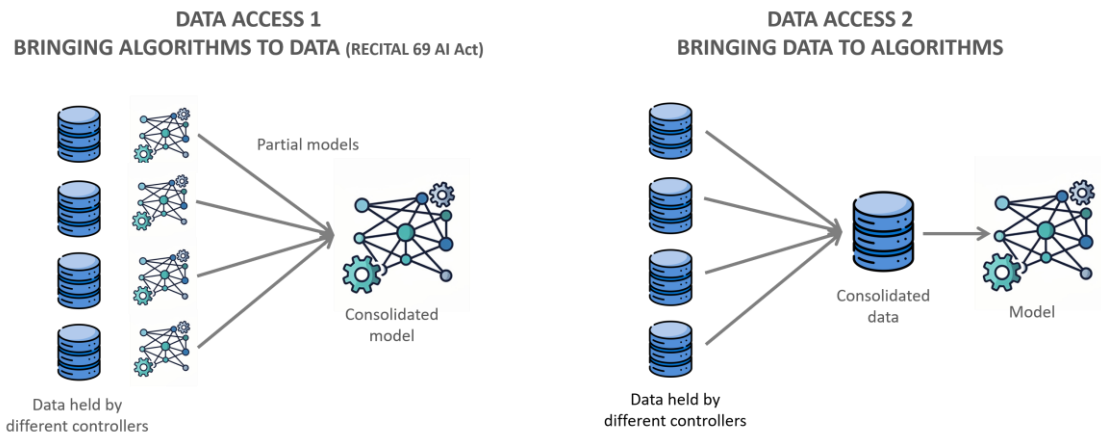


Figure 14 On the left is the access model based on bringing algorithms to data "compute to data", which in AI system development is called federated learning. Its advantages are not limited to data protection, but also to guaranteeing data sovereignty. In some cases, this access model is mandatory, for example, in the case of secondary uses in the European Health Data Space (secure processing environment).

Data access can be a stand-alone processing – for example, for the creation of data catalogues – or integrated as an operation within a processing operation for the development of an AI system.

Before initiating a process of access or mass data collection, the legitimacy of such access must be assessed. To this end, it must be determined, among other things, whether the data sets are **necessary**, i.e. whether they are **adequate, relevant and limited** to what **is necessary** in relation to the **purposes**, as established by the principle of minimisation-GDPR. This implies that an assessment of the quality of the data at source must be made and assess whether they are of sufficient quality, on their own or in conjunction with other datasets.

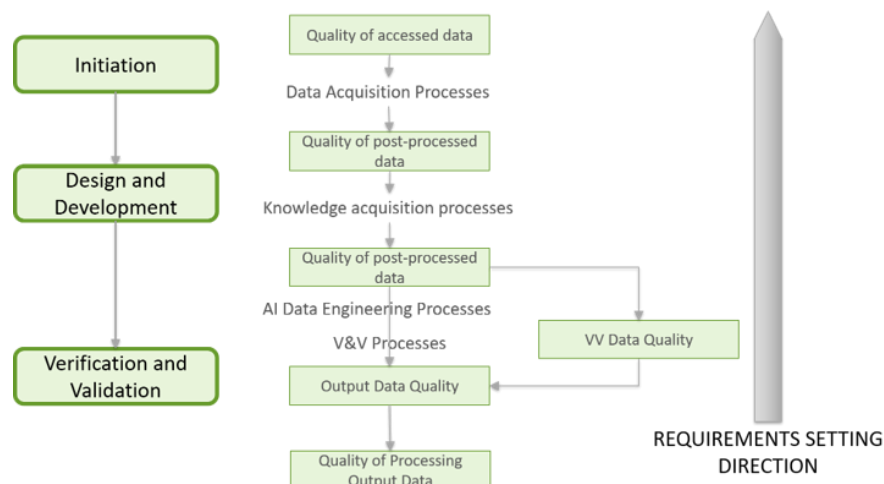


Figure 15 Determining the performance requirements of the AI system in the intended contexts of use is the starting point for assessing the necessary quality of the data to be accessed.

This analysis could be done both in the context of access to data for the development of AI systems or models, and for the execution of any other processing.

G. INADEQUATE QUALITY REQUIREMENTS

Defining **data quality requirements** based on **objective criteria and verifiable evidence** is an essential element for the development of AI systems and models. The establishment of

insufficient quality requirements could **jeopardise the suitability** of the development processing by preventing the system finally obtained from achieving the performance or satisfying the design requirements envisaged.

