WHITE PAPER

Robustness and performance of AI applications

Trustworthy artificial intelligence

3 October 2023

Sebastian Scher*, Andreas Trügler*, Simone Kopeinik*, Tomislav Nad**, and Dominik Kowald*

* Know-Center GmbH

** SGS Digital Trusts Services GmbH

Partners of SGS

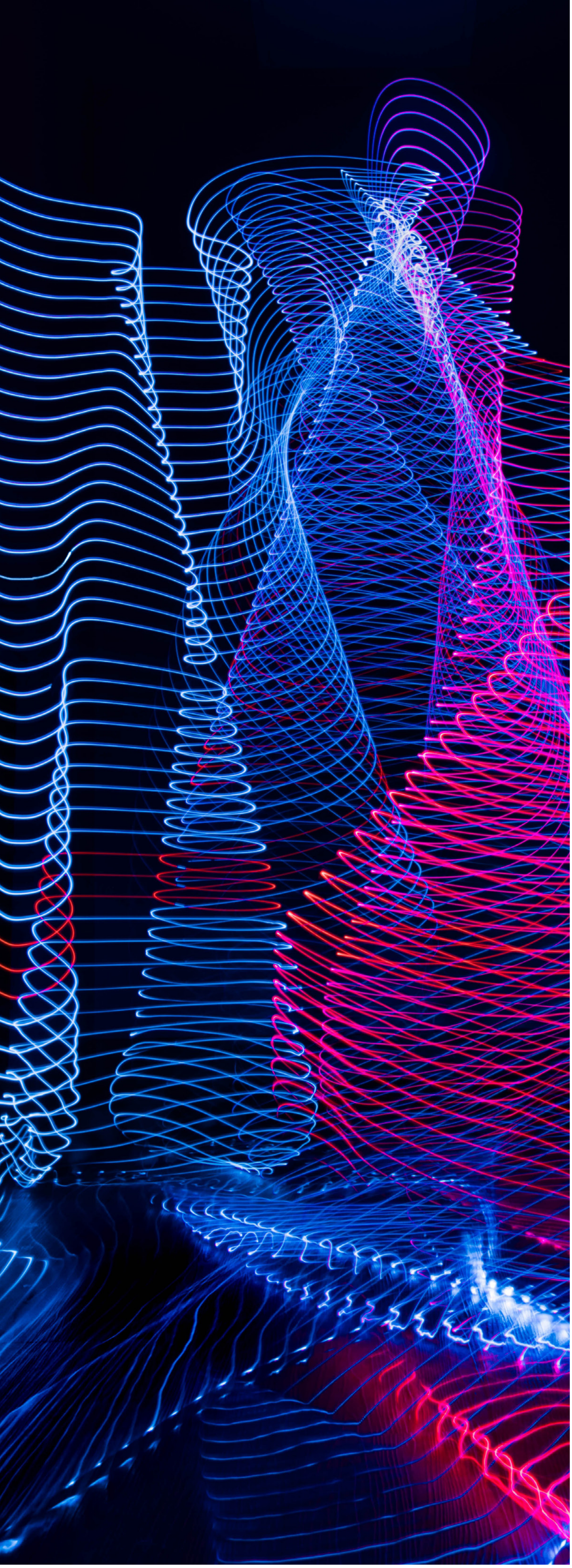
Contents:

1 Introduction

There has been an unprecedented surge in the use of artificial intelligence (AI) in recent years. AI systems are increasingly adapted in applications, both in industry and in everyday life. Regulators struggle to keep pace with the speed of development – and so does the field of testing and auditing.

AI systems are now being used in critical applications – in medicine where life is at stake [link], in factories where safety concerns are paramount, or in business settings where the wrong decisions can lead to financial losses.

In this white paper, two central topics of AI systems will be discussed, namely performance and robustness. An overview of current practice for the measurement of AI accuracy and robustness and how to ensure performance and robustness in applications will also be provided. In conclusion, an overview of open issues and challenges is given.



Definition of terms

For the purpose of this white paper, performance and robustness need to be defined in general terms.

