Technical Assessment of High-Risk AI Systems: State of Play and Challenges

Introduction

The European AI Regulation, also known as the EU AI Act (AIA), was established to provide a harmonised and horizontal legal framework for the use of AI models and systems in the European Union. Its primary goal is to maximise the benefits of AI technologies while minimising the risks associated with their use. The AIA classifies AI systems into various risk categories: prohibited AI, high-risk AI, low-risk AI, and minimal-risk AI. High-risk AI systems are subject to stringent technical and regulatory requirements.

In addition to documentation obligations, transparency rules, data requirements, quality and risk management protocols, technical criteria, such as accuracy, robustness, cybersecurity or bias (in the AI Act covered with regard to data governance), are critical in the evaluation of AI systems. This whitepaper delves into these criteria, outlining their significance and challenges in their technical evaluation.

In this paper, we explain the technical testing criteria accuracy, robustness and bias, and broadly discuss different methods to test AI systems w.r.t. these criteria.[1] Furthermore, we outline challenges that arise when testing AI systems, arguing that, while testing these criteria is technically feasible, it crucially depends on the knowledge about the use case and domain in which the system is deployed. To this end, the TÜV companies can leverage their sector-specific expertise to successfully enable the adaption of AI technology in high-risk applications.

What is Accuracy, Robustness and Bias?

Accuracy in the context of the AIA refers to how well an AI system can deliver correct predictions (Art. 15(1), EU AI Act). A system with high accuracy minimises errors or incorrect predictions, which is particularly crucial in safety-critical applications such as medical diagnostics or machinery. The AIA requires that high-risk AI systems are regularly tested and monitored to ensure that their accuracy

[1] Technical criteria related to cybersecurity are out of scope of this document.

remains at a high level. This includes the use of valid testing procedures and metrics to ensure that the models integrated into such systems perform reliably.

Next, robustness refers to how resilient an AI system is to disturbances or changes in input data (Art. 15(1), EU AI Act). A robust system must remain stable and reliable even under changing conditions or deviations in the data. This is crucial to prevent AI systems from being compromised by small errors or manipulations. The AIA therefore requires that AI systems are regularly tested for their robustness, particularly regarding their ability to function reliably even under adverse conditions.

Finally, bias in the context of this paper refers to a consistent deviation in the model's prediction, which can lead to systematic errors favouring certain outcomes over others (including, but not limited to, outcomes produced for certain user groups). This can occur due to statistical biases where estimators do not accurately reflect the true parameters, or model biases stemming from assumptions that oversimplify the underlying relationships in the data. Algorithmic biases can also skew results if the algorithm used to train or construct a model has inherent limitations that favour certain types of solutions. Generally, bias can lead to predictions that do not align with desired outcomes, making the AI system inaccurate and, thus, unreliable. While bias is mentioned in the AIA specifically in the context of data governance (Art. 10, EU AI Act), testing for bias in the sense of systematic model errors represents a criterion that not only requires inspection of the data at hand but also the model behaviour. These tests may, in turn, inform the data collection or preparation process. Notably, the AI Act does not specify, which measures should be used to detect, prevent and mitigate possible biases (Art. 10(2g), EU AI Act).

Measuring Accuracy, Robustness and Bias

Technical assessment of an AI system requires the computation of adequate performance metrics. The following section will present several metrics that are most commonly used to measure the performance of an AI system with respect to the criteria mentioned above. Notice that there exists a plethora of metrics, which will not be fully covered in this paper.

First, metrics for accuracy are considered. One commonly used metric to assess the accuracy of an AI system is its *precision*, which measures the proportion of true positive predictions among all positive predictions. On the other hand, *recall* (or *sensitivity*) measures the proportion of true positive predictions of all actual positive cases. There also exist aggregated metrics such as the *F1 score*, which presents the harmonic mean of precision and recall, thereby balancing the two metrics. These metrics can be used to set requirements and/or reason about the performance of a given system. For example, an AI system diagnosing breast cancer must ensure a high F1 score to minimise both false negatives and false positives. In the case of a spam filter, high precision ensures that legitimate emails are not wrongly marked as spam.

