

Short Note

Not peer-reviewed version

Requirements on Interpretation Tools for AI Systems

[Stefan Haufe](#) *

Posted Date: 10 March 2025

doi: 10.20944/preprints202503.0383.v3

Keywords: interpretation tools; explainable AI; uncertainty quantification; European AI Act



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Short Note

Requirements on Interpretation Tools for AI Systems

Stefan Haufe ^{1,2,3}

¹ Technische Universität Berlin; haufe@tu-berlin.de

² Physikalisch-Technische Bundesanstalt, Berlin

³ Charité - Universitätsmedizin Berlin

Abstract: This work formulates generic requirements on AI interpretation tools as alluded to in the European AI Act. Methods to quantify predictive uncertainties and methods referred to as so-called explainable AI may be considered examples of interpretation tools. The requirements aim to ensure that the deployer of an AI system can use such tools to assess and assure the quality and proper functioning of the system. I argue that the concrete purpose of an interpretation tool needs to be specified, the information provided through its output needs to be unambiguously defined, the utility of the output for serving the specified purpose needs to be substantiated, and evidence needs to be provided that the output is accurate and precise and, hence, that the tool's intended purpose can be fulfilled.

Keywords: interpretation tools; explainable AI; uncertainty quantification; European AI Act

1. Introduction

In the following, the terms “high-risk AI system”, “provider” (of the AI system), and “deployer” (of the AI system) are used as defined in the European AI Act [1].

The AI Act regulates the use of products comprising artificial intelligence (AI) components in the European Union's market. Article 13 demands that “*High-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately*”. Whereas Article 14 demands that “*High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons*”, which entails that “*natural persons to whom human oversight is assigned are enabled [...] to correctly interpret the high-risk AI system's output, taking into account, for example, the interpretation tools and methods available*.”

Such rules aim to ensure that deployers of high-risk AI systems can assess and assure their quality. However, it is currently unclear *how* providers can implement such rules as the harmonized European standards addressing the AI Act are just currently being drafted [2]. While the use of interpretation tools is not strictly mandated by the AI Act, there is still a need to define the scope of such tools, and to formulate specifications on the design and use of such tools and the reporting on such tools.

Importantly, the mere use of an interpretation tool is insufficient to enable quality assurance for any AI system (high-risk or not). Interpretation tools are often statistical estimators providing uncertain and sometimes not even well-defined output quantities. As such, these tools require quality assurance themselves rather than being able to serve quality control purposes. This document collects high-level requirements that aim to ensure that interpretation tools can indeed provide quality control for AI systems.

The document partially builds on prior work [3].

2. Requirement Specification

2.1. Interpretation Tools

Interpretation tools are tools that provide additional information about an AI system beyond its output and technical documentation. The provided information can, for example, characterize the AI system in general or its components including the AI model and its parameters, the training data, a given test input or group of inputs, the model's behavior in general, its behavior on data in or outside the training data distribution, and its behavior on a given test input or group of inputs.

Examples for interpretation tools are tools quantifying or estimating uncertainty in the system's outputs or model parameters [4], and tools designed to provide so-called explanations for the system, its output, or any of its components [5].

2.2. Purposes of Interpretation Tools

Interpretation tools can address purposes including but not limited to the following:

- i. Enabling the deployer to reject or correct training or test data on the basis of insufficient data quality. Data quality issues can include: unwanted confounding, data imbalance, bias, presence of noise, artifacts, and outliers.
- ii. Enabling the deployer to reject or correct an AI model on the basis of insufficient training data quality or inappropriate model behavior. Inappropriate model behavior can include: unwanted reliance on confounding information in data, unacceptable levels of uncertainty, bias, unfair decision making.
- iii. Enabling the deployer to reject, scrutinize (e.g. by cross-checking with the output of a second model or the opinion of a human expert), or correct outputs of a model on a given input or group of inputs on the basis of insufficient quality of test inputs, test inputs being outside the training distribution, unacceptable levels of uncertainty, or other reasons.
- iv. Selecting certain training or test data inputs, or input dimensions, for further inspection, e.g., to confirm the presence of noise or artifacts in inputs, or to assess the predictive value of individual input dimensions.
- v. Recommending certain dimensions of a test input for external intervention, for example with the goal of simulating a model's output (e.g., a credit risk score) or predicting a real-world quantity approximated by the model (e.g., a health outcome) in a counterfactual setting.