Performance: performance describes how "good" the outputs (often referred to as "predictions") of AI systems are, e.g. whether they are correct, or how close the predictions are to true values. There are many different metrics available, for different applications and settings, which will be discussed later in this document ([Tharwat+21, Sokolova+06]). In everyday language, the term "accuracy" is often used interchangeably with "performance". In a technical sense, however, accuracy is just one particular performance metric.

Robustness: the robustness of an AI system describes how good the model performs in unexpected circumstances: "[Robustness describes]... risks that arise from a minor change or disturbance to a use case that the AI application is expected to be able to handle without error under normal circumstances" [FH23].

Relevance and potential risks

Having defined the principal terms, the question of why ensuring performance and robustness is essential when using AI systems needs to be addressed. Robustness, and performance even more so, are core requirements for any AI system. If it is not accurate, its main purpose is not met. Therefore, a non-accurate AI system is equal to a non-functioning system. Real-world examples of AI systems that are dangerous because of a lack of performance are AI-based COVID diagnosis tools that have widely been tested, but due to a lack of skill, did more harm than good. [link]

In general, AI systems which perform inadequately or fail to be robust, may also include risks for health, life or property, depending on their applications. For example, typical issues of AI systems that are untrustworthy also include the violation of consumer and data privacy, biased results, low reliability of model predictions, or the missed interpretability of results and of the models themselves. Also see National Institute of Standards and Technology (NIST) Taxonomy of AI risks and the EU Ethics guidelines for trustworthy AI in this regard.

Performance

How to measure performance

There are many different metrics available to measure the performance of AI systems. Principally, there are different types of AI algorithms, that are distinguished by the type of output they have. The most important ones are:

- Classification algorithms
- Regression algorithms
- Ranking algorithms
- Natural Language Processing (NLP) algorithms

For each of these application types, different performance metrics are applicable. The most well-known performance metric for classification applications is "accuracy". "Accuracy" in the literal sense is the fraction of outputs of the AI system that are correct. While this is a seemingly intuitive and easy-to-understand metric, it can be highly unintuitive and misleading for certain applications. For example, if the system tries to diagnose a rare disease that only 1% of the tested persons have, then a system that deems all persons as healthy would obviously not be useful – as it did not detect any of the sick cases – but still the accuracy metric would be 0.99 (or 99%), which would give the impression of a highly skillful system. Therefore, for such cases, different metrics are better suited, such as precision and recall. Many more metrics exist, e.g. F-score, ROC-curves, confusion matrices, etc. [Tharwat+21, Sokolova+06] and expert knowledge is usually required to choose the best metric for a specific application. Precision, for example, is a very useful measure for the success of prediction when classes are very imbalanced (e.g. fraud detection). Recall evaluates the model's ability to avoid false negatives and thus in medical diagnosis recall should be very high. For ranking applications on the other hand, recall might not be the best choice, since for a ranking of 1000 items, the position of the hits might also be important, where the Mean Reciprocal Rank would serve as a better metric. In order to determine whether an AI system is accurate and useful, it is thus very often not enough to report any single metric. Instead, for the given application and problem setting, the right metric(s) need to be determined. Only with the correctly chosen metrics can the accuracy of the AI application be judged. An overview of common performance metrics ([FH23]) is given in Fig. 1.

AI Applications

Type of AI application	**Regression**	**Classification**	**Ranking**	**Clustering**	**Computer vision**	**Language processing**
Performance measures	Mean squared error	Accuracy	Mean reciprocal rank	Silhouette value	Peak signal-to-noise ratio (SNR)	Perplexity score
	Mean absolute error	F1 score	Discounted cumulative gain	Adjusted mutual information score	Structural similarity index	Blue score
	…	Precision and recall	…	Completeness score	Intersection over index (IoU)	…
		Sensivity and specificity		…	…	
		AUC value				
		…				

Fig. 1 overview of performance metrics for different types of AI applications (based on [FH23])

In addition to choosing appropriate performance metrics, the right dataset to test the AI system needs to be carefully chosen from the whole dataset available – most of the data will be used for training, but a part needs to be held separately for evaluation. Deciding which part of the data to separate is a difficult task. There are many different approaches – e.g. random sampling – and again there is no general right or wrong approach; instead, it depends on the application. If an incorrect strategy is chosen for the application at hand, the accuracy scores will be misleading.

