Towards the Certification of AI-based Systems

Philipp Denzel*, Stefan Brunner*, Yann Billeter,
Oliver Forster, Carmen Frischknecht-Gruber, Monika Reif,
Frank-Peter Schilling, Joanna Weng, and Ricardo Chavarriaga
* Co-first author.
Zurich University of Applied Sciences ZHAW
Winterthur, Switzerland
{denp,brrt,bily,foro,frsh,reif,scik,wenj,char}@zhaw.ch

Amin Amini, Marco Repetto, and Arman Iranfar
CertX AG
Fribourg, Switzerland
{amin.amini,marco.repetto,arman.iranfar}@certx.com

***Abstract*—Certifying the trustworthiness of Artificial Intelligence (AI)-based systems based on dimensions including reliability and transparency is crucial given their increased uptake. Likewise, as regulatory requirements are established, actionable guidelines for certification will be useful for developers and certification bodies to ensure trustworthiness of AI. Here, we present an ongoing effort to develop a validated AI certification scheme which is a framework for assessing the trustworthiness of AI systems including specific objectives with their corresponding means of compliance (i.e. process, documentation or technical methods). Importantly, the scheme makes an explicit link between legal requirements and validated techniques for assessing the compliance of AI systems, resulting in the implementation of a workflow to support AI certification. We explain the rationale for developing the certification scheme and demonstrate the assessment of an example use case with a concrete workflow traversing from objectives to corresponding means, focused on reliability and transparency.**

***Index Terms*—Artificial Intelligence, Machine Learning, Certification, Reliability, Transparency**

I. Introduction

Artificial intelligence (AI) and machine learning (ML) are transforming numerous domains, including those with significant implications for human safety, asset protection, and environmental impact. Ensuring the safe and reliable use of AI technology is therefore of utmost importance. Although there is a large number of ethical guidelines for AI [1], their translation into specific methods and practices remains poorly defined. We address this gap by proposing a certification scheme for AI systems. This scheme consists of two parts:

(i) a framework outlining existing regulatory requirements, criteria, and measures needed to certify an AI/ML system;
(ii) guidance linking these requirements to a set of technical and scientific methods for assessing certification-relevant characteristics of AI/ML systems including, among others, reliability and transparency.

Altogether, the certification scheme provides a comprehensive workflow to identify and apply methods and processes for assessing compliance with emerging AI regulations. This work is based on existing efforts on the standardization and regulation of AI systems, and on initial groundwork for AI certification processes. Complementarily, it also draws on recent work in the AI/ML research community on algorithmic methods for determining and verifying relevant properties of ML models.

In the following sections, we provide an overview of existing norms, standards (Section II), and related work (Section III). We then describe our approach for developing the certification scheme (Section IV), and give a brief summary of existing technical methods for assessing aspects of transparency and reliability of AI systems (Section V), complemented by an example of their implementation within a concrete assessment workflow (Section VI).

II. Overview of AI legislation and standards

There is a global push for regulatory frameworks for AI. The European Union (EU) has taken a leading role in this, launching efforts for establishing a comprehensive regulatory framework for AI systems in the EU (AI Act) [2]. In fact, as of December 2023, the European Parliament and Council agreed upon a first draft of a legislative proposal [3].

The proposed legislation classifies AI systems into four risk categories based on their potential harm to individuals, society, and the environment: *unacceptable risk, high risk, limited risk*, and *minimal risk*. AI practices with unacceptable risk, in particular biometric classification, emotion recognition in the workplace, subliminal manipulation to circumvent free will, and social scoring will be prohibited by the legislation, while high-risk AI systems will be subject to additional obligations, such as an assessment of the impact on fundamental rights, especially with regard to the environment and marginalized groups. Moreover, certain AI systems intended for use in high-risk sectors, healthcare, transport, and energy among others, must undergo a conformity assessment process and be certified by a designated third-party assessment body before they can be placed on the EU market. The certification process will assess the AI system's compliance with mandatory requirements such as transparency, accuracy, and robustness, as well as the potential risks associated with its use. For limited-risk AI applications there will be an essential obligation to be transparent about their purpose, function, and decision-making processes. Similar regulatory frameworks are expected to be introduced in other regions of the world. Indeed, in October 2023, the US followed suit by issuing an executive order on the development of new standards for safe, secure, and trustworthy AI [4].

