TÜV AUSTRIA

REQUIREMENTS AND PROPOSED SOLUTIONS FOR THE ASSESSMENT OF AI SYSTEMS WITH REGARD TO THE EU AI ACT

A TÜV AUSTRIA position paper

REQUIREMENTS AND PROPOSED SOLUTIONS FOR THE ASSESSMENT OF AI SYSTEMS WITH REGARD TO THE EU AI ACT.

A TÜV AUSTRIA position paper

AI is one of the fastest growing technologies and will have a lasting impact on our society and working world. Often, the entry of new breakthrough technologies is also associated with the occurrence of new challenges and risks. **The EU AI Act, which was recently passed by the European Parliament, attempts to regulate the risks of AI applications.**

However, from the point of view of a testing organization that launched the world's first test catalog for AI applications back in 2020 and has accumulated considerable experience in auditing machine learning models, the **known risks that can arise from the use of AI are not adequately addressed by the current regulatory trials:** for example, the foundations for future conformity assessments with regard to the functional requirements for these systems, such as robustness, transparency and reliability, are not sufficiently taken into account in the standardisation committees and the EU AI Act. There is also a lack of concrete specifications for uniform risk assessment procedures and possible opt-out procedures for stand-alone systems harbour the danger of unequal treatment of similar products. **In its current form, the EU AI Act thus fails to achieve the goal of ensuring quality through functional trustworthiness and correct allocation of responsibilities.**

It also seems alarming to us that not only the current draft of the EU AI Act, but also the accompanying standardization efforts in CEN/CENELEC have retreated to the position that real functional guarantees of AI systems would be supposedly unrealistic and too complex. **The adoption of a conformity assessment procedure, especially for safety-critical applications, which simulates trust in inadequately evaluated AI systems, must be firmly rejected.**

Thus, the discussions in standardization circles to define what should constitute a representative data set for the training and testing of applications seem nonsensical to us. The current definition means that, for example, there could be different conformity assessments by different organisations entrusted with testing such systems.

Current scientific papers on the fundamentals of AI certification, such as those by Hochreiter, Nessler, Doms [https://arxiv.org/abs/2310.02727], on the one hand share these concerns about an emerging misalignment for the regulation of AI, but also show how an appropriate verification of AI systems, especially for safety-critical applications, can be carried out in the future.

We demand that so-called low-profile certifications or purely paper- and document-based reviews of risky AI applications must be avoided. Our experience from the certification projects already carried out also shows that the frequently expressed fear of some stakeholders that test procedures with sufficient depth of testing would lead to over-regulation is a bogus argument, because even audits for complex models and applications, taking into account functional safety requirements, have not exceeded an economically justifiable scope.

Due to our extensive knowledge of possible problems in the auditing of AI systems, we know how to effectively address them in the context of AI audits. This applies, for example, to the input data spaces in which the state of an AI system can range for common tasks performed by these systems, which often gives the impression that exhaustive tests are not feasible. In addition, the behavior of these systems also depends heavily on the data they have been trained with. And ensuring that this data is independent or tamper-free is sometimes difficult. The complexity, as we know it from deep neural networks, for example, also makes it more difficult to detect malfunctions or manipulations from the outside and to remedy them.

In order to ensure a safe, robust and transparent use of AI, especially in safety-critical applications, functional requirements as well as concepts, methods and tools for appropriate testing are required. Since much of the rapid technological change in AI is still in the area of basic research, the experts in the standardisation committees and legislation often lack reliable empirical values on which to base the definition of a regulatory framework.

This lack of foundations was also the reason for the collaboration between Johannes Kepler University (JKU), Software Competence Center Hagenberg (SCCH) and TÜV AUSTRIA, which began in 2020, to develop a TRUSTED AI test procedure based on scientific principles for Machine learning applications.

This TRUSTED AI method has already been successfully verified in the field of industrial applications, automotive assistance systems and medical diagnostic systems. These projects have shown that it is particularly important to determine the reliability of AI systems by means of appropriate quantitative test criteria, i.e. to be able to certify functional trustworthiness.