Thus, in order to generate trust in the performance of an AI system, (1) the metric – or metrics – appropriate for the application need to be chosen and (2) need to be computed on the correct dataset. Finally, it has to be shown that the value of the metric is good enough for the application – again something that can only be done on an application-by-application basis.

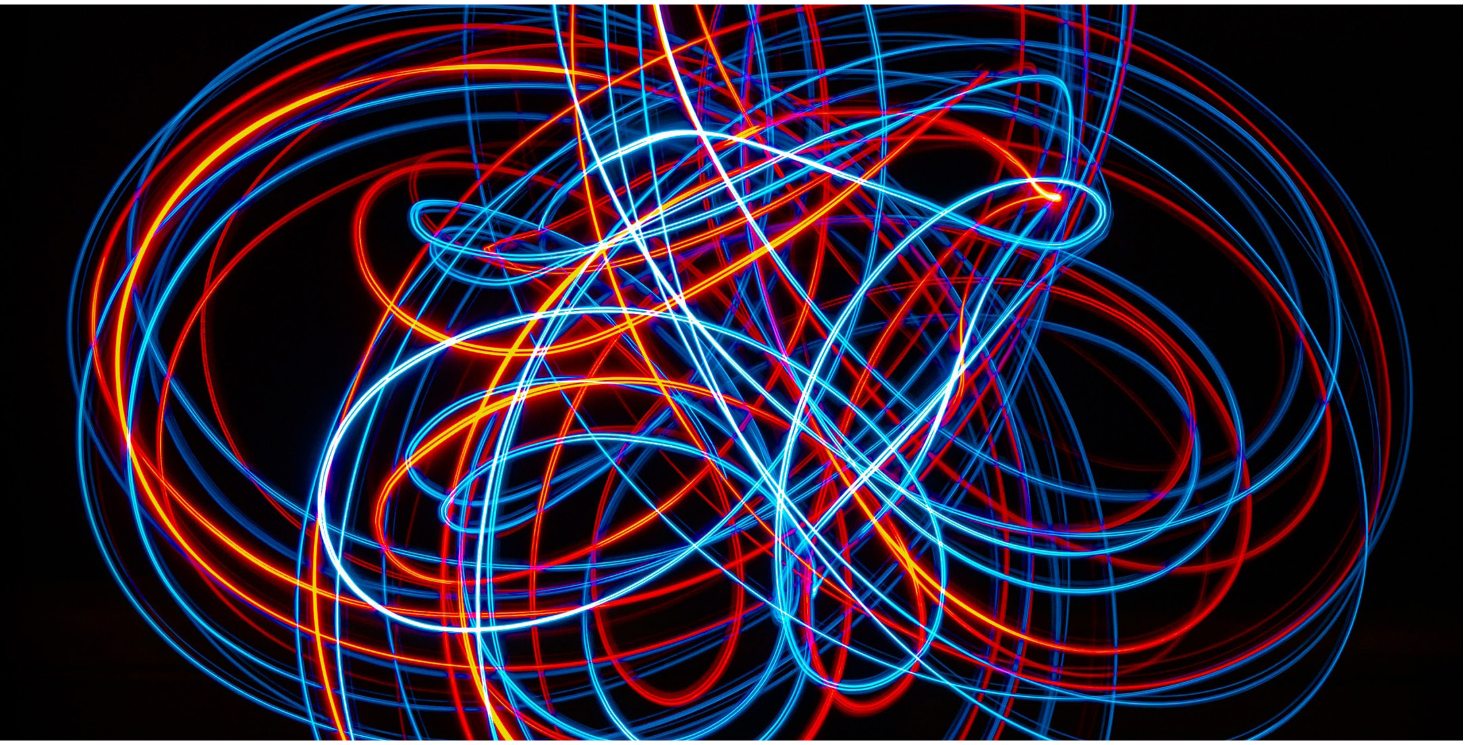
How to ensure best performance

Ensuring the best performance of an AI system goes hand in hand with best practices of (AI) application development. First, the intended task of the AI system needs to be clearly defined ([FH23]):

- Which problem does the AI application solve? (What exactly does it "do"?)
- What input data is provided and of what type is it?
- What are the outputs of the AI application and of what type are they?