Standards support binding laws and regulations by documenting the state of the art and best practices, and providing a basis for demonstrating compliance and certification. Several organizations and initiatives such as ISO/IEC, IEEE and NIST are currently working on developing relevant AI standards. The standards developed by ISO/IEC [5] cover a wide range of AI aspects, including terminology, performance metrics, data quality, ethics, and human-AI interaction. The IEEE P7000 series of standards [6] focuses on the ethical implications of AI technologies, while the NIST framework [7] provides guidance on managing risks, ensuring data quality, and promoting transparency and accountability in AI systems.

III. Related work

Various national and international organizations are working on initiatives to support the certification of AI systems. DIN/DKE provide comprehensive recommendations for standardization across all AI topics to establish a common language, principles for development, use, and certification [8]. In this context, the Fraunhofer Institute has developed a guideline for assessing the trustworthiness of AI [9] which proposes to base such evaluations on six dimensions: fairness, autonomy & control, transparency, reliability, safety & security, and privacy. The LNE's AI certification programme provides objective criteria for the selection of trustworthy AI systems based on their evaluation according to compliance with a set of criteria related to ethics, safety, transparency and privacy [10]. Similarly, the IEEE has established a certification programme to assess the transparency, accountability, bias and privacy of AI-related processes [11]. While these initiatives provide detailed description of processes, they remain rather vague when it comes to the evaluation of AI systems with concrete technical methods.

EASA is the first European agency to publish guidance on the safe use of ML [12]. Its purpose is to assist aviation stakeholders in the development and implementation of ML systems with low levels of automation and covers the entire life cycle, including topics such as the development process, data collection, model selection and testing, and the use of ML operations (MLOps). As such, it serves as one of the main precursors to the development of a comprehensive AI certification scheme outlined in this contribution.

Companies such as IBM, Meta, Seldon, Microsoft, and Google offer open source toolboxes for AI evaluation such as [13]–[20]. These toolboxes aim to detect and mitigate bias, explain model decisions, ensure the robustness, and assess the uncertainty of AI systems.

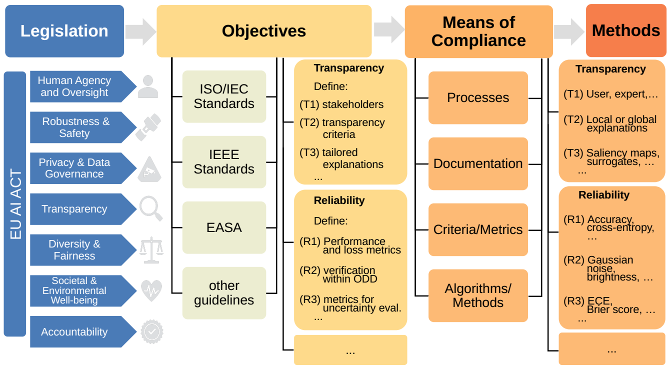


Fig. 1: Certification scheme for AI systems. It enables identification of application-specific requirements, respective means of compliance and technical methods for conformity assessment (see text).

IV. Towards a Certification Scheme

Our goal is to design an actionable guide for regulators and developers to be able to assess and certify the trustworthiness of AI-based products and their associated processes, including data, development, model, testing, and (continuous) operation. This requires considering all stakeholders, including users, developers, auditors, and authorities, as well as the entire life cycle of the AI-system from its conception to its deployment, and in principle even its decommissioning.

While numerous guidelines for the establishment of responsible AI exist (see Section III), most lack detailed prescriptions for actual measures to guarantee trustworthiness. The majority of these guidelines base the definition of trustworthiness on the same ethical principles which only differ in the importance with which they weight each dimension. Our research provides clear guidance on means for assessing the following trustworthiness dimensions:

- *Reliability* ensures that the AI system can operate effectively under various conditions and resist (input) errors, bias or malicious attacks.
- *Transparency* ensures that different stakeholders can understand the decision-making process of an AI-system and thus can judge the extent to which they can trust it.
- *Safety* is crucial for AI systems in critical applications like healthcare or self driving vehicles to prevent any harm or unintended consequences.
- *Autonomy and control* is related to the autonomy of the AI system ("Human-in/on-the-Loop" vs "Complete Autonomy" [21]) as well as the level of support of human decision making.

For brevity, in this paper we focus in the following on the reliability and transparency dimensions.

The certification scheme is illustrated in Fig. 1. The scheme identifies application-specific *objectives* for achieving conformity with the requirements of EU legislation. Then, the associated *means of compliance* are identified starting from the guidelines summarized in Section III and taking also the state of the art into account. Currently, the certification scheme comprises more than 230 *objectives* for certification and about 250 *means of compliance*, based on the analysis of 38 documents on regulations and standards.