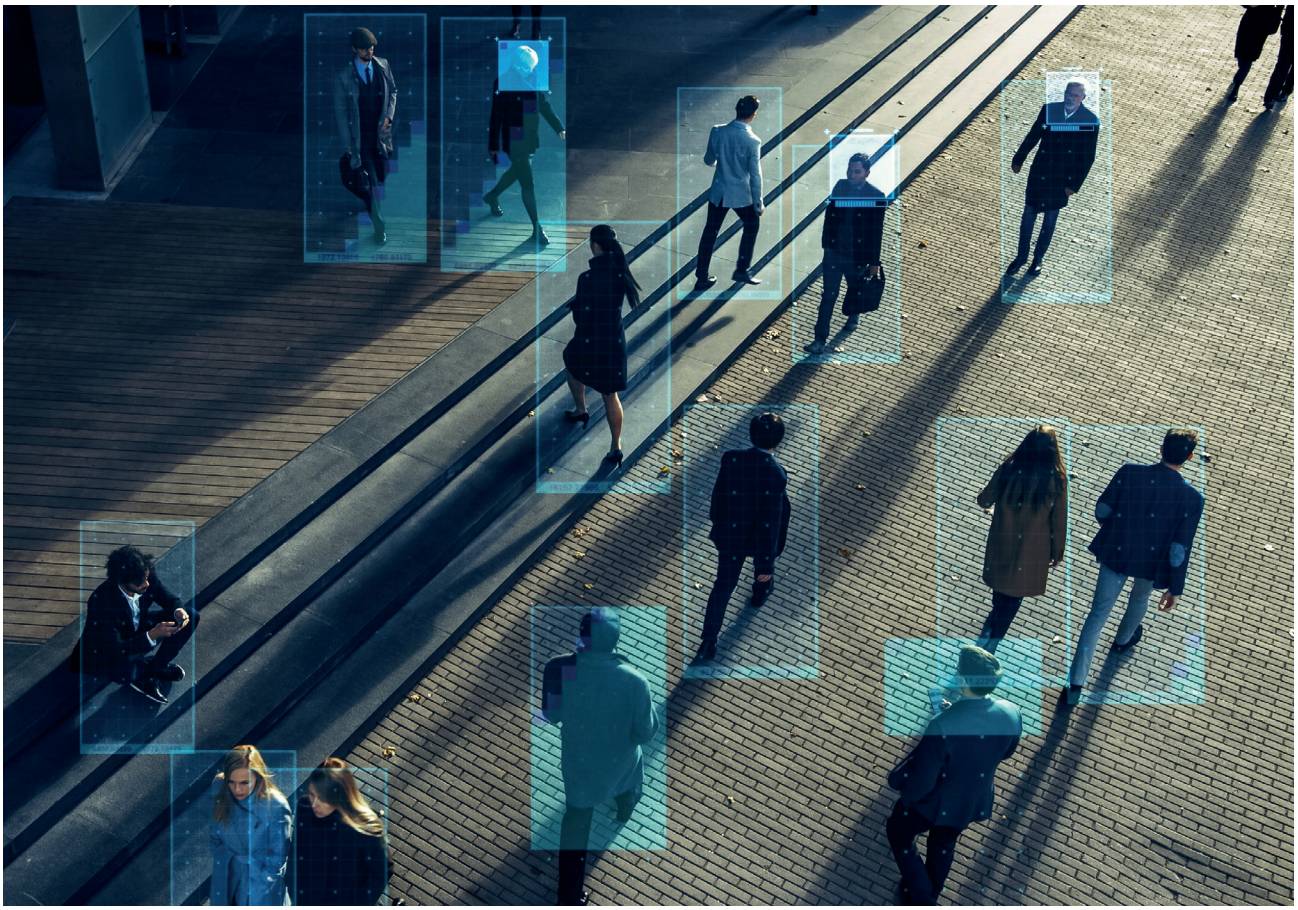
The quality of a trustworthy AI decision-making system can be determined primarily by the correct statistical testing on randomly selected samples and in the precision of the definition of the field of application, which makes the drawing of representative samples possible in the first place. **Therefore, we demand that a reliable evaluation of the statistical functional properties of an AI system according to the known scientific state of the art must form the indispensable and binding core of a conformity assessment.**

