Accounting AI Measures as ISO/IEC 25000 Standards Measures

Andrea Trenta [1]

[1] *UNINFO UNI TC 533 Technical Committee Artificial Intelligence, Turin, Italy*

Abstract

This paper is a part of a set of papers showing how newly defined data and software quality measures can be described in ISO 25000 format. In the first group of papers [3], [1], [28], [2], we discussed with the help of some examples, the general approach of conformance when new quality measures are defined, and in the last paper [20] how to build practical ISO/IEC 25000 compliant product quality measures for AI, starting from measures developed in several public projects. In this paper we continue to show, through some examples, that standards and research coming from the scientific community on the topic of AI measures can be easily accounted as ISO/IEC 25000 measures. Moreover, the paper can be considered for the works in AI standardization area.

Keywords

product quality, measures, accuracy, ISO, ISO/IEC 25059, ISO/IEC 5259, ISO/IEC 24029, ITU-T F.748.11, metric, AI, ML, Machine Learning, Artificial Intelligence

1. Introduction

Policy makers, industries, and academia are facing the problem of building trust in AI; in this paper we show, following the approach of a previous paper [20], how some AI measures taken from non-ISO standards and research literature, can be accounted as ISO/IEC 25000 AI product quality measures.

The items considered for AI product quality measures are recalled in the following "shopping list".

For the following, it is useful to recall definitions given in [20].

The implementation I is defined as a function of

1) I= I(method, algorithm(library, parameters), training(dataset, process))

where:

'method' is the high-level categorization [7] like decision tree, k-means clustering, neural networks,…

'algorithm' is the type of method[2] (es. ResNet for method=NN)

'library' contains the code to be invoked for evaluation (see machine learning process in [11])

'parameters' are the configuration data of the algorithm.

'training' includes dataset (ImageNet, MNIST,…) and process (initialization, retraining,…).

Then, we can define for the i-characteristic and the j-measure, the measurement

2) $M_{ij}=M_{ij}(I)$

and taking into account 1):

3) $M_{ij}=M_{ij}$ (method, algorithm(library, parameters), training(dataset, process))

Proceedings of 5th International Workshop on Experience with SQuaRE series and its Future Direction IWESQ@APSEC, December 4th, 2023, Seoul, Korea

EMAIL: andrea.trenta@dataqualitylab.it

[2] Note: for 'algorithm' it is intended the categorization of the code that perform the task, e.g. for the classification task, the 'algorithm' can be either a neural network, or a decision tree, or a support vector machine, or other.