Corresponding processes, algorithms, and technical *methods* (see next section) are identified in the next step, which were tested and validated to assess the certification objectives. The choice of appropriate methods can vary significantly based on several factors, including the data type, model type, the phase of the life cycle, number of components in the AI system, and the stakeholders involved. The certification scheme incorporates all of these factors and provides a clear guidance on how to proceed.

To confirm the suitability of the technical and algorithmic methods for not just an academic setting, but for certification of real-world applications, we validate the methodology on use cases featuring actual AI applications provided by industrial partners[1]. This validation process ensures the certification scheme is comprehensive, reliable, and effective.

Among the means of compliance for the existing regulatory objectives, there are in particular:

- use of appropriate quantifiable metrics or qualitative criteria refining the objective;
- documentation of system design, implementation, validation, deployment and monitoring;
- adherence to processes defined by best practices and guidelines;
- use of technical and algorithmic methods for verification and testing of certain properties of the AI system.

V. Technical Methods for AI Certification

The term technical and algorithmic method refers to methods discussed by the AI scientific literature, open-source code implementing these algorithms (e.g. code repositories accompanying publications, or reference implementations by third parties), or commercial and non-commercial software solutions for verification and testing of AI systems. The AI research community is actively engaged in proposing and advancing such technical and algorithmic methods to assess the behavior, robustness, interpretability, and overall trustworthiness of AI systems.

In particular, research in methods to assess AI systems along the transparency and reliability dimensions is rapidly advancing. Methods for transparency generally assess the behavior of the model with respect to the interpretability of its input and output, as well as the human comprehensibility of the model's internal mechanisms leading to predictions [22]. Regarding reliability, areas of concern include model accuracy and uncertainty, dataset coverage or bias, model generalization properties, and robustness with respect to noise, labeling errors or adversarial attacks [23].

Despite considerable efforts from the communities in these fields, many problems still persist for the certification of AI systems. Since most systems requiring certification are deployed in real-world environments, the suitability of the technical methods is barely known. Mostly due to the lack of realistic benchmarks, new algorithms are often tested in an academic setting against idealized datasets.

A. Methods for Transparency

ML techniques like Deep Learning (DL) scale well with the amount of data which feed their learning process. This scalability often comes at the cost of increased model complexity, which in turn makes the rationale behind the models' predictions opaque to users and experts alike. Yet, certification will include objectives for model interpretability and explainability, related to the capacity of providing explanations for different stakeholders, e.g., users, affected citizens, domain experts, or government regulators [2].

Moreover, different types of models and data modalities potentially demand different methods or variations thereof.

Contrary to intrinsically interpretable models such as decision trees or linear regression, models (e.g. DL models) whose internal behavior and predictions cannot be understood by looking at their parameters, are considered *black-box models* in this context. Consequently, they require a additional methods to gain insight into their decision-making process.

These techniques for instance analyze the dependence of the model's output on the data feature statistics, visualize model internals, or find *surrogate* models, interpretable models which approximate the model's behavior *locally*, on single data samples, or *globally*, on data features in general and the entirety of the data domain.

A comprehensive taxonomy of well-established explainability methods provides organizing principles to identify appropriate methods for a given problem. E.g., [22], [24] propose such a taxonomy encompassing many well-established methods in ML research. Both schemes are presented in the form of decision trees and essentially only differ in the sequence of their questions, thus we list them here in arbitrary order:

(i) *Data or model?* Does the explanation pertain to the data or the model? Data explanations can take form as samples [25], features [26], or distributions.

(ii) *Intrinsic or post-hoc?* Intrinsically interpretable models usually have no need for additional explanations, whereas black-box models involve auxiliary methods employed after training.

(iii) *Local or global?* Does the explanation need to describe the behavior of the entire model (global) or an individual prediction (local)? Can the explanations take into account a certain class of model? E.g., methods which work only for neural networks are model-specific, whereas model-agnostic methods assume a black-box model and analyze the relation between input and output without access to the model's internal structure.

(iv) *Interpretation type?* Some cases require a specific form of explanation such as feature summary statistics [27], saliency or attribution maps [28]–[30], visualizations of features or model internals [31], exemplar explanations, so-called prototypes [25], or explanations through surrogates [32].

These taxonomies facilitate the classification and search for transparency methods that meet the requirements of the certification scheme. Selected methods, in combination with explainability