On the other hand, identifying **quality requirements that are too high** not only **contradicts the principle of minimisation-GDPR** by implying access **to a larger volume of data** or to data with greater extension, granularity, precision or frequency than needed. In addition, **it would make access** to datasets, **data governance process** management²², and **other AI system development processes** more expensive. In the worst case, this **could lead to the discarding** of business or research initiatives on the erroneous assumption that a dataset of sufficient quality is not available or that it is not possible to obtain it (for other legal, economic or commercial reasons).

IV. FINAL THOUGHTS

In conclusion, the following remarks are formulated on data quality and the principle of GDPR-accuracy in processing, both where AI systems and models are deployed and where processing is carried out for their development or evolution:

- When the result of a processing operation consists of a decision, profiling, inference or enrichment of information relating to an identified or identifiable natural person, this result will be considered personal data. Consequently, it must comply with the principles of accuracy and minimisation provided for in the GDPR, in relation to the purpose of the processing, guaranteeing in all cases respect for the rights and freedoms of natural persons²³.
- Personal data used as input in processing must also comply with the principles of accuracy and minimisation provided for in the GDPR, to the extent that this is necessary to ensure compliance with the purpose of the processing and respect for the rights and freedoms of natural persons when the result of such processing affects them.
- The principle of accuracy provided for in the GDPR requires that the personal data used as input in a processing operation meet the degree of quality necessary fit for the purpose of the processing. This may require different levels of accuracy, currency, precision, granularity, frequency, or other attributes. Consequently, the principle of accuracy is not limited to requiring the veracity and updating of the data, but these must be guaranteed only to the extent that they are necessary to the purposes for which the personal data are processed, particularly insofar as they may affect natural persons, but not in all cases.
- In accordance with the principle of minimisation provided for in the GDPR, only input personal data may be processed if they are adequate, relevant and limited to what is necessary for the purpose of the processing. Accordingly, the GDPR does not legitimise the processing of personal data with a degree of accuracy, granularity, frequency, or other characteristics that exceed what is necessary for the fulfilment of that purpose.
- The principles of accuracy and minimisation provided for in the GDPR delimit the range of quality that personal data must meet in order to ensure compliance with the purpose of the processing and the proper performance of the different operations that make it up.
- The data quality requirements and the principle of accuracy provided for in the GDPR are closely related, but not equivalent, concepts. Data quality has a broader scope, as it also extends to non-personal data and includes requirements that go beyond compliance with data protection regulations.

²² For example, of the 22 processes that, for the government, management and management of data quality, are developed in the UNE 0080:2023 standard.

²³ Although it has already been pointed out above, in particular, when the result is related to an identified or identifiable natural person and in relation, among others, to compliance with the principles of accuracy and minimisation.

- In the context of the processing of personal data, the non-personal data used must also meet the appropriate quality requirements, insofar as their use may affect the suitability of the processing in relation to its purposes, or the results of the processing when these may affect the rights and freedoms of natural persons.
- It must be assumed that there is no perfect processing. Deficiencies in the collection of input data, errors in processing operations, design deficiencies or other incidents can degrade the quality and suitability of the results obtained. Therefore, it is necessary to determine the minimum acceptable degree of quality that the results of the processing must meet in order to fulfil its purpose, to define metrics that allow it to be evaluated and to incorporate into the design of the processing monitoring, control and incident management mechanisms that guarantee the suitability of said processing.
- There is also no perfect system for implementing processing operations, particularly when AI systems are employed. However, there is an AI system suitable for its deployment in a processing operation when its strengths and weaknesses are known and duly documented, and the latter can be effectively managed by designing and implementing the necessary safeguards in said processing.
- In a processing operation, ignoring the use of quality metrics of the output prevents guaranteeing and being able to demonstrate²⁴ whether the purposes of the processing are really being fulfilled effectively and appropriately.
- The determination of the minimum level of data quality necessary for the result of the processing to be suitable allows rational decisions to be made on the design of the processing, its different operations, the relationship between them, the selection of the most appropriate systems for its execution according to their performance and the safeguards to be implemented. In particular, in relation to AI systems.
- For machine learning processing using personal data, the dataset used must be of GDPR-compliant quality and accuracy to enable the development of an AI system that meets the performance requirements for deployment.
- In machine learning, the quality of the dataset, considered in aggregate, is more relevant than the quality characteristics of each individual piece of data.
- In order to determine the quality of the dataset necessary for the development or evolution of an AI system, the pre-processing of the data, the development methodology, the type of deployment, the context of use and the safeguards envisaged for its integration into the processing should be taken into account, among other factors.
- The quality of the outputs of an AI system, which is necessary for it to meet the performance requirements when deployed in a processing operation, is different from the quality required of the data used during its development. Consequently, both must be evaluated differently.
- Datasets that do not meet the quality necessary for the development or evolution of an AI system or model shall not be required for such processing. Consequently, access to them will not be justified, except to the extent strictly necessary to assess their quality.
- The establishment of excessive data quality requirements for processing for the development of AI systems can compromise compliance with the principle of minimisation set out in the GDPR, unnecessarily increase the costs of development processing and, in the worst case, lead to the cancellation of research initiatives or projects.
- The organisation's data governance, as well as the management processes derived from it, must integrate the governance requirements applicable both to the processing of personal data and to the processes of development, acquisition and evolution of AI systems.

