2025 IMEKO TC-6 International Conference on
Metrology and Digital Transformation - **M4DConf 2025**
Benevento, Italy, September 3-5, 2025

Metrological AI Reliability Verification

Volker Zeuner, Gulian Couvreur

Federal Institute of Metrology METAS, Lindenweg 50, 3003 Bern-Wabern, Switzerland
Volker.Zeuner@metas.ch, Gulian.Couvreur@metas.ch

***Abstract* – Applications based on Artificial Intelligence (AI) are becoming useful tools in Metrology. The ability of e.g. managing big amounts of devices or analysing Big Data using AI Applications (in the sense of Software) raises speditivity and accuracy of gaining metrological results as well as facilitates metrological analyses on a new scale.**

When AI is involved in measurement, decision making, classification etc., it must be assured that the results are transparent, reproducible, comparable and traceable, corresponding to the established metrological measurement requirements.

Hence, the operation of AI Applications in Metrology should be subject to prior and periodic verification, regarding a classification of the AI types and their use scenarios (e.g. in Legal Metrology), their usage and the expected results thereof.

Keywords:
Metrology, Legal Metrology, Artificial Intelligence, reliable measurement, CRA [1], AIA [2], Cybersecurity, secure operation, conformity, trustworthiness

I. INTRODUCTION

AI Applications open a fundamentally new opportunity on how big amounts of data can be processed in a fast way in order to achieve decisions, verdicts and results regarding a given purpose – even with the possibility to retrieve continually improved ones out of a continuously learning system.

At the same time, the great computational complexity an AI System brings to play calls for a lot of necessary verification steps in order to make sure that such a system keeps its warranted properties during operation.

It is imaginable that for given use cases in metrology the operation of AI Applications looks advantageous, but at the same time lacks the reliability of approved measuring equipment. For specific use cases, the corresponding good practise requirements may be accomplished and verified easily. However, each set of such singleton use cases needs assessment regarding conformity of its elements to be able to assign AI Applications to their appropriate use case classifications.

Resulting from this request, is the subject of classifying AI Applications. The classical instruments of standardisation seem comparably slow and narrow facing a software product with a near indefinite number of possible use cases, product cycles with a fraction of the duration of standardisation phases and implementational degrees of freedom exceeding the coverage by "must, shall and should".

Furthermore, AI Applications are characterised by a certain opaqueness and, consequently, the calls for transparency, explainability or other clarifying actions to be taken in this respect exist. But without an organised approach, how can transparency be achieved?

II. APPROACH

The current state of the art

Currently, no initiatives of qualifying AI Apps for metrological systems are public.

PTB's TraCIM service provides a selection of test data sets to validate evaluation algorithms for metrological software [3]. (cf.: https://tracim.ptb.de/tracim/tenant.xhtml?
fragment=quickguide.xhtml)

The verification scheme for Metrological AI Applications for given metrological use cases might profit thereof.

Regarding conformity assessment norms, the status as of January 30, 2025, of corresponding activities of "CEN / CLC / JTC 21 - Artificial Intelligence" for the AI Conformity assessment framework was "Forecasted voting date 2026-07-29".

Under the assumption that the knowledge of a system's properties (regardless of being focused on ethical, metrological, legal or cybersecurity issues) is the paramount requisite to use or to manage it (cf. Lord Kelvin and also Mr. Peter Drucker), the development of appropriate Norms and Standards lags behind the speed of the development cycles of AI Applications as well as in regard of the presumably needed high number of use case specific documents.

At the same time, the approach to applicable specifications governing the AI issues is not "unified".

Standardisation groups and for example the authors of the EU AIA, as a specification for regulation, surely mean the same, but operate with different semantics.

Still, management guidelines do represent an approach to good practise of managing AI Applications, although it may not be stringently clear what the managed systems' properties really are and whether these stay impaired.

Appreciable examples are for instance

- NIST AI 100-1, Artificial Intelligence Risk Management Framework (AI RMF 1.0) [4]
- ISO/IEC 42001, Information technology — Artificial intelligence — Management system [5]

whereas the first document seems at least to address the shortfall of unknown system properties in its section 5.

A very interesting aspect with well-defined use cases is the usage of AI Applications for medicinal purposes. An obvious advantageous use case lies in supporting imaging technology. Unfortunately, the many degrees of freedom to select and operate AI Applications creates the need for categorisation in order to purchase or develop a suitable solution and also how to qualify the obtained results out of the operation (Jonas Richiardi, Patrick Omoumi1 et. al: "To buy or not to buy—evaluating commercial AI solutions in radiology (the ECLAIR guidelines)" in IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE [6]. This approach to a classification of procurable AI Applications bases on a lot of lessons learnt during research and operation of AI Applications. Here, it seems practical to offer a standardised pre-qualification, because this might reduce the number of "not to buy" cases and the spent effort at the same time.