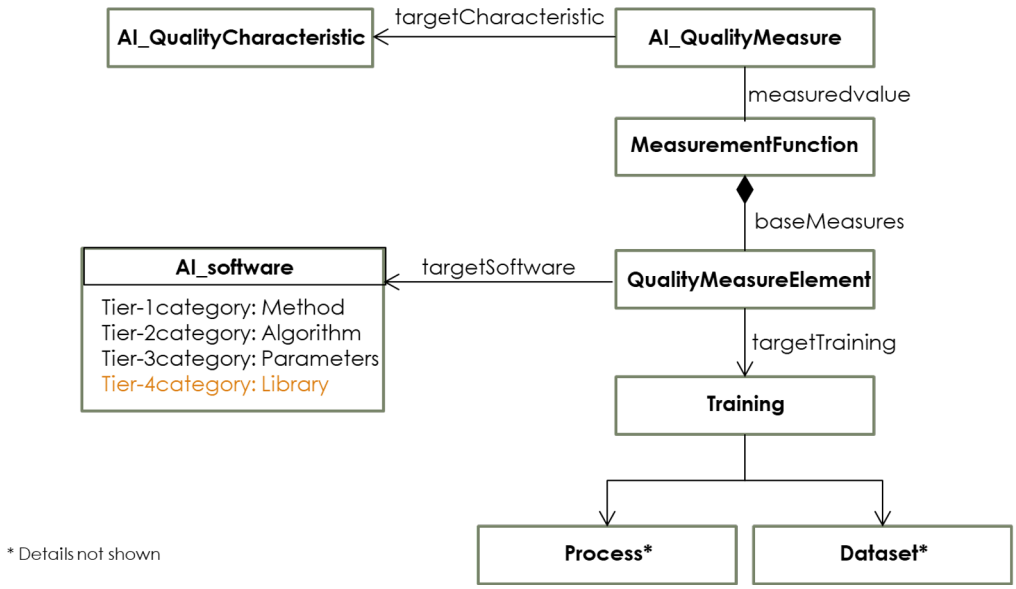

Figure 1 AI quality measure shopping list (UML-like)

With those definitions, benchmark B_{ij} is the best value M_{ij} for the time being (e.g. for a full year) for the i-characteristic and the j-measure[3] among all the K implementations of I_k

4) $B_{ij}=\max_k M_{ij}(I_k)$ k=1,..,K

Starting from those definitions, we map some existing measures to ISO 25000 measures when 1) holds. In the following, we pick those existing measures from

A. ROC curve metric [24]
B. Recommendation ITU-T F.748.11 [21],
C. Holistic Evaluation of Language Models [19]

and explain how they can be accounted as ISO/IEC 25000 measures and make some more consideration about the perspectives of the ongoing standardization work in the relevant bodies on the topic of AI product[4] evaluation and assessment.

2. AI Standardization (2023 UPDATE)

Policy makers have addressed the issue of AI trustworthiness mainly, but not only, to the international standardization body ISO/IEC SC42 and to the European standardization body CEN/CENELEC JTC21 that have in charge the drafting of technical standards in support of industry and of lawful rules. For the scope of this paper, we consider, among the others, the standards based on ISO 25000 series that define or contribute to define product quality for an AI product [8]. The assessment of product quality, possibly together with the assessment of process quality [9], will be performed in the near future on voluntary or mandatory basis, in the former case to promote trustworthiness in AI systems, in the latter case to get compliance to rules [10]. In the following, we focus on ML based AI systems [4].

The work of ISO\IEC SC42 in the last years has given birth to a set of standards on AI that are covering topics such as quality, testing, risk, management system, data, application according to the non-official scheme of figure 2.

It is to be noted that SC42 has developed and is developing extensions [5], [23] to standards of the series ISO/IEC 25000 and this appears at the moment the most mature approach to the AI product evaluation, as it relies on the core SQuaRE standards developed since 2008. Indeed, the ISO/IEC 25000 itself foresees the possibility to extend the model to specific technologies like AI, through the definition of new characteristics and new measures. This view and its reasons are also well explained in the ISO/IEC news given in https://www.iec.ch/blog/new-international-standard-ensuring-quality-ai-systems.

At the moment, the ISO 25000 extensions for AI are the technical specification for AI product

[3] in 4) the j-measure is supposed as scalar; if the j-measure is a vector or a matrix, the expression 4) should be adapted.

[4] Note: the topic of product measurement is distinct from the topic of the process measurement.