²⁴ Article 24 of the GDPR

- Data governance processes will need to clearly define who and how data quality requirements will be determined in the responsible organisation. Those requirements shall be objectively defined and ensure compliance with the principles of accuracy and minimisation set out in the GDPR.
- The quality of the data must be continuously monitored throughout the life cycle of the processing, through the incorporation of mechanisms for periodic review and adaptation to changes in the context, purposes, scope and risks of the processing, as well as to variations in the behaviour of the systems used.
- In the design of a processing operation, the documentation of data quality requirements as well as compliance with the principles of accuracy and minimisation set out in the GDPR is an essential element of the principle of accountability. In the case of processing for the development of AI systems, such documentation shall take into account the foreseeable uses of the systems.
- The controller, when selecting the AI systems to be used to implement processing operations, shall take into account the requirements documented by the provider in relation to data quality and, where applicable, the safeguards and other aspects that must be incorporated into the design of the processing to ensure adequate deployment.
- A DPO or a data protection adviser must know and understand the implications of all data flows and operations that occur in AI model and system development processing, as well as those that occur in processing involving AI. This knowledge will enable the DPO and the adviser to provide solutions and contribute to resolving dilemmas that may hinder both innovation and the protection of fundamental rights.
- In the development of data-driven AI systems and models, it is essential to have a team that incorporates at least data scientists and data protection specialists. This team must establish data quality requirements based on objective criteria and verifiable evidence throughout the entire life cycle of personal data processing. In this way, it helps to ensure both the success of the projects and the protection of fundamental rights and freedoms.

Finally, it should be noted that this technical note is related to other material published by the AEPD on artificial intelligence and data protection, among others:

- [Agentic Artificial Intelligence from the perspective of Data Protection](#) [feb 2026]
- [Audit Requirements for Personal Data Processing Activities Involving AI](#) [jan 2021]
- [GDPR compliance of processing that embed Artificial Intelligence. An introduction](#) [feb 2020]

V. ANNEX: EVOLUTION OF THE CONCEPT OF ACCURACY IN SPANISH LEGISLATION

Directive 95/46/EC, regarding to personal data, established in its first section, “*Principles relating to data quality*” the concept of data quality that data should be:

*c) adequate, relevant and not excessive **in relation to the purposes** for which they are collected and/or further processed;*

*d) **accurate and, where necessary, kept up to date**; every reasonable step must be taken to ensure that data which are inaccurate or incomplete, having regard to the purposes for which they were collected or for which they are further processed, are erased or rectified;*

The Directive linked the concepts of accuracy and minimisation to be interpreted in the context of the purpose of the processing. Spanish Organic Law 15/1999 of 13 December on the Protection of Personal Data (LO 15/1999), which transposed the Directive, in TITLE II "Principles of data protection", in its article 4 "Data quality", enunciated a series of principles, close to those established in article 5(1) of the GDPR. In particular:

1. Los datos de carácter personal **sólo** se podrán recoger **para su tratamiento**, así como someterlos a dicho tratamiento, **cuando sean adecuados, pertinentes y no excesivos en relación con el ámbito y las finalidades** determinadas, explícitas y legítimas para las que se hayan obtenido [Personal data may **only** be collected **for processing**, as well as subjected to such processing, **when personal data are adequate, relevant and not excessive in relation to the scope and** the determined, explicit and legitimate purposes for which they are collected]

3. Los datos de carácter personal serán **exactos** y puestos al día de forma que **respondan con veracidad a la situación actual del afectado** [The personal data will be **accurate** and kept up to date in such a way that **they truthfully respond to the current situation of the data subject**]

The wording of LO 15/1999 disassociated the concept of data accuracy from that of the need to the purposes for which the personal data are processed, defining the accuracy of personal data in absolute terms in relation to the veracity of the current situation of the data subject. From this wording it could be concluded:

- That the data could be considered accurate, even if it affected the purpose of the processing.
- And that in order for the data to be considered accurate, it had to correspond truthfully to the data subject's current situation, even if this circumstance did not affect the purpose of the processing and, therefore, it was not necessary.