Another valid solution path

"Conformance testing — an element of conformity assessment, and also known as compliance testing, or type testing — is testing or other activities that determine whether a process, product, or service complies with the requirements of a specification, technical Standard, contract, or regulation." [7]
(https://en.wikipedia.org/wiki/Conformance_testing)

Under the assumption that a conformity assessment body (CAB) is qualified to work with an appropriate specification of properties to determine that these properties are present and valid in the test object, it seems appreciable, to make the sponsor of the testing bring a requirement list instead of waiting for a Norm to become applicable.
Likewise, for different phases of an AI Application's life cycle, different specific properties may have to be verified – via adequate sets of specification.
And after all, if the relevant properties are being regarded as "assets" in the meaning of "entities that the owner of a ToE [i.e. Target of Evaluation in the CC] presumably places value upon" [8], there is obviously a way to specify sets of requirements for the properties that matter. Should there be vulnerable assets in the set, the specification of protective functionality and the verification of its correctness and effectiveness would then be part of the conformance testing, ensuring the unimpaired presence of the properties at the time of the testing.

Basic starting condition

A testing scheme for trustworthiness of AI Applications was developed covering the three following areas (respective lists not exhaustive) over all relevant life cycle phases beginning with development specifications and ending with de-commission:

1. 3 Tiers
   a. Operational environment of the AI Application
   b. AI behaviour
   c. Competence of AI User

2. 5 Assets
   a. Transparency
   b. Fairness (non-discrimination)
   c. Beneficence (non-malfeasance)
   d. Accountability
   e. Privacy

3. Assessment of asset handling comprising at least
   a. Contents and structure of output data
   b. Creation of output data
   c. Implementation aspects

On the first view, this classification seems very abstract in regard of a Metrological AI Application. In fact, the categories were defined for a broader scope of possible AI Applications. Still, the classification scheme will prove helpful to characterise AI Applications so that a more focused scope can be defined.

Being elaborate systems based on Software, all AI Applications face justified reservations about their trustworthy functioning – just like all SW.
Combined with generic SW life cycle phases, the three areas above can be projected onto:

i. design of the application and procurement of 3rd party components
ii. development, implementation and training
iii. deployment, commissioning, operation
iv. maintenance, update, upgrade
v. end-of-operation, disposal

During the life cycle, different roles appear, exercised by different legal entities: software developers, system integrators, operators end users etc.

The operational contexts of AI Applications, are characterised by many degrees of freedom. These contexts may span from DIY-developed systems for research purposes over purchased solutions operated for research or service provision over in-house metrological support SW to a marketed Metrological AI Application enhancing an until then customary measurement instrument.

Regardless of the context, all AI Applications will have been examined, whether they grant their intended "warranted properties".

The basic assumption that a purchased product maintains the warranted properties, however, is not valid for AI Applications. They are prone to change their behaviour just after commissioning and during operation.

Consequently, the equivalent of instrumental bias or drift needs to be controlled for a Metrological AI Application and every other one with legal of financial liabilities in case of malfunction. That may be covered sufficiently by in-house solutions for research purposes. For marketable AI Applications a conformity test with a quality seal (or else) may be advisable. Likewise, two different Metrological AI Applications, maybe also with different origin, for the same metrological purpose could be comparable.

III. PROOF-OF-CONCEPT SITUATION

If one would assume a Metrological AI Application similar to a measurement instrument, then each result an AI Application has contributed to should have an equivalent quality as one achieved with a classical measurement instrument alone.

Operating a metrological AI Application in the field, it seems obvious that there is a need for qualifying it via conformity testing.

Operating one in the field of Legal Metrology, it is not imaginable to run it without prior AI Reliability Verification.

Metrological AI Application Use Cases comprise:

- Measurement
- Calibration
- Analyses
- Predictive Maintenance
- Control, Alarming, Decision Making

Goals for AI Applications' Reliability in Metrology are:

- they do what they should
- they keep their (warranted) properties during operation
- they do not do what they shouldn't

Taking into regard the sections 2.44 and 2.45 of BIPM JCGM 200:2012 [9]

- The set of Metrological AI Applications needs to be classified in a way that they can be assigned to their AI class.
- The scope of the requirements for a given AI class can be matched with an AI Application under assessment.
- The fulfilment of the requirements can be verified by applying a defined verification scheme.

Thus, each AI Application can be subject to Conformity Assessment, which should be congruent with the "verification process" according to section 2.44.

A Reliable Metrological AI Application, hence, will have to undergo assessment focusing on:

- measurement requirements,
- safe and secure operation,
- results that are transparent, reproducible, comparable and traceable.

Here, the generic approach as described in section II. needs to be refined.

Purpose, expected benefit and operational environment must be known before the implementation starts.

Depending of these aspects, some "assets" (cf. II.2.) may be excluded, whilst other ones must be added.

After all, if there is a comprehensive list of assets characterising a given AI Application, appropriate measures can be taken to conserve these properties during the AI Application's life cycle.

Conformity testing against such a set of properties, derived from the identified assets seems an appropriate tool to assess, whether the assets are covered and maintained in the phases of design, specification, implementation, training, commissioning, operation and maybe even at the end of operations (cf. previous section).

A. SELECTION OF SPECIFIC REQUIREMENTS

The AI Application should be analysed in regard of the intended warranted properties, for example:

- Measurement: The AI Application shows reproducible output including reliably low uncertainty.
- Data Analytics: The AI Application generates reproducible results.
- Machine Learning; The AI Application's output keeps continuity over time or improves whilst adding new data.