3.3. Requirements on Interpretation Tools

Interpretation tools are often statistical estimators and exhibit bias and variance. The correct interpretation of their outputs moreover typically rests on assumptions on the AI model, its training data, and test data [5]. Consequently, these methods are vulnerable to violations of their assumptions such as model misspecification, noise or artifacts in the data, and adversarial attacks. So-called explainable AI tools often in- or explicitly rely on simplistic assumptions on the causal dependency structure of the variables of the AI system such as mutual independence of the input dimensions or a strictly causal dependency of the predicted variables on the model's inputs [7]. Often, outputs of explainable AI tools can not be interpreted as estimates of well-defined properties of an AI system or its components [6,7]. These outputs can easily be misinterpreted [8,9], limiting their value for AI quality assurance [7]. As such, interpretation tools themselves need quality assurance, and their capability to systematically provide the same for an AI system needs to be enforced through appropriate requirements collected in the following.

If a provider makes an interpretation tool available and recommends its use for assessing or assuring the quality of an AI system, the following information *shall* be provided to the deployer:

- i. The intended purpose of the tool.

- ii. The information provided by the tool, defined as the concrete interpretation of the tool's output. The output shall correspond to well-defined unambiguous properties of the AI system in general or its components.
Example: A tool may provide 95% confidence intervals for outputs of an AI system.
- iii. A logically sound line of argument stating how the provided information enables the tool to fulfill its intended purpose when used by the deployer according to provided instructions.
- iv. The technical constraints including assumptions on the components of the AI system (e.g., model class, training data, test input) affecting the accuracy and precision of the tool with respect to providing correct information about the AI system or its components and with respect to serving its intended purpose.
- v. The expected accuracy and precision of the tool with respect to providing correct information, and the expected accuracy and precision of the tool with respect to serving its intended purpose.
Reported accuracies and precisions shall be based on either of the following, or both:
 - Theoretical guarantees taking into account the technical constraints and assumptions of the tool and the properties of the AI system and its components.
 - Empirical results obtained using large enough and sufficiently representative sets of test inputs.
Example: Uncertainties are typically required be well-calibrated. For a tool providing 95% confidence intervals for the outputs of an AI system, this would mean that the true value to be approximated by the model output (which is unknown during deployment) is contained in the provided interval for 95% of the test inputs. Thus, evidence should be provided that the provided confidence intervals have this property to a sufficient accuracy and precision.
- vi. Instructions on when and how to use the tool, including instructions on how to act upon observing the tool's output to fulfill its purpose.
- vii. A risk assessment including the discussion of possible failure modes of the tool and possible consequences of failures for the appropriate use of the AI system.
The following information *should* also be provided:
- viii. (A reference to) the technical specification of the interpretation tool.
- ix. Technical details of the experiments conducted to determine the accuracy and precision of the information provided by the tool and to determine its fitness for purpose.
- x. Details on the derivation of theoretical guarantees for the accuracy and precision of the information provided by the tool and for its fitness for purpose.

3. Conclusions

It is suggested that European standards covering the use of interpretation tools to address transparency and human oversight requirements should specify technical and non-technical requirements to ensure the fitness of such tools to provide quality control for AI systems, possibly lending inspiration from this document. Adherence to requirements such as those outlined here may be considered good practice not only for providers of high-risk AI systems but also in other contexts in which AI interpretation tools may be used (e.g., scientific studies).

References

1. <http://data.europa.eu/eli/reg/2024/1689/oj>
2. https://standards.cencenelec.eu/dyn/www/f?p=205:22:0:::FSP_ORG_ID,FSP_LANG_ID:2916257,25&cs=1827B89DA69577BF3631EE2B6070F207D
3. DIN SPEC 92001-3:2023-04, 2023. Artificial intelligence – life cycle processes and quality requirements – part 3: Explainability.
4. Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3), 457-506.

5. Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.
6. Weber, R. O., Johs, A. J., Goel, P., & Silva, J. M. (2024). XAI is in trouble. *AI Magazine*, 45(3), 300-316.
7. Haufe, S., Wilming, R., Clark, B., Zhumagambetov, R., Panknin, D., & Boubekki, A. Position: XAI needs formal notions of explanation correctness. In *Interpretable AI: Past, Present and Future*.
8. Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J. D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, 87, 96-110.
9. Wilming, R., Kieslich, L., Clark, B., & Haufe, S. (2023, July). Theoretical behavior of XAI methods in the presence of suppressor variables. In *International Conference on Machine Learning* (pp. 37091-37107). PMLR.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.