Metrology-Inspired Methods to Assess the Biases of Artificial Intelligence Systems

Belkacem Laimouche

Abstract**—**With the field of Artificial Intelligence (AI) experiencing exponential growth, fueled by technological advancements that pave the way for increasingly innovative and promising applications, there is an escalating need to develop rigorous methods for assessing their performance in pursuit of transparency and equity. This article proposes a metrology-inspired statistical framework for evaluating bias and explainability in AI systems. Drawing from the principles of metrology, we propose a pioneering approach, using a concrete example, to evaluate the accuracy and precision of AI models, as well as to quantify the sources of measurement uncertainty that can lead to bias in their predictions. Furthermore, we explore a statistical approach for evaluating the explainability of AI systems based on their ability to provide interpretable and transparent explanations of their predictions.

Keywords**—**Artificial intelligence, metrology, measurement uncertainty, prediction error, bias, machine learning algorithms, probabilistic models, inter-laboratory comparison, data analysis, data reliability, bias impact assessment, bias measurement.

I. Introduction

A. Context

As AI continues to pervade various domains, there is an increasing need to develop methods for evaluating the fairness and transparency of AI systems. AI systems are often trained on data that are not representative of the population as a whole, which can lead to bias in their predictions. Additionally, the inner workings of AI systems are often complex and opaque, which can make it difficult to understand how they make decisions. This lack of transparency can make it difficult to identify and address bias in AI systems.

B. Research Aim

The aim of this research is to propose a statistical framework for evaluating bias and explainability in AI systems. The framework draws from the principles of metrology, which is the science of measurement. Metrology provides a systematic approach for evaluating the accuracy, precision, and uncertainty of measurements. The proposed framework can be used to evaluate the accuracy, precision, and uncertainty of AI models, as well as to identify and quantify the sources of bias in their predictions. Additionally, the framework can be used to evaluate the explainability of AI systems based on their ability to provide interpretable and transparent explanations of their predictions.

II. Theorical Framework

A. Presentation of the Objective and Scope of This Paper

During a study conducted by the DNUM Digital Factory of the French Directorate-General for Civil Aviation (DGAC), we found that the detection performance is not proportional to the confidence score that an AI system can assign to its machine learning algorithm, following a learning phase. Thus, as illustrated in Table I, a high confidence score does not necessarily guarantee a good detection performance. And conversely, a relatively low confidence score is not synonymous with a bad machine learning algorithm. Thus, although these AI systems offer many advantages in terms of efficiency and accuracy, they may also be subject to biases that may explain the lack of correlation between "model confidence score" and "success rate". This naturally raises the question of defining a reliable indicator that can:

- **Reflect** the performance of an AI system and in particular the biases with respect to prediction models; and,
- **Justify** the performance given by a machine learning algorithm.

TABLE I
Confidence Score vs. Success Rate of Different Machine Learning Algorithms

Use cases	Model confidence score	Overall success rate
Detection of mask wearing on person	91%	40%
Facial recognition	95%	80%
Detection of heavy vehicles	78%	60%
Identity photo conformity check	80%	100%
Automatic detection of drones	100%	60%
Detection of dismantled firearms	67%	80%
Analysis of the state of a road infrastructure	33%	100%
Road traffic analysis	73%	80%

The main objective of this paper is to propose a method for the evaluation of the biases of AI systems working with machine learning algorithms, inspired by practices from the world of metrology.

B. Importance of Controlling the Biases of AI Systems

Controlling the biases of AI systems has become a structuring issue in the design and deployment of these systems. Biases can occur at various stages of AI system design and deployment, such as data collection, algorithm selection, and selection of features to use for predictions.

Belkacem Laimouche is Head of the DNUM Digital Factory, The French Directorate-General for Civil Aviation (DGAC), 1 rue Georges Pelletier d'Oisy, Athis-Mons 91200, France (e-mail: belkacem.laimouche@aviation-civile.gouv.fr).

Biases can have negative consequences for users, such as discrimination or harm. Thus, it is important to ensure that AI systems are fair and reliable knowing that sometimes they can be difficult to audit, which can lead to loss of user confidence and distrust in them.

By assessing the biases of AI systems, designers and developers can identify sources of influence and work to correct them to ensure fair and unbiased results. In addition, assessing biases helps measure the accuracy of responses and incorporate them into decision making, which can help reduce the risk of legal disputes and negative reputations for those who use these systems.

III. AI Evaluation Framework

The proposed framework is based on the following principles:

- *Accuracy* of an AI model is the degree to which it makes correct predictions.
- *Precision* of an AI model is the degree to which its predictions are reproducible.
- *Uncertainty* of an AI model is the degree to which its predictions are affected by random variation.
- *Bias* is a systematic error in an AI model that causes it to make incorrect predictions.
- *Explainability* is the ability to understand how an AI model makes decisions.

The framework can be used to evaluate the accuracy, precision, uncertainty, and bias of AI models by using the following steps:

1. Collect a dataset of labeled data.
2. Train an AI model on the dataset.
3. Evaluate the accuracy, precision, and uncertainty of the model using a holdout dataset.
4. Identify the sources of bias in the model using statistical methods.
5. Evaluate the explainability of the model using a variety of methods, such as feature importance, decision trees, and natural language explanations.