When quantifying the robustness of an AI system, one can compute its *adversarial robustness*, *out-of-distribution robustness*, *robustness against noise* or its *robustness against input variations*. Adversarial robustness measures the system's resilience against deliberate input manipulations bounded by a distance metric capturing the difference between the original input and the adversarial example. Out-of-distribution (OOD) robustness evaluates the system's performance on real data points that lie outside the training distribution. Robustness against noise assesses performance stability when (domain-specific) noise is added to the input data. Lastly, robustness against input variations evaluates how well the system can cope with systematic variations of the input, such as rotation or changing brightness in the case of image data. These metrics are relevant in use cases ranging from autonomous driving, where systems must perform reliably in adverse weather conditions, to speech recognition, where robustness ensures accurate results e.g. in the presence of accents or background noise.

Lastly, the presence of bias in an AI system can be detected using metrics that are similar to those used for measuring accuracy but, importantly, applied in a different context. The key difference to accuracy testing lies in the selection of the data instances or scenarios, which should represent the test cases for which we expect equal performance levels. Especially in applications that serve as tools for access control to social or economic benefits (e.g. recruitment systems), specialised metrics such as *demographic parity*, *equal opportunity*, and *predictive equality* can be used to assess whether different groups receive similar predictions (or benefits) from the model.

Testing in Practice: Technical Challenges

Testing an AI systems w.r.t. the criteria mentioned above raises several technical challenges. These challenges are amplified by further, more general issues, such as the lack of concrete guidelines and standards for AI testing or the ambiguity in legal terminology, which are not addressed in this paper.

The technical challenges have implications for the complexity of the method used for computing a specific metric w.r.t. a given criterion. In this work, we heuristically characterise the complexity of a method by the degree of how well it can be automated and the degree of implementation efforts demanded by the method. An overview of this categorisation is presented in Figure 1. Note that in this context, automatability refers to the extent to which a testing process can be executed without human

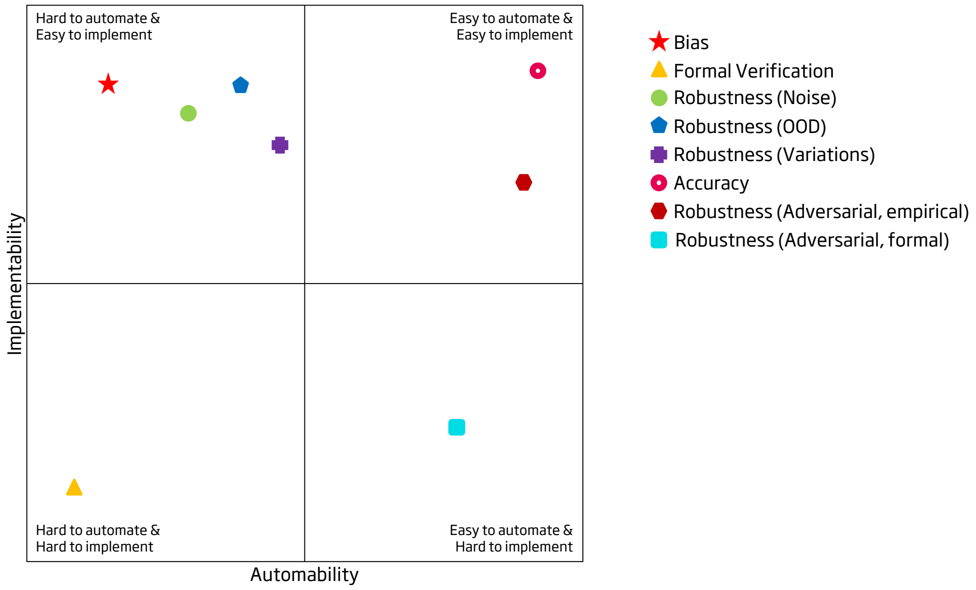

Figure 1: Overview of testing methods and categorisation based on their complexity. In this context, complexity is characterised by means of the degree of automatability (x-axis) and degree of implementability (y-axis).

intervention, relying instead on computational tools, predefined algorithms, or workflows. In contrast, processes that require significant human intervention and expertise, such as tailoring tests to specific domains, interpreting ambiguous results or setting custom parameters, are less automatable. In the following, we will further specify some of the challenges that arise from computing metrics for each testing criterion and the implications of these challenges for the complexity of the methods used.