The GDPR does not refer to the term "*data quality*",²⁵ although it does establish the principles of minimisation and accuracy in its articles 5(1)(c) and (d):

"1. Personal data shall be:

c) **adequate, relevant and limited** to what is necessary **in relation to the purposes** for which they are processed («data minimisation»);

d) **accurate** and, where **necessary, kept up to date**; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes **for which they are processed**, are erased or rectified without delay («accuracy»);".

The GDPR aligns with the definitions of the Directive by linking minimisation and accuracy to the purpose of the processing.

In the LOPDGDD, in its article 4, the wording is different, although making explicit reference to article 5(1)(d):

Artículo 4. Exactitud de los datos

1. Conforme al artículo 5.1.d) del Reglamento (UE) 2016/679 los datos serán **exactos** y, si fuere necesario, actualizados.

2. A los efectos previstos en el artículo 5.1 d) del Reglamento (UE) 2016/679, no será imputable al responsable del tratamiento, siempre que éste haya adoptado todas las medidas razonables para que se supriman o rectifiquen sin dilación, **la inexactitud de los datos personales, con respecto a los fines para los que se tratan**, cuando los datos inexactos:

a) Hubiesen sido obtenidos por el responsable directamente del afectado.

b) Hubiesen sido obtenidos por el responsable de un mediador o intermediario en caso de que las normas aplicables al sector de actividad al que pertenezca el responsable del tratamiento establecieran la posibilidad de intervención de un intermediario o mediador

²⁵ Except in article 47 "Binding Corporate Rules" in paragraph 2 letter d) "The Binding Corporate Rules referred to in paragraph 1 shall specify at least: d) the application of the general data protection principles, in particular purpose limitation, data minimisation, limited storage periods, data quality, data protection by design and by default, ...".

que recoja en nombre propio los datos de los afectados para su transmisión al responsable. El mediador o intermediario asumirá las responsabilidades que pudieran derivarse en el supuesto de comunicación al responsable de datos que no se correspondan con los facilitados por el afectado.

c) Fuesen sometidos a tratamiento por el responsable por haberlos recibido de otro responsable en virtud del ejercicio por el afectado del derecho a la portabilidad conforme al artículo 20 del Reglamento (UE) 2016/679 y lo previsto en esta ley orgánica.

d) Fuesen obtenidos de un registro público por el responsable.

[Article 4. Accuracy of data

1. In accordance with Article 5(1)(d) of Regulation (EU) 2016/679 the data shall be accurate and, if necessary, updated.

2. For the purposes provided for in Article 5(1)(d) of Regulation (EU) 2016/679, the controller shall not be held liable for the inaccuracy of personal data, with respect to the purposes for which they are processed, provided that the controller has taken all reasonable steps to ensure that such data are erased or rectified without delay, where the inaccurate data:

a) Have been obtained by the controller directly from the data subject.

b) Had been obtained by the controller from a mediator or intermediary in the event that the rules applicable to the sector of activity to which the controller belongs provide for the possibility of intervention of an intermediary or mediator who collects the data of the data subjects on their own behalf for transmission to the controller. The mediator or intermediary shall assume the responsibilities that may arise in the event of communication to the controller of data that do not correspond to those provided by the data subject.

c) Were subject to processing by the controller because they were received from another controller by virtue of the exercise by the data subject of the right to portability in accordance with Article 20 of Regulation (EU) 2016/679 and the provisions of this Organic Law.

d) Were obtained from a public record by the controller].

The LOPDGDD requires accuracy without defining what accuracy is, although the reference to 5(1)(d) should be understood to be linked to the fulfilment of the purposes. However, perhaps due to the cognitive resistance bias²⁶, in many cases the concept defined in the framework of L.O. 15/99 that "accuracy" means to respond truthfully to the current situation of the data subject is dragged along.

Such an interpretation could affect:

- To the rights of the data subjects, insofar as it requires data that could go beyond the needs of the processing, contradicting the principle of minimisation.
- To the rights of the data subjects, since the wording of the LOPDGDD would allow the use, in certain cases, of data that are not appropriate for the purpose of the processing and, therefore, contradict the provisions of article 5(1)(d) of the GDPR.
- Controllers and processors, and in particular developers and organisations responsible for the evolution of AI systems, by demanding levels of veracity and timeliness that may not be necessary, could go against the quality of the data itself to develop systems, and in addition to demanding the attention of rights beyond what is established in the GDPR²⁷.

²⁶ Or perseverance bias, the resistance to accept a change in ideas even if there is evidence for it.

²⁷ Art. 16 of the GDPR also links rectification to the purpose of the processing.