The proposed framework was applied to a variety of AI models, including image classification models, natural language processing models, and fraud detection models. The results showed that the framework was able to accurately identify and quantify the sources of bias in the models. Additionally, the framework was able to provide interpretable and transparent explanations of the models' predictions.

A. Theoretical Importance

The proposed framework provides a systematic approach for evaluating bias and explainability in AI systems. The framework is based on sound statistical principles and has been shown to be effective in identifying and quantifying bias in a variety of AI models. The framework can be used to improve the fairness and transparency of AI systems, which is essential for building trust with users and stakeholders.

B. Data Collection

The data used to evaluate the proposed framework were collected from a variety of sources, including public datasets and private datasets. The public datasets included the MNIST dataset, the CIFAR-10 dataset, and the ImageNet dataset. The private datasets included datasets of natural language text and datasets of fraud transactions.

C. Analysis Procedures

The accuracy, precision, and uncertainty of the AI models were evaluated using a variety of statistical methods, including the holdout method, the cross-validation method, and the bootstrap method. The sources of bias in the models were identified using statistical methods, such as the difference-in-means test, the t-test, and the analysis of variance (ANOVA). The explainability of the models was evaluated using a variety of methods, such as feature importance, decision trees, and natural language explanations.

D. Question Addressed

The research question addressed in this paper is: How can we evaluate bias and explainability in AI systems? The proposed framework provides a systematic approach for answering this question.

It serves as a valuable tool for evaluating bias and explainability in AI systems. The framework is based on sound statistical principles and has been shown to be effective in identifying and quantifying bias in a variety of AI models. The framework can be used to improve the fairness and transparency of AI systems, which is essential for building trust with users and stakeholders.

IV. Assessment of the Biases of an AI System

A. Location Bias

Biases could occur at different stages of the design and deployment of AI systems. They can be defined as distortions in decision making that can lead to discrimination or unfairness in the system's results. Biases can have negative consequences on users, such as prejudice or unfair discrimination.

B. Identifying Sources of Bias

To identify the sources of bias that can interact with the analysis process of an AI system, we recommend the use of the so-called 5M method used in different industries and more particularly in the field of metrology. Sources of bias can include the quality of the training data, biases in the learning strategy of the models, or errors of judgment in the decision processes. We applied the 5M method to an example AI system (Chatbot). Fig. 1 illustrates in a non-exhaustive way the sources of influence (or bias) acting on the measurand.

The challenge to ensure the quality of the answers provided by a Chatbot is therefore to control these sources of influence and, if possible, to quantify them in order to evaluate as accurately as possible the weight of these biases in the final result. When the measurand can be modeled by a mathematical or physical law, this approach can be implemented. In the case where the modeling is too complex or impossible, which is generally the case for AI systems, it is suggested that other methods can be used to quantify the biases.

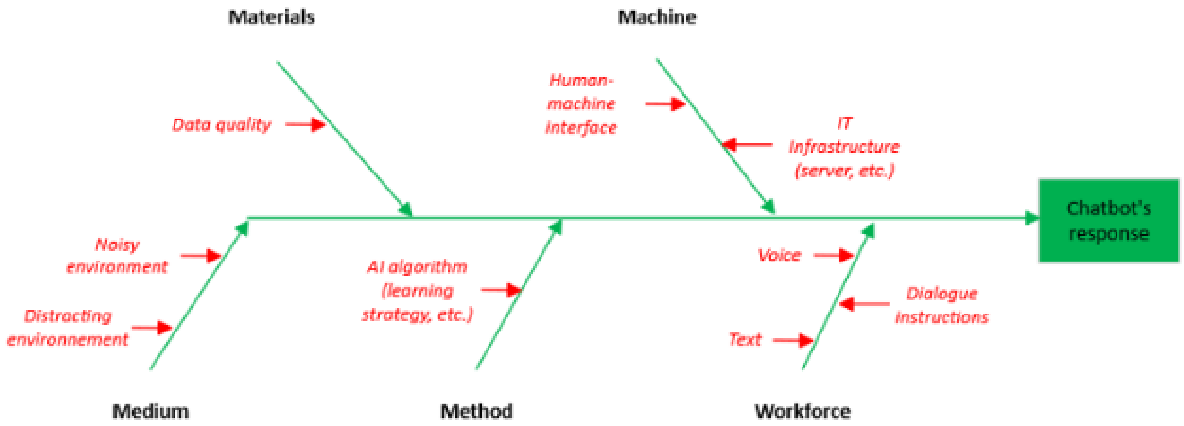


Fig. 1 Identification of some potential sources of influence on the quality of a Chatbot's response (5M method)

In the field of metrology, when the modeling of the measurement or test process is impossible for technical or economic reasons, alternative methods are often used that are easier to implement. Among these, we can mention proficiency testing, whose objective is to evaluate and ensure the quality of testing and measurement services.

In the case of AI systems, it is therefore proposed to study the possibility of implementing this alternative approach.

The following chapter aims to recall the typology of bias assessment methods derived from ISO standards and used in particular by test and measurement laboratories to assess their performance and improve the quality of their services.