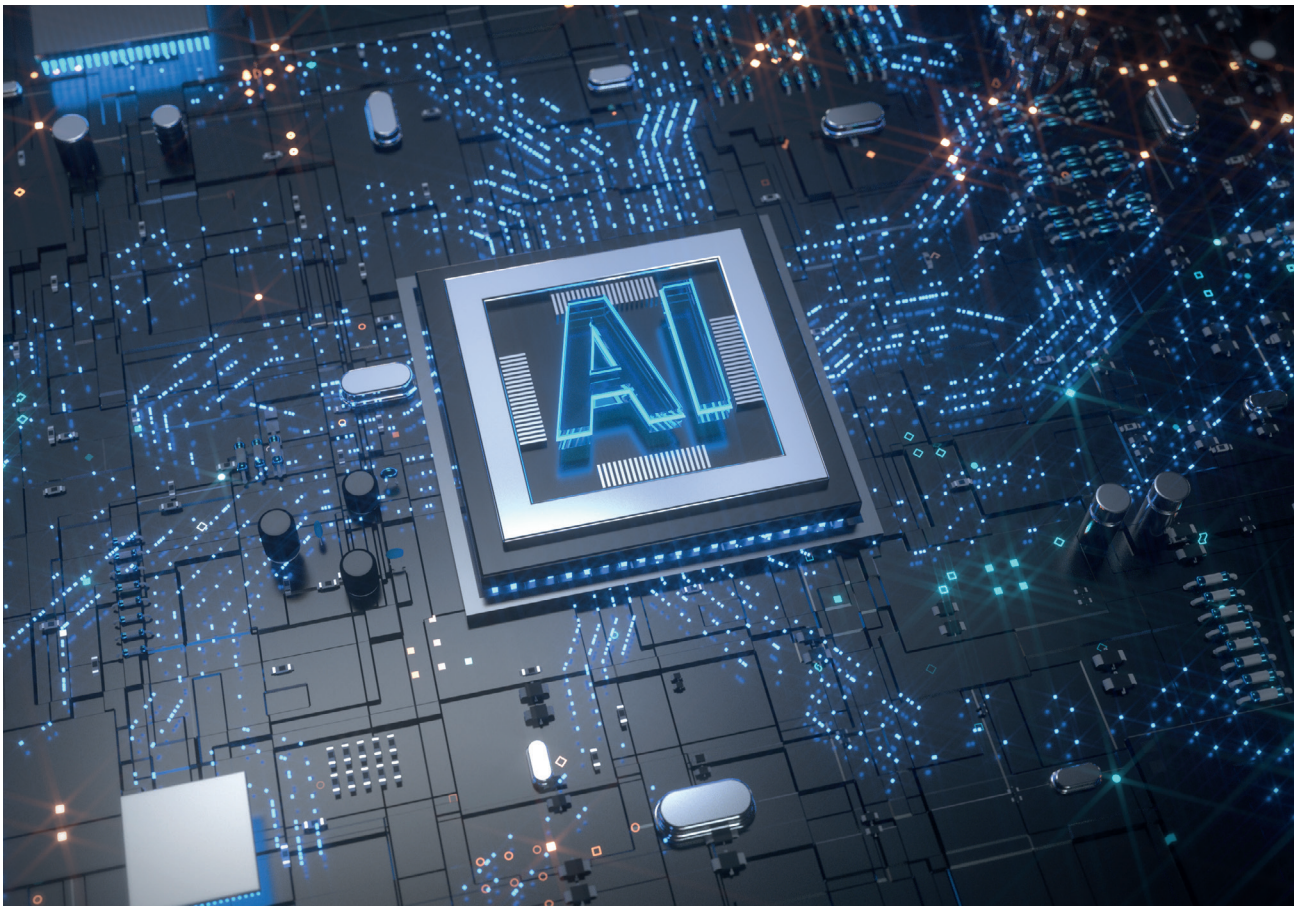
The **3 necessary elements to establish a reliable, functional trustworthiness** are therefore
- a precise definition of the intended use, formalized as a technical distribution that allows for random selection,
- the collection of risk-based minimum performance requirements and
- statistically valid testing on the basis of independent random samples.

The TRUSTED AI testing procedure [https://www.tuv.at/trusted-ai-by-tuev-austria/] developed by TÜV AUSTRIA, JKU and SCCH and verified in practice can already fully cover these elements and guarantee their fulfillment.

In our opinion, the handling of the use of artificial data and simulations for training and testing purposes also needs further investigation before they should be used safely in the development of safety-critical applications. In addition, the handling of domain gaps, also known as distribution shifts, which arise when the data distribution for learning and application do not match, which can occur in real applications, must be adequately covered by a test catalog. A typical scenario for this is, for example, learning and training the system using artificial, simulated data and then applying the model in the real world with at least slightly different data distributions.

Since, compared to conventional software, AI applications are usually self-learning systems that are less static and predictable in their behavior, the question also arises which additional requirements must be provided, for example, for the detection of distribution errors.





There is also the question of how the threshold values for uncertainty in particular should be defined and ultimately determined for safety-critical applications. Most machine learning models currently do not provide a well-founded confidence measure for the uncertainties of the model.

However, the majority of these addressed problems can already be adequately addressed today - also contrary to other expert opinions - with the TRUSTED AI procedure audit procedure and functionally validated within the scope of model tests.

For this reason, the TÜV AUSTRIA Group, in cooperation with its scientific partners from the Institute for Machine Learning at the Johannes Kepler University Linz and the Software Competence Center Hagenberg, recommends the use of a certification procedure and criteria catalogue according to the latest scientific findings, as defined in the TRUSTED AI testing catalogue, for appropriate regulation.

We take a **holistic approach** to analysing machine learning applications from multiple perspectives **to assess and verify aspects of functional requirements, data quality, secure software development, risk handling and ethics.**

With regard to ethical aspects, it is of utmost importance to make a clear distinction between predictive models that aim to reflect reality in an unbiased way and decision-making models that aim to make an appropriate choice with regard to a specific definition of the goal, i.e. to optimize the decision-making process with regard to the goal. Ethical considerations must be explicitly taken into account when designing the goal definitions and should never be implemented through artificial bias of the input data or artificial inductive bias of the ML.

The TRUSTED AI test method can be applied to both supervised and unsupervised learning applications. **We are happy to share our criteria, experiences and evaluation results with the EU Commission and its standardization bodies.**

November 2023

TÜV AUSTRIA HOLDING AG
Deutschstraße 10, 1230 Vienna

Bilder: Shutterstock (Ryzhi | Advance Designer | Gorodenkoff | HelloRF Zcool | your)