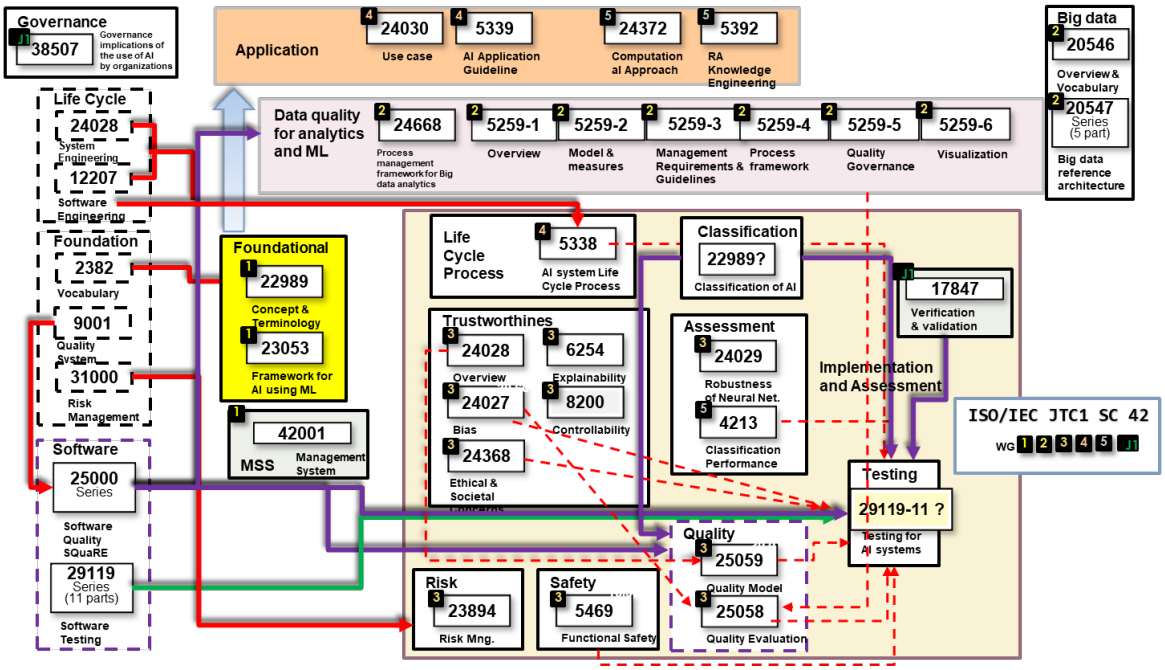

Figure 2 Non-official AI ISO standards by topics

quality evaluation [23] that is under development, and the quality model for AI already published [5], that is to be read in conjunction with [23].

The considerations of this paper are supporting the current set of ISO standards.

3. Example: AUC (Area Under Receiver Operating Curve)

A receiver operating characteristic (ROC) curve is a graphical method for displaying true positive rates and false positive rates across multiple thresholds from a binary classifier [7].

ID	Accu-ML-9-1
Name	Accuracy of Neural networks for pneumonia detection
Description	AUC Area Under ROC curve
Measurement function	X= L(I, O) L= integral of Receiver Operating Curve I_1= implementation (NOTE2) O= dataset ChestX-ray14
NOTE1 X=1 means maximum accuracy, X=0.5 means random accuracy (minimum)	
NOTE 2 I_1 = I_1 (method, algorithm (library, parameters), training (dataset, process))	
where: method = {Neural Network} algorithm$_1$ ={NSGANetV1-A3} library = {library_url} parameters$_1$ = {parameters_set} training = {ChestX-rayNIHCC, one-step training}	

Table 1 Accuracy measure – AUC in ISO 25000 format

To express performance across all thresholds, the area under the receiver operating characteristic curve (AUC) can be calculated. Higher AUCs indicate more robust performance, ranging from 0 (worst) to 1 (best). Classifiers that perform no better than chance will have an AUC of 0.5

AUC is an example of how statistical methods for assessing NN can be accounted as quality ISO 25000 measure (Table 1).

In conclusion, an AI measure like AUC well known in scientific literature and classified according to [24] into the category of statistical methods, can be represented in an ISO/IEC 25000 format[5].

4. Example: Rec. ITU-T F.748.11

The Rec. F.748.11 [21], proposes, among the others, metrics for AI applications.
The approach is to define benchmarks, as the history of processors evolution that includes both architecture, clock, energy consumption, and more parameters, has shown that it is impossible to make comparisons without a common challenging metric, like e.g. FLOP/s. At the same manner, for each triplet composed by

- Application (e.g. Image classification, Speech recognition,..),
- Dataset (e.g. Imagenet, LibriSpeech,..),
- ML model (e.g. ResNet, DeepSpeech2,..),

[5] For the scope of this paper, we don't discuss the characteristic to which the measure of table 1 is referred; as hypothesis, it could be referred to Functional correctness.

it is defined the benchmark

- Accuracy

with a specific metric for each triplet (e.g. Word Error Rate for Speech recognition implemented with DeepSpeech2 and tested against dataset LibriSpeech).
The ML model is further detailed with neuron layers, input size and source code, e.g.