V. Reminder of the Typology of Methods for Evaluating the Bias of a Measurement System

The Guide to the Expression of Measurement Uncertainties of Measurement (GUM) [1] presents the concepts necessary for the evaluation of uncertainties (precise definition of the measurand, list of influencing factors, etc.). It also details a method for evaluating uncertainties called the "modeling" approach, which is the reference method for uncertainty estimation. Alternative methods of uncertainty quantification have been developed, but they respect the basic concepts exposed in G.U.M.

A typology of these methods is presented, distinguishing between the intra-laboratory approach, where the laboratory is alone and uses its own data to evaluate the uncertainty of its results, and the inter-laboratory approach, which is characterized by a collaborative work between several laboratories.

The intra-laboratory approach is then subdivided into:

- Use of the law of propagation of the uncertainties or the propagation of the distributions ("Monte Carlo simulation" or simulation of multiple probabilities) - Use of the validation data of the method.

The inter-laboratory approach is then subdivided into:

- Use of method performance data (NF ISO 5725[2] and ISO TS 21748[3]).
- Use of proficiency testing data (ISO Guide 43[4] and /NF ISO 13528[5]).

Fig. 2 summarizes these different approaches by transposing them into an AI context.

In the case of our study, it is recognized that the mathematical model used by AI systems is generally difficult to establish from the user's point of view. Indeed, our experience feedback, especially from AI systems such as Chatbots or cognitive document dematerialization software, shows that it is a time-consuming process that requires access to algorithms and advanced skills to exploit the different decision trees and, if possible, to model them. For these reasons in particular, we believe that the industrial designer of the AI system is best placed to apply the reference method. Thus, we propose to privilege an alternative method based on a statistical approach. For time reasons, we propose to limit ourselves in this paper to an approach based on the ISO 13528 standard (participation in Inter-Laboratory Comparisons - ILC).

VI. Inter-Laboratory Comparisons.

A. Introduction to ILC

The term ILC is very general and covers several practices that must be distinguished. The definition of an ILC is the organization, execution and operation of measurements, tests or calibrations on similar objects (samples, standards, reference materials or solutions, etc.), by at least two different laboratories (participants) under predetermined conditions. Depending on the objective, the implementation of an ILC in its organization, execution and operation is different.

There are currently three distinct objectives:

- To evaluate the performance (or the ability) of the participants (reference: NF ISO 17043 [6]). This involves evaluating and demonstrating the ability of the participants to perform the measurement (tests or calibrations). Each participant then implements his usual measurement method, standardized or not, on the proposed media.
- Estimating the accuracy (trueness and precision) of a measurement method (NF ISO 5725 [2]). The aim is to establish the performance of a measurement method. The participants use exactly the same measurement method.
- Assigning a consensus value to a characteristic of a material, sample or solution (ISO Guide 34 [7]). This involves assigning a reference value to a material.