After defining the task of the AI system, there are two key aspects to get an accurate AI system: data quality, and correct model training and selection. Data quality is a broad topic, and always needs to be viewed in the context of the application at hand, as stated in the EU AI Act: "Training, validation and testing data sets should be sufficiently relevant, representative and free of errors and complete in view of the intended purpose of the system." (AI ACT). Model training and selection consists, among other things, of correct data splitting and evaluation approaches.

Only with the correctly chosen metrics can the accuracy of the **AI** application be judged.



3 Robustness

The robustness of AI applications is a complex and wide topic, as there are many different aspects of it. Each one of them is a whole research field on its own. Here is a brief overview of them.

Common corruptions and perturbation robustness describes how much (or little) the output of an AI system changes, if the inputs are slightly altered in a random way [Hendrycks+19]. This could, for example, be small errors in images.

Robustness in relation to distribution shift describes how well an AI system works in the case that the data used for training the model does not completely represent the data that will be used in the deployed system. This is a widely studied subject (e.g. [QSS+08], [TE11], [FTP21], [HBK+21]). An example would be a system trained on data from a specific camera, which is then used on data from another camera type.

Domain generalization deals with issues which are similar – but not identical to – robustness in relation to distribution shift. Here the question is whether a system trained on one domain (i.e. a certain application) also works for another domain [MBS13].

Finally, **adversarial robustness** is robustness in relation to so-called adversarial attacks. These are perturbations of the input data that are created specifically to "fool" the model, thus making the model produce an incorrect output [XML+20, Yuan+19]. An example would be small modifications to traffic signs, that cause an autonomous vehicle to misidentify them.

How to measure robustness

The robustness of predictions describes in general how good the model predictions are in unexpected circumstances and includes risks of erroneous AI behavior that arise from a minor change or disturbance to a use case [FH23]. What these minor changes or "unexpected circumstances" are, depends on the type of robustness considered. Perturbation robustness for example is measured through applying random perturbations to the test data, and measures how this changes the performance of the model. The same approach holds for robustness in relation to common corruptions, where typical perturbations can be applied to the test data, and again the corresponding influence on the performance of the model is measured.

Robustness in relation to distribution shifts is usually measured by comparing models with in- and out-of-distribution data based on performance metrics (e.g. [KSM+21]). To determine domain generalization on a test domain, three strategies are proposed in the literature [WLL+22]:

- Test-domain validation: which utilizes parts of the target domain as validations
- Leave-one-domain-out cross-validation: which works when training data contains multiple sources, and leaves one training source as the validation while treating the others as the training part
- Training-domain validation: where each source is split into two parts: the training part and the validation part. The most common strategy is where all training parts are combined for training while all validation parts are combined for selecting the best model

How to ensure robustness

Perturbation robustness can be increased with adapted training, which modifies the training algorithm by including typical perturbations in the training [RSZ+20]. The same approach also works for common corruptions [RSZ+20]. For domain generalization a wide range of methods exist ([GL20], [WLL+22]) which can be grouped into three categories known as data manipulation, representation learning and adapted learning strategies. With respect to noise in training data, there are no specific methods for increasing robustness in relation to this kind of noise, apart from general best practices for model development. Adversarial robustness can be ensured with adversarial training [Weng+18], where adversarial examples are generated during the training process, and included as new training material, so that the model learns to classify the adversarial examples correctly. This makes the system resistant to the attack method that was used to create the adversarial examples during the training, however, the trained model might still be vulnerable to unseen attacks.

4 Open issues and current limitations

The topic of performance evaluation metrics is very well established in statistics and machine learning, and computing them is not difficult. However, the choice of metrics, and defining thresholds for the metrics, is much less straightforward. While best practices exist, no formal guidelines are available. Additionally, despite the wide range of established performance metrics, highly specific applications might need new performance metrics specifically developed. Ensuring performance is the core of machine learning and artificial intelligence, and thus, methods for ensuring that AI applications are accurate go hand-in-hand with the development of and improvement of machine learning (ML) algorithms.

For robustness, there are many more open issues and limitations. Robustness is a broad term, and current research does not necessarily cover all aspects. Sometimes no specific methods (apart from best practices) are available for increasing robustness (e.g. robustness in relation to noise). A particular problem of adversarial attacks is the "cat and mouse game": if a specific attack method is known, AI systems can be made robust by including the attack in the training procedure – known as adversarial training. However, this does not guarantee that the system will also be robust in the event of a new attack that has not been considered in the training.