- ML model detailed (e.g. ResNet_50)
- ML model source code (e.g. https://github.com/KaimingHe/deep-residual-networks)

This corresponds to the use case 1 Accuracy showed in [20] and can be represented as in Table 2 for the triplet Image Classification, Imagenet, ResNet, implemented with ResNet_50 with source code https://github.com/KaimingHe/deep-residual-networks

ID	Accu-ML-1-50
Name	Accuracy of Neural networks for prediction
Description	Prediction accuracy
Measurement function	X= L (I_{50}, O, Q) L is the Prediction accuracy (see F.748.11) I_{50} is the 50-implementation NOTE 1 O is the set of observations Q is the set of predictions
NOTE 1 I_{50} = I_{50} (method, algorithm (library, parameters), training (dataset, process)) where: method = {Neural Network} $algorithm_{50}$ ={ResNet_50} library = {https://github.com/KaimingHe/deep-residual-networks} $parameters_{50}$ = {$parameters_{50}$} training = {Imagenet, one-step training}	

Table 2 Accuracy measure – Prediction accuracy in ISO 25000 format

So, we can conclude that ITU-T F.748.11 [21] measures can be accounted as ISO/IEC 25000 conforming measures.

5. Example: LLM (Large Language Models)

In this clause we try to show how the measures performed in the research [19] can be accounted as ISO/IEC 25000 measures. To do this, we note that in [60], the following criteria are applied, that are the same criteria used for defining ISO 25000 compliant measures [20].

Firstly, [19] taxonomizes the LLM applications, as proposed in [20].

Secondly, the characteristics of the model are identified; new characteristics are introduced (calibration, toxicity) that are not present in models [5], [6] but can be handled as ISO 25000 conforming [27].

Thirdly, the measures contain the same description used in [20]:

M_{ij}=M_{ij} (method, algorithm(library, parameters), training(dataset, process))

With those considerations, it is easy to identify the full description of the measures according the ISO/IEC 25000 format.

For example, we consider the measure of detection of toxic text[6] and draw the table 3 below.

ID	Txtclass-ML-1-1
Name	Text classification by toxicity
Description	Detection of toxic text in LLM input prompts
Measurement function	X= L(I, O) L= Perspective API I_1= implementation (NOTE1) O= RealToxicityPrompts (NOTE2)
NOTE1 I_1 = I_1 (method, algorithm (library, parameters), training (dataset, process)) where: method = {Large Language Models} $algorithm_1$ ={Generative Pre-trained Transformer} library = {GPT-3 davinci v1} $parameters_1$ = {parameters_set} training = {CivilComments, one-step training}	
NOTE2 https://ai2-public-datasets.s3.amazonaws.com/realtoxicityprompts/realtoxicityprompts-data.tar.gz.	

Table 3 Toxicity - Detection of toxic text in ISO 25000 format

In conclusion, also the measures [7] for LLM presented in [19] can be mapped to ISO 25000 quality measure.

6. Measurements in Operation

As highlighted in [25], when an ML implemented with neural networks uses continuous learning, its hyperparameters are evolving, and the measurement of characteristics of the NN can be different (and assessed worse or better) from the measurement taken in the initial state. This is also the reason why the AI medical devices are deployed and sold as "frozen" [26], giving a guarantee to the user-buyer that the behaviour of the ML will be the same all the time.