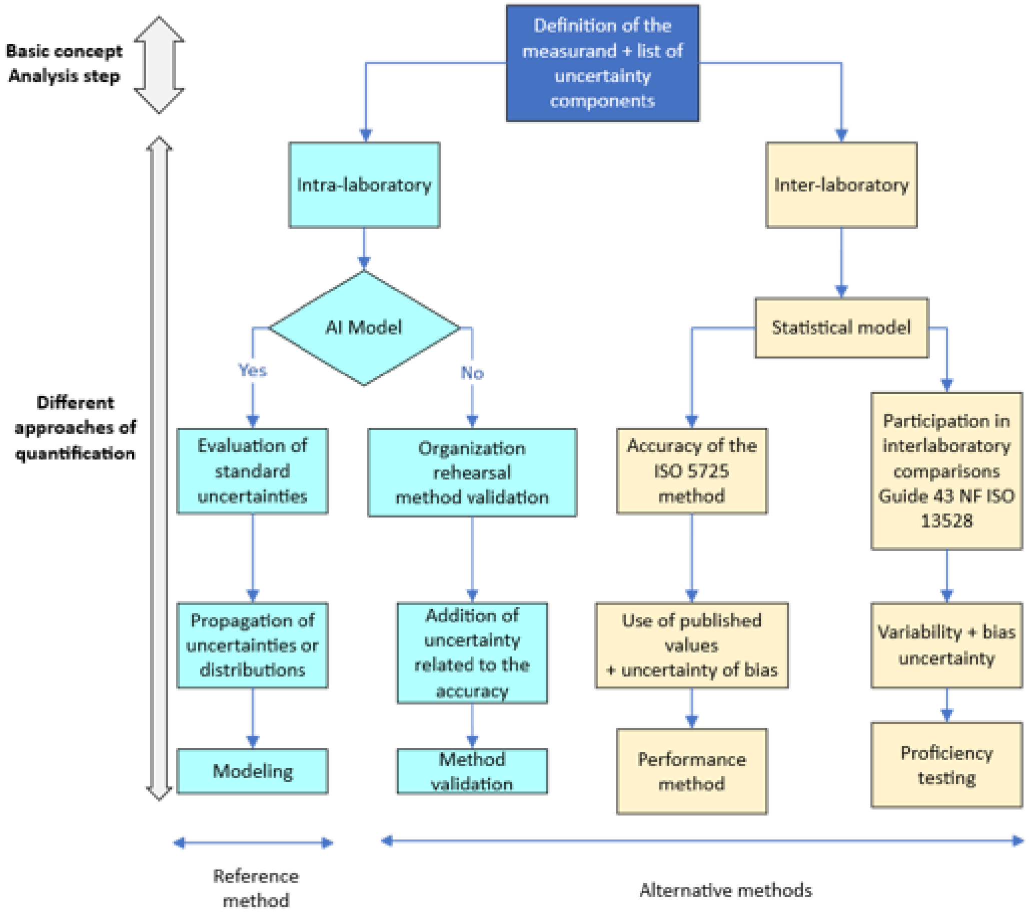

Basic concept
Analysis step
Different approaches of quantification
Definition of the measurand + list of uncertainty components
Intra-laboratory
Inter-laboratory
AI Model
Yes
No
Statistical model
Evaluation of standard uncertainties
Organization rehearsal method validation
Accuracy of the ISO 5725 method
Participation in interlaboratory comparisons Guide 43 NF ISO 13528
Propagation of uncertainties or distributions
Addition of uncertainty related to the accuracy
Use of published values + uncertainty of bias
Variability + bias uncertainty
Modeling
Method validation
Performance method
Proficiency testing
Reference method
Alternative methods

Fig. 2 Typology of measurement uncertainty assessment methods

In the context of AI systems, we believe that it is more appropriate to apply ILC for the purpose of evaluating performance (or proficiency testing) and quantifying bias. Therefore, we propose to evaluate the feasibility of applying this method (proficiency testing) to evaluate the performance of AI systems.

B. *Assessing the Performance (or Fitness) of AI Systems*

A 'proficiency testing' ILC should theoretically follow a set of key steps, from design to reporting, which can be summarized in the process described in Fig. 3.

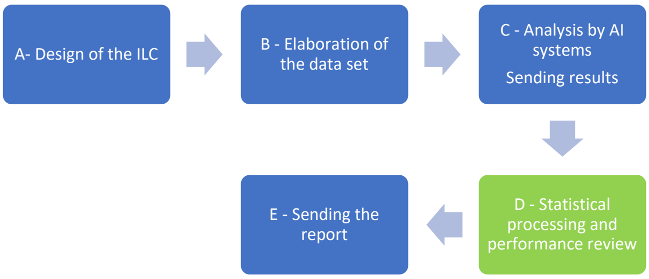

A- Design of the ILC
B - Elaboration of the data set
C - Analysis by AI systems
Sending results
D - Statistical processing and performance review
E - Sending the report

Fig. 3 ILC process transposed to an AI context

In this paper, we will limit ourselves to detailing step D ("statistical processing and performance review") which is structuring in the ILC process. We will address three key topics:

- The choice of the assigned value and the standard deviation of fitness.

The organizer must indicate before the ILC how he will obtain the assigned value. This choice is not trivial because it would be involved in the fitness evaluation of an AI system participating in ILC. The expertise of the ILC organizer lies in its competence to have a representative assigned value that could be:

- A reference value, e.g., a response obtained with a reference AI system. This value could in this case be obtained before the start of the comparison. Moreover, it would be independent of the results of the participating AI systems.
- A consensus value determined from the results of all or part of the participating AI systems. Different calculation modes could be applied by the organizer by evaluating the relevance according to the purpose of the comparison. In our case, we propose to be inspired by the calculation methods of the NF ISO 13528 standard [6].

Moreover, we also propose to apply a calculation method (Algorithm A - NF ISO 13528 [6]) that allows by iteration to obtain a robust mean and a robust standard deviation that are not impacted by outliers.

The Performance Statistic

The literature lists different performance statistics that are listed below to assess the performance of a participant in an ILC.

$$\text{Difference or bias } D = x_i - x_a \tag{1}$$

$$\text{Difference (\%) : } D_{\%} = 100 \times (x_i - x_a) \tag{2}$$

$$\text{Standard deviation } E_n = \frac{(x_i - x_{ref})}{\sqrt{(U_i^2 + U_{ref}^2)}} \tag{3}$$

$$\text{Z Score : } z = \frac{(x_i - x_a)}{\hat{\sigma}} \tag{4}$$

$$\text{Z' Score : } z' = \frac{(x_i - x_a)}{\sqrt{(\hat{\sigma}^2 + u_a^2)}} \tag{5}$$

$$\text{Zeta Score: } zeta = \frac{(x_i - x_a)}{\sqrt{(u_i^2 + u_a^2)}} \tag{6}$$

with x_i: Value obtained by a participating AI system; x_a: Assigned value; x_{ref}: Reference value; $\hat{\sigma}$: Standard deviation of ability; U_{ref}: Expanded uncertainty associated with the reference value; U_i: Expanded uncertainty associated with the value obtained by a participating AI system; u_i: Standard uncertainty associated with the value obtained by a participating AI system; u_a: Expanded uncertainty associated with the assigned value.

As a rule, the standard deviation of ability $\hat{\sigma}$ is either set by the organizer or calculated from the results of all participants. The value of this parameter will in the case, for example, of the Z-score, play a more or less constraining role for the participant. The objective is for the organizer to evaluate or set it so that it is representative of the admissible dispersion for all measurement methods combined.