In essence, the task of testing for accuracy involves straightforward comparisons between predicted outputs and actual labels, which is computationally inexpensive and conceptually simple. Accuracy metrics are well-defined and standardised, and they typically only require labelled testing data and access to the system outputs. In addition, they have clear mathematical formulations; hence, no specialised methods or algorithms are needed. This makes the accuracy testing methods relatively easy to implement, as a simple routine or function can be used to compute relevant metrics. Furthermore, the process of testing can easily be automated, as it does not require manual intervention once the datasets are provided. The key challenge when testing for accuracy lies in obtaining high-quality testing data, which is representative of the task at hand, and in choosing the appropriate metrics. This is by no means a trivial task; however, we regard the provision of appropriate datasets and accuracy metrics as given for this analysis.

While accuracy testing does not require simulating complex scenarios, this becomes imperative when testing for robustness of an AI system. However, the complexity of these scenarios differs between different notions of robustness (e.g. adversarial robustness *vs.* robustness against environmental changes). Furthermore, it should be noted that different approaches exist to test for robustness, that is, one can use empirical or formal methods, where the former involves generating counterexamples and the latter seeks to obtain formal robustness guarantees using advanced reasoning or optimisation techniques. Testing for adversarial robustness in an empirical fashion is easy to automate due to the availability of well-defined metrics such as adversarial accuracy. Moreover, it involves generating adversarial examples and evaluating model performance, which can be fully scripted. However, it is slightly harder to implement than accuracy testing due to the need for designing and configuring adversarial attack methods, selecting appropriate parameters (e.g., perturbation magnitude, iterations), and ensuring coverage of diverse attack scenarios, all of which require a deeper understanding of adversarial techniques and computational trade-offs.

On the other hand, testing for out-of-distribution (OOD) robustness and robustness against variations is easy to implement because both rely on straightforward processes: evaluating system performance on predefined OOD datasets or datasets with systematic variations (e.g., noise, transformations, or corruptions). Given such datasets, standard accuracy metrics can be used to measure robustness. However, automating these tests is conceptually challenging as it requires generating a wide range of OOD or perturbed datasets, which often involves domain-specific transformations or synthetic data augmentation.

Unlike empirical methods, formal verification of robustness is both hard to implement and automate because it requires highly complex algorithms and tools to handle the non-linearity and high dimensionality of AI models. Additionally, robustness properties must be carefully defined based on the specific safety specifications of the system, which varies significantly across applications. In contrast, when testing for adversarial robustness, the property is more generic as it focuses on model behaviour within a defined perturbation range and does not require domain-specific adaptation. Once implemented, the verification method can be used across domains, making it easier to automate.

Lastly, we discuss the complexity of testing potential bias in AI systems. Testing for bias is easy to implement because it relies on well-defined metrics such as common accuracy metrics or specialised metrics such as disparate impact, demographic parity, or equalised odds, which can be computed directly using labelled datasets. Bias metrics involve straightforward statistical comparisons between groups, making manual implementation relatively easy with adequate data at hand. However, bias testing is hard to automate because it often requires identifying and labelling the attributes used for defining relevant groups of data instances, which may not always be explicitly available or straightforward to define. Additionally, bias is extremely context-dependent, especially when human individuals are affected by the outcomes, requiring human judgment to determine the appropriate metrics and thresholds for specific applications, making fully automated pipelines challenging to design and deploy.

Conclusion: Technical AI Assessment needs Domain Experts

The technical evaluation of AI systems is a cornerstone of compliance with the AI Act. Accuracy, robustness and bias are essential criteria that determine the safety, reliability, and trustworthiness of these systems. At the same time, several challenges exist with respect to the technical assessment of AI systems, especially when it comes to automation. These challenges make it difficult to design off-the-shelf testing frameworks or software to perform automated tests of AI systems.

This applies to the general inspection of the testing data or choice of appropriate performance metrics but is further amplified for criteria such as robustness or bias, where the former requires capturing relevant changes in the system environment and the latter requires the definition of what is considered an undesirable bias before testing can be performed. Thus, testing these criteria demands the involvement of domain experts from the respective area of use in the testing procedure.

Considering this, the availability of knowledge about the specific use case and domain in which the AI system is deployed becomes a major bottleneck for those testing AI systems. With longstanding expertise in testing and certification of products in several high-risk domains, such as medical devices or machinery, the TÜV companies are well equipped to provide comprehensive testing schemes for AI-based products used in high-risk applications. The central task is to integrate AI-specific testing methods, such as those outlined in this paper, into existing certification processes, and to adapt them to the use case at hand.