Anyway, additional requirements (e.g. operational performance not worse than tested ones) and measurement can be satisfied, so

[6] In NLP applications, there is the general task of text classification, and among them there is the specific task for the machine to detect prompts with toxic text (e.g.. biased questions, hate speech,…)

[7] For the scope of this paper, we don't discuss the characteristic to which the measure of table 3 is referred; as hypothesis, it could be referred to Functional correctness.

enlarging the field of evaluation, both along the time and the post-training data and perform a further assessment of the ML in the operational mode.

7. Formal Methods

It is to be noted the awareness of the scientific community for the need of an a-priori guarantee of the robustness of NN: many papers ([12], [13], [14], [15], [16], [17], [18]) contain the word "certification" or "formal guarantees" or "verification" or "provably", as they research the proof of a target performance.

To understand how the topic is presently addressed and to complete the landscape of AI measures, we recall the approach represented by formal methods.

According to the classification given in [24], formal methods can successfully answer the question whether or not, for a given Neural Network, input and output (e.g. input image of airplane and output label "airplane"), a modification of the input leads to the same output or a different one (e.g. input image of airplane with noise, output label "helicopter"). This question can be formulated as a formal mathematical statement that is either true or false for a given neural network and image.

Based on the research that have proven that for Neural Network using the linear piece-wise activation function (ReLU), it is possible to measure robustness in terms of lower bound of minimal adversarial distortion for given input data points [13]. The results for ReLU, were successively extended to NNs with common activation functions like sigmoid [13]. The results of various research on this topic are summarized in §6.2 [25] that supports formal methods as engineering or quality evaluation of some NNs and characteristics.

Even if formal methods are at the moment considered [25] a quality approach complementary to ISO/IEC 25000, we could consider the math function that defines distortion bounds[8] as a SQuaRE measurement function and then account any formal method as an ISO/IEC 25000 measure.

[8] Formal methods are based on the theorem that, in certain conditions, there exist Upper and Lower bounds for an m-layer neural network function f with n_m neurons so that for $\forall j \in [n_m]$ and $\forall x \in (|x - x_0| \leq \varepsilon)$ holds:

$$f_j^L(x) \leq f_j(x) \leq f_j^U(x)$$

Where x is the perturbed input vector, centered in the reference data point x_0 and bounded in a sphere with ray ε.

8. Proposal

The proposal in this paper completes the proposal in [20]; there we showed how to design and document a product quality measure that includes algorithm, training dataset, library code and parameters; here we show in a sort of reverse engineering, how to account and represent measures from standard and scientific literature into the ISO/IEC 25000 format.

Finally, some investigation areas (formal methods and operational measures) and relevant considerations are presented in the perspective of an even wider application of the present and [20] paper proposals.

9. Conclusion

The role of ISO/IEC 25000 in measurement and assessing of AI product quality is widely recognized and of growing interest.

A further confirmation comes from the similarity between measurement methods developed in scientific literature and projects and the ISO 25000 conforming measurement method as shown in [20] and in this paper.

The paper analyzed this similarity and came to the conclusion that most of the measurement methods used for AI can be easily mapped into ISO 25000 format.

10. References

[1] A. Trenta, Data bias measurement: a geometrical approach through frames, Proceedings of IWESQ@APSEC 2021. URL: http://ceur-ws.org/Vol-3114/

[2] A. Trenta: ISO/IEC 25000 quality measures for A.I.: a geometrical approach, Proceedings of IWESQ@APSEC 2020. URL: http://ceur-ws.org/Vol-2800/

[3] D. Natale, A. Trenta, Examples of practical use of ISO/IEC 25000, Proceedings of IWESQ@APSEC 2019. URL: http://ceur-ws.org/Vol-2545/