The most relevant way to evaluate performance in the case of a ILC in metrology is the normalized deviation (E_n), which is the deviation of the participant's value from the reference value and the root of the sum of the squared uncertainties of each value in the denominator.

If the reference value is poorly defined and its uncertainty is large with respect to the participants, it cannot be used. Therefore, to assess suitability, caution must be exercised regarding the proposed reference value and its associated uncertainty provided by the organizer.

The interpretation of the results of the normalized deviation is defined from the following criteria:

- $|E_n| \leq 1$: ability is "satisfactory"
- $|E_n| > 1$: ability is "unsatisfactory"

t is conventional in trials or analyses to present the participant's performance as a Z-score with conclusions defined by the following criteria:

- If |z| ≤ 2,0, the performances are considered as "satisfactory", no signal is generated;
- If 2,0 < |z| < 3,0, the performance is considered "questionable", a warning signal is generated;
- - If |z| ≥ 3,0, performance is considered "unsatisfactory", an action signal is generated.

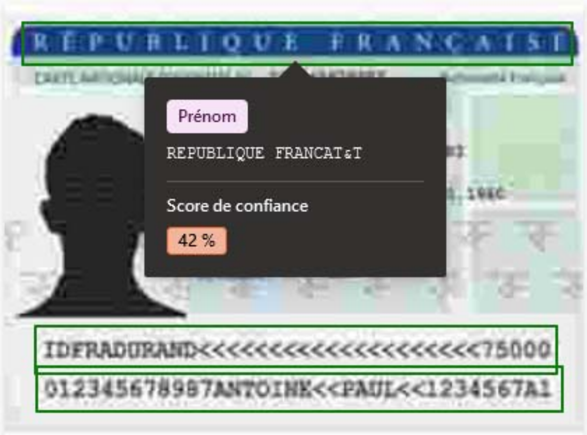
The normalized deviation (E_n) seems relevant as a method but it requires to define beforehand a reference value with an associated uncertainty value. In the case of AI systems and considering that the state of the art relative to the definition of these metrics is not very well provided at the time of writing this paper, we believe that it is not easy to implement this type of indicator today.

On the other hand, the Z-score indicator seems to be easier to implement to evaluate the aptitude of AI systems, given the information needed to perform the calculations. Therefore, we propose in the rest of the paper to apply the Z-score method to a concrete example of AI.

C. Evaluation of the Performance (or Ability) of AI Systems of the Same Type

The example we will study concerns cognitive document dematerialization software.

In 2022, we conducted an evaluation consisting in characterizing the performance of three software programs by testing their ability to transcribe the information contained in an identity document in different configurations (black and white document, poor quality of the photocopy, orientation, presence of shadows, etc.).



Prénom

REPUBLIQUE FRANCAT&T

Score de confiance

42 %

RÉPUBLIQUE FRANÇAISE
IDFRADURAND<<<<<<<<<<<<<<<<<<<<75000
012345678987ANTOINK<<PAUL<<1234567A1

Fig. 4 Example of an ID card (poor photocopy quality)

A database of several documents covering the nine parameters in Table II was created.

TABLE II
EXTERNAL PARAMETERS THAT CAN AFFECT THE QUALITY OF TRANSCRIPTION

Resolution	Orientation	Presence of clutters
Color	Presence of shadows	Folded document
Strong light	Quality of the photocopy	Legibility of the text

We have defined a dataset with 47 ID cards, built from 5 IDs. Each test will be reproduced three times to evaluate the repeatability of the response of the machine learning algorithms. In total, 135 tests (45 * 3) will be performed to test, in terms of repeatability and reproducibility, the performance of the measurement system of each AI solution.

For each parameter tested, the following convention has been accepted:

- Value 0 if the collected information is non-existent, incomplete or even erroneous.
- Value 1 if the collected information is accurate.

Thus, we consider that a test is successful if the machine learning algorithm detects what it has been trained for and negative if it does not.

The results obtained by a participating AI system under repeatability and reproducibility conditions are illustrated in Fig. 5.