When the specific properties are defined, functional or architectural objectives need to be formulated. Their achievement can be seen as specific assets to be protected and can be added to the list in II.2.

From a metrological point of view, the traceability of the results of an AI Application seems worth a deeper consideration.
The common interpretation of traceability in an AI Application focuses rather on the safeguarding of the prevention of bias or drift in the output results by being able to trace causal interrelations in the AI Applications' functionality.
If the properties of a measurement instrument are well known, it might be possible to model a possible drift of the measurement results, thus being able to operate under known conditions – and even with changing measurement results.
Consequently, given a sufficient transparency regarding an AI Application's functionalities, a traceability of biased or drifting results might also become possible, because the applied assessment scheme would make it explainable.

B. DELINEATION OF AI APPLICATION TYPES

Likewise, the operational specifics must be analysed and specified, for example in regard of:

- generative vs. discriminative results
- predictive capabilities
- decision supporting vs. stand-alone decision-making
- potential liabilities for manufacturers or operators
- in-house vs. public operation

If objectives or assets can be formulated, their achievement can be seen as specific assets to be covered and can also be added to the list in II.2.

C. ASSESSMENT SCHEME PHASES

During the life cycle phases the list of assets will not change. The measures taken surely will differ, but always shall follow the idea to keep all assets unimpaired.
Subject to the conformity testing scheme will always be the list of assets, assumed sources for impairment thereof, processes and functions implemented to protect all given assets against the feared impairment and specific steps taken to ensure the AI Application's producing the desired answers being stimulated by given sets of data.

Non-exhaustive lists of aspects to be covered during different phases:

i. design of the application and procurement of 3rd party components:

- purpose of the AI Application & AI type
- input data in training and reality
- output
- 3rd party components
- models and algorithms
- training data qualification

ii. development, implementation and training:

- training data
- output
- warranted characteristics
- verification processes
- acceptance process

iii. deployment, commissioning, operation

- output vs. training data
- output vs. reality data

iv. maintenance, update, upgrade
same as iii.

v. end-of-operation, disposal

- removal of non-disclosable modules
- re-use of pooled data

D. ASSESSMENT STEPS

In order to assess the correctness and effectiveness of the implementation as well as the rendering of the warranted characteristics by an AI Application, for each assessment step a qualified and verifiable description of how the (relevant) assets are protected must be documented.
There are three tiers:

- ease-of-use (i.e. safe and secure operational environment)
- regular functioning (the AI Application only does, what it is supposed to do)
- competent user (e.g. like a driver's license)

and five life cycle phases, which defines 15 testing steps.
In the testing steps, the defined set of assets, their freedom from impairment and the taken measures will be analysed. The analysis should at least contain results in regard of how the functionality regarding

- Contents and structure of output data
- Creation of output data
- Implementation aspects

maintains the assets impairment-free.

IV. RESUME AND RATIONALE

This conformity testing scheme for AI Applications is under development. It is characterised by treating the desired properties as assets and testing their presence in the AI Application under assessment and the AI Application's capability to uphold these properties over time in an acceptable way.

The properties are partly non-technical. The manufacturers are free to specify their interpretation of the adequate properties and their protective measures taken.

The testing follows a standardised but use case specific scheme. The result is certifiable by an accredited CAB.

An AI Application as a measurement instrument in the field of metrology would represent a comparably small test object with relatively focused properties and, hence, would represent an ideal proof-of-concept use case for the Metrological AI Reliability Verification. At the same time, a method to qualify AI Applications for use in the field of Legal Metrology will provide room for digital transformation.

Ongoing research regarding this subject will be concluded. The objective is to gain a conceptual approach to be published.

In the meantime, co-operative contributions or partnerships are welcome.

REFERENCES

[1] Cyber Resilience Act, REGULATION (EU) 2024/2847 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

[2] AI Act, REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

[3] PTB's TraCIM service; https://tracim.ptb.de/tracim/tenant.xhtml?fragment=quickguide.xhtml

[4] NIST AI 100-1, Artificial Intelligence Risk Management Framework (AI RMF 1.0) https://doi.org/10.6028/NIST.AI.100-1; January 2023

[5] ISO/IEC 42001, Information technology — Artificial intelligence — Management system First edition 2023-12

[6] Patrick Omoumi, Alexis Ducarouge, Antoine Tournier, Hugh Harvey, Charles E. Kahn Jr, Fanny Louvet-de Verchère, Daniel Pinto Dos Santos, Tobias Kober, Jonas Richiardi; To buy or not to buy—evaluating commercial AI solutions in radiology (the ECLAIR guidelines); European Radiology https://doi.org/10.1007/s00330-020-07684-x

[7] Conformance testing; https://en.wikipedia.org/wiki/Conformance_testing

[8] Common Criteria for Information Technology Security Evaluation; CCMB-2017-04-001 Part 1: Introduction and general model April 2017, Version 3.1, Revision 5

[9] JCGM 200:2012, International vocabulary of metrology – Basic and general concepts and associated terms (VIM) 3rd edition, 2008 version