[4] International Organization for Standardization, ISO/IEC 22989:2022 Information technology — Artificial

intelligence —Artificial intelligence concepts and terminology. URL: https://www.iso.org/standard/74296.html
[5] International Organization for Standardization, ISO/IEC DIS 25059 Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality Model for AI-based systems. URL: https://www.iso.org/standard/80655.html
[6] International Organization for Standardization, ISO/IEC CD 5259-2 (under development) Artificial intelligence — Data quality for analytics and ML — Part 2: Data quality measures. URL: https://www.iso.org/standard/81860.html
[7] International Organization for Standardization, ISO/IEC 23053:2022 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML). URL: https://www.iso.org/standard/74438.html
[8] D. Natale, Extensions of ISO/IEC 25000 quality models to the context of Artificial Intelligence, Proceedings of IWESQ@APSEC 2022. To appear.
[9] International Organization for Standardization, ISO/IEC 42001 (draft) Information technology — Artificial intelligence — Management system. URL: https://www.iso.org/standard/81230.html
[10] European Commission, COM/2021/206 'Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts', 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206
[11] International Organization for Standardization, ISO/IEC 23053:2022 Framework for Artificial Intelligence (AI) URL: https://www.iso.org/standard/74438.html
[12] M. Hein and M. Andriushchenko, "Formal guarantees on the robustness of a classifier against adversarial manipulation," in NIPS, 2017.
[13] Zhang H., Weng T.-W., Chen P.-Y., Hsieh C.-J., Daniel L. Efficient Neural Network Robustness Certification with General Activation Functions. Neural Information Processing Systems Conference. 2018, 31, 4944–4953
[14] A. Sinha, H. Namkoong, and J. Duchi, "Certifiable distributional robustness with principled adversarial training," ICLR, 2018
[15] A. Raghunathan, J. Steinhardt, and P. Liang, "Certified defenses against adversarial examples,"ICLR, 2018.
[16] T.-W. Weng, H. Zhang, H. Chen, Z. Song, C.-J. Hsieh, D. Boning, I. S. Dhillon, and L. Daniel, "Towards fast computation of certified robustness for relu networks," ICML, 2018.
[17] T. Gehr, M. Mirman, D. Drachsler-Cohen, P. Tsankov, S. Chaudhuri, and M. Vechev, "Ai2: Safety and robustness certification of neural networks with abstract interpretation," in IEEE Symposium on Security and Privacy (SP), vol. 00, 2018, pp. 948–963.
[18] Mirman M., Gehr T., Vechev M. Differentiable Abstract Interpretation for 1020 Provably Robust Neural Networks. Proceedings of the 35th International Conference on Machine Learning. 2018, 80, 3575–3583
[19] P.Liang, R. Bommasani, T. Lee et al., Holistic Evaluation of Language Models, Stanford Institute for Human-Centered Artificial Intelligence (HAI), Stanford University, 2022
[20] A. Trenta: ISO/IEC 25000 and AI Product Quality Measurement Perspectives Proceedings APSEC IWESQ 2022 (https://ceur-ws.org/Vol-3356/, SSN 1613-0073,https://dblp.org/db/conf/apsec/iwesq2022.html#GirirajHH22)
[21] ITU-T F.748.11 Metrics and evaluation methods for a deep neural network processor benchmark, 2020
[22] ITU-T F.748.12 Deep learning software framework evaluation methodology, 2021
[23] International Organization for Standardization, ISO/IEC DTS 25058 Software engineering — Systems and software Quality Requirements and Evaluation (SquaRE) — Guidance for quality evaluation of artificial intelligence (AI) systems
[24] International Organization for Standardization, ISO/IEC TR 24029-1:2021 Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview
[25] International Organization for Standardization, ISO/IEC 24029-2:2023 Artificial intelligence (AI) — Assessment of

the robustness of neural networks — Part 2: Methodology for the use of formal methods
[26] M. van Hartskamp, S. Consoli er al., Artificial Intelligence in Clinical Health Care Applications: Viewpoint, Interactive journal of medical research, 2019
[27] International Organization for Standardization, ISO/IEC DIS 25002 Systems and Software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Quality models overview and usage
[28] A. Simonetta, A. Trenta, M. C. Paoletti, and A. Vetrò, “Metrics for identifying bias in datasets,” SYSTEM, 2021.