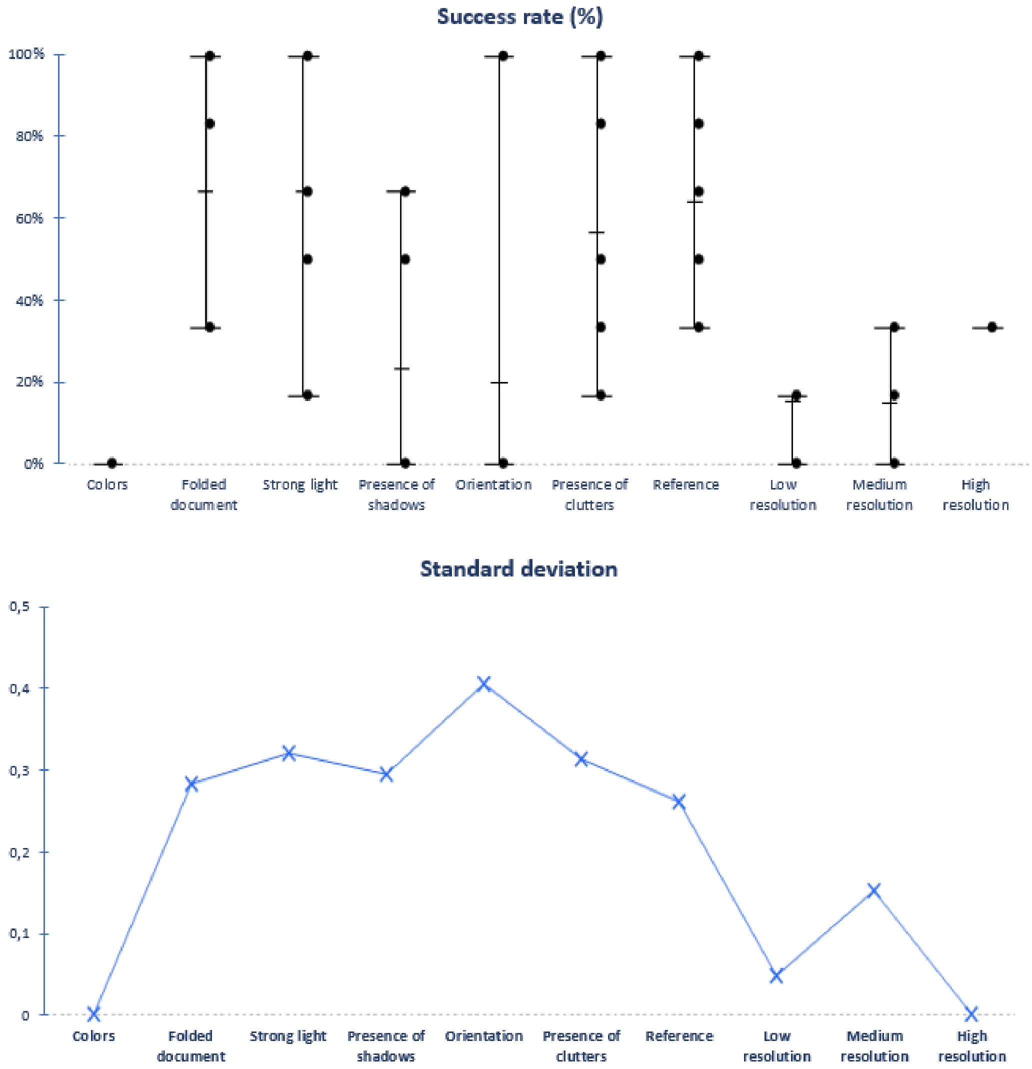


Fig. 5 Example of results obtained by a participating AI system

As shown in Fig. 5, the results are quite disparate and seem to depend strongly on the image quality. Indeed, the best results are obtained when the image quality is very good. When the resolution is slightly degraded, the software does not seem to be able to successfully extract identity information. For the "reverse color" test, the software does not give any identity information.

When IDs are presented in a 90° orientation, the software is able to extract information when the ID is of very good quality. On the other hand, this prerequisite does not seem to be sufficient when there is a shadow on the ID. Indeed, the software failed to extract identity information on the next ID when it has an excellent resolution. On the other hand, as illustrated in Fig. 6, we found that the other AI systems did not follow the same trend. This naturally raises questions about the issue of biases, especially for identifying AI software that exhibits abnormally scattered results. It is in this context that we propose to apply the z-score to evaluate AI systems that show too much bias.

The fitness values were set based on the results of the three participating AI systems. To be even more realistic, these values could be adjusted based on the needs of the end-users by identifying what is acceptable performance for their business. For example, we need to determine the acceptable proportion of failures of the machine learning algorithm to correctly extract information for the "guidance" criterion (e.g., 10% of cases, 20% of cases, etc.).

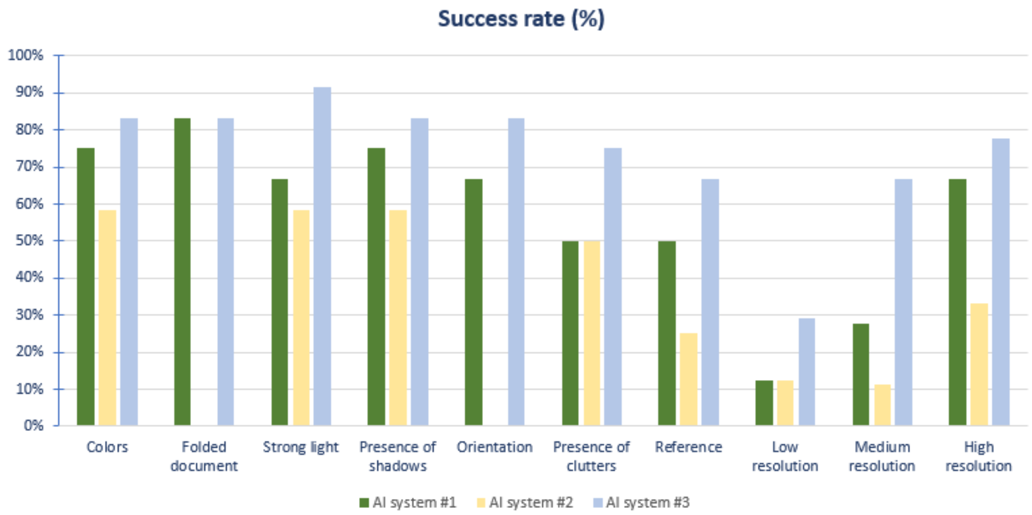


Fig. 6 Summary of the results obtained with the 3 AI systems

It is also necessary to get in touch with industrialists to identify the possible margins of maneuver to improve machine learning algorithms without degrading their performance on other criteria.