An additional open issue is how to assure performance and robustness when multiple AI components are combined, or in situations where an AI evaluation is part of a larger product/solution, or where performance and robustness need to be assured in evolving (learning) AI systems that are constantly updated (in some cases with every single use).

In general data quality is very important for AI systems and influences performance and robustness. Currently no unified quality concept is available, but basic automated tests are possible. Additionally, model training and selection influence performance and robustness, where best practices are at least available.

5 Summary

In this white paper, a brief overview of issues surrounding the performance and robustness of AI systems has been provided. Both performance and robustness have been defined and evaluation best practices have been discussed. For ensuring performance, the challenges are not so much on the technical side – if one knows what one wants to measure, it is relatively straightforward to measure it. The difficult part is to know which measures and metrics are relevant for the application at hand. Issues related to robustness, on the other hand, are the subject of ongoing research (e.g. adversarial training).

6 Acknowledgements

Know-Center is a leading European research center for big data, artificial intelligence (AI) and data-driven business models. Know-Center is a COMET Centre within COMET – Competence Centers for Excellent Technologies. This program is funded by the Austrian Federal Ministries for Climate Policy, Environment, Energy, Mobility, Innovation and Technology (BMK) and for Labor and Economy (BMAW), represented by Österreichische Forschungsförderungsgesellschaft mbH (FFG), Steirische Wirtschaftsförderungsgesellschaft mbH (SFG) and the Province of Styria, Wirtschaftsagentur Vienna and Standortagentur Tyrol GmbH.

References

[FH23] Fraunhofer IAIS, Guideline for Designing Trustworthy Artificial Intelligence – AI Assessment Catalog, 2023. *https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefkatalog/Fraunhofer_IAIS_AI_ASSESSMENT_Catalog_Web.pdf*

[FTP21] Federici, M., Tomioka, R. and Forré, P. (2021). An information-theoretic approach to distribution shifts. Advances in Neural Information Processing Systems, 34, 17628-17641.

[Hendrycks+19] Hendrycks D., Dietterich T. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations, ICLR 2019.
https://arxiv.org/abs/1903.12261

[MBS13] Muandet, K., Balduzzi, D. and Schölkopf, B. (2013, February). Domain generalization via invariant feature representation. In International Conference on Machine Learning (pp. 10-18). PMLR.

[Sokolova+06] Sokolova, M., Japkowicz, N., Szpakowicz, S. (2006). Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In: Sattar, A., Kang, Bh. (eds), AI 2006: Advances in Artificial Intelligence. AI 2006. Lecture Notes in Computer Science, vol 4304. Springer, Berlin, Heidelberg.
https://doi.org/10.1007/11941439_114

[TE11] Torralba, A. and Efros, A. A. (2011, June). Unbiased look at dataset bias. In CVPR 2011 (pp. 1521-1528). IEEE.

[Tharwat+21] Tharwat, A. (2021), "Classification assessment methods", Applied Computing and Informatics, Vol. 17 No. 1, pp. 168-192. *https://doi.org/10.1016/j.aci.2018.08.003*

[HBK+21] Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., ... & Gilmer, J. (2021). The many faces of robustness: A critical analysis of out-of-distribution generalization. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 8340-8349).

[XML+20] Xu, H., Ma, Y., Liu, H.-C., Deb, D., Liu, H., Tang, J.-L. and Jain, A. K. (2020). Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. International Journal of Automation and Computing, 17(2), 151–178.

[Yuan+19] Yuan X., He P., Zhu Q. and Li X., "Adversarial Examples: Attacks and Defenses for Deep Learning," in IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 9, pp. 2805-2824, September 2019, doi: 10.1109/TNNLS.2018.2886017.
https://ieeexplore.ieee.org/ielaam/5962385/8809853/8611298-aam.pdf

Robustness and performance of AI applications

SGS.COM/DIGITAL

CONTACT US

SGS

Emerging Technology

✉ Enquiry.Emerging-Technology@sgs.com

KNOW CENTER

Leading Research and Innovation Center for Trustworthy AI

🌐 https://know-center.at/

✉ info@know-center.at

WHEN YOU NEED TO BE SURE