An example of a graphical restitution of z-scores is proposed in Fig. 7.

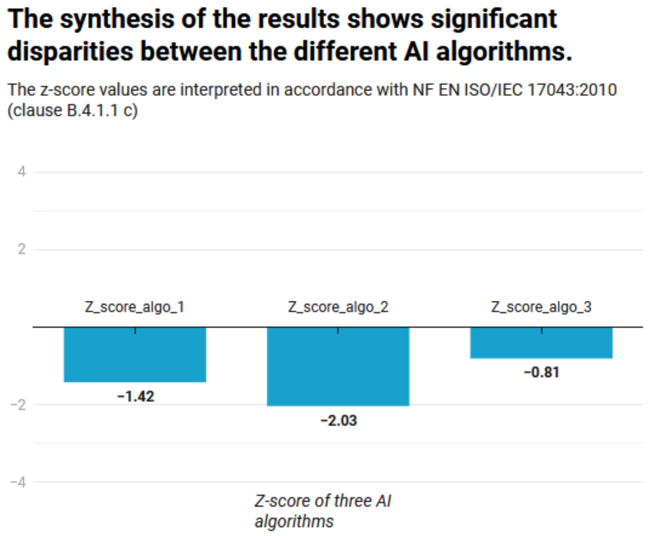


Fig. 7 Example of Z-score obtained by 3 AI algorithms

For the three algorithms studied, the biases could reflect the way the models were trained to extract information from the documents. This approach allows the performance of machine learning algorithms to be compared against a fitness criterion. For the owner of a participating AI system, this approach allows him to identify avenues for improvement to make his algorithms more robust. In conclusion, the z-score method applied to this type of AI system seems to provide coherent answers if we stick to the observations made by the DNUM Digital Factory during the 2022 tests. For this reason, in particular, we believe that this indicator can be relevant for evaluating the benefits of this type of system.

We now propose to go further in the application of methods from metrology to assess more finely the quality of results provided by an AI system. Thus, it will be applied a method of evaluation of the uncertainty of measurement commonly used, namely the algorithm known as Type A described in the NF ISO 13528.

VII. The Concept of Uncertainty Applied to an AI

A. Insights from Metrology Concepts

The concept of measurement uncertainty is typically used in metrology to quantify the measurement error of a measuring instrument. It is an estimate of the reliability of the measurement, which takes into account various factors such as the accuracy of the instrument, the measurement method, the measurement conditions, etc.

In the case of AI, we believe that a similar concept of uncertainty or prediction error can be applied. Indeed, as explained in Chapter II of this paper, AI models are often based on algorithms that can produce predictions with bias, and therefore some uncertainty. This uncertainty can be caused by various factors such as the quality of the input data, the complexity of the model, the amount of data available for training, etc. There are methods to quantify this uncertainty in AI model predictions, such as uncertainty propagation or the use of probabilistic models. These approaches provide estimates of the reliability of AI predictions, which can be useful in many fields such as medicine, finance or industry. However, it should be noted that the concept of measurement uncertainty in metrology is often linked to strict standards and regulatory requirements, which is not yet the case at the time of writing for AI models. Quantifying uncertainty in AI predictions is still an active area of research and challenges remain in standardizing this practice and fully integrating it into decision-making processes.

B. Evaluation of the Uncertainty of Measurement of an AI System by Statistical Analysis of the Results

It is proposed to apply the so-called Type A algorithm described in NF ISO 13528 which is an evaluation method commonly used in the field of metrology to:

- quantify the repeatability uncertainties of measurement results.
- quantify the reproducibility uncertainties of measurement results.
- give an estimate for an input quantity and its uncertainty from a series of n measurements.

The uncertainty of repeatability or reproducibility of the results of measurements of the same measurand x_i is determined by the calculation of the experimental standard deviation noted $s(x_i)$ is defined by $s(x_i) = \sqrt{\frac{\sum_{k=1}^{n}(x_{ik}-\overline{x_l})^2}{n-1}}$ with $x_i = \frac{\sum_{k=1}^{n} x_{ik}}{n}$ where x_{ik} correspond to the n results of measurements of the same measurand x_i.

The experimental standard deviation characterizes the dispersion of the observed values x_{ik} around their mean. The number of degrees of freedom is $v = n - 1$.

The Type A method for the estimation of a quantity X_i consists in performing a series of n measurements x_{i1}, …, x_{in} of the quantity X_i. The arithmetic mean is obtained by $x_i = \frac{\sum_{k=1}^{n} x_{ik}}{n}$. The standard uncertainty $u(x_i)$ of its estimate is defined as:

$$u(x_i) = \frac{s(x_i)}{\sqrt{n}} = \sqrt{\frac{\sum_{k=1}^{n}(x_{ik}-\overline{x_l})^2}{n(n-1)}} \tag{7}$$

The number of degrees of freedom of $u^2(x_i)$ is $v = n - 1$.

Remarks

A variance estimate_$s^2(x_i)$ performed on a set of K series of independent observations of the same quantity x_i is obtained from: $s^2(x_i) = \frac{\sum_{i=1}^{K} v_i s^2{}_i(x_i)}{\sum_{i=1}^{K} v_i}$ where $s^2{}_i(x_i)$ is the experimental variance of the ith set of ni independent repeated observations with number of degrees of freedom $v_i = n_i + 1$.

The number of degrees of freedom of $s^2(x_i)$ is $v = \sum_{i=1}^{K} v_i$; let us posit, $m = \sum_{i=1}^{K} n_i$.

The experimental variance $u^2(x_i) = \frac{s^2(x_i)}{m}$ of the arithmetic mean of m independent observations characterized by the variance estimate s^2 constructed from a data set also has v degrees of freedom.

- The number of degrees of freedom should always be given when Type A estimates of the components of the combined standard uncertainty are provided.
- For convenience, $u^2(x_i)$ and $u(x_i)$ valuated in this way are called Type A variance and Type A standard uncertainty.

In our study, we determined the standard uncertainties associated with the values obtained by two systems and for each parameter tested:

TABLE III
TYPE A STANDARD UNCERTAINTIES OF TWO AI SYSTEMS

Parameter tested	Type A Standard Uncertainties	
	System 1	System 2
Color	0,036	-
Fold	0,052	0,075
Brightness	0,045	0,085
Orientation	0,052	0,107
Cleanliness	0,062	0,083
Shading	0,077	0,078
Low resolution	0,033	0,011
Medium Resolution	0,043	0,033
Control resolution	0,036	0,069

The standard uncertainty reflects the errors/approximations produced during the algorithmic process. It is thus the quantity that could give an idea of the confidence that can be attributed to an answer provided by an AI system (did it answer with certainty? etc.).

For the type of AI system studied in this paper, it is suggested that the total uncertainty is obtained by the quadratic sum of the different contributions. Thus, for AI system 1 (SIA*1*), this could give:

$$u\,(SIA1) = \sqrt{\begin{array}{c} u^2(SIA1_{\,color}) + \, u^2\big(SIA1_{\,fold}\big) \\ + \cdots \\ + \, u^2(SIA1_{\,control\,resolution}) \end{array}}$$

$$\boldsymbol{u\,(SIA1) = 0,151}$$

The transposition of the concept of measurement uncertainty to the AI system leads us to believe that, in this given example, if a test is performed, the results provided by this AI system are true within 15.1%.

C. Discussion of the Implications of the Bias Assessment Results

The implications of the bias assessment results are important for users and developers of AI measurement systems. The results highlight the need to consider potential biases in the design and deployment of these systems. Developers must work to reduce sources of bias to ensure fair and equitable results for all users. Users should also be aware of potential sources of bias in AI measurement systems and the associated risks.

D. Recommendations for Reducing Bias

We have identified several sources of bias for cognitive document dematerialization software. Based on our evaluation results, we propose the following recommendations to reduce these biases:

- More diverse data collection: to avoid over-representation of certain categories in the training data, more diverse data collection is recommended.
- Pre-processing of data: it is recommended to pre-process the data to eliminate potential biases, such as missing or erroneous data, that could negatively affect the performance of AI measurement systems.

- Use of unbiased learning algorithms: it is recommended to use unbiased learning algorithms to ensure fair evaluation of all user groups.

VIII. Conclusion

In this paper, we have studied the feasibility of transposing to AI systems, a bias evaluation method (Z-Score) commonly used to test the performance of measurement systems in the metrology domain.

The results obtained are promising. They show a possible transposition of the Z-Score method in order to compare the performance of AI systems. In terms of form, the documentary repositories describing, for example, the organizational methods of the ILC must be adapted to the context of AI systems. To do this, a cross analysis (measurement system vs. AI system) could be carried out to identify and translate certain key concepts of metrology to the context of AI.

Moreover, two important elements must be remembered for a ILC to be truly successful in an AI context. First, it is important to ensure that the participants have the necessary skills to interpret the results obtained and to monitor the ILC process. Second, in order for the participants to get the most out of their participation, the organizer of the ILC will have to ensure that the results are made available, for example through an evaluation report.

Finally, we studied the concept of measurement uncertainty for AI systems. The results obtained could give some indication of the mastery of the algorithmic process. This reference could thus be used to quantify the gains linked to new versions of AI systems due to, for example, the improvement of machine learning algorithms.

References

[1] JCGM GUM-6:2020 Guide to the expression of uncertainty in measurement Part 6: Developing and using measurement models

[2] NF ISO 5725-2 (1994) Accuracy (trueness and precision) of measurement results and methods. Part 2 Basic method for the determination of the repeatability and reproducibility of a standard measurement method

[3] ISO/TS 21748: Guide to uncertainty estimation using interlaboratory study data

[4] Guide ISO 43 Proficiency testing of laboratories by intercomparison

[5] NF ISO 13528 December 2005, Statistical methods used in proficiency testing by interlaboratory comparisons

[6] NF EN ISO/CEI 17043 (April 2010), Conformity assessment - General requirements for proficiency testing

[7] FD ISO GUIDE 34 May 2010 General requirements for the competence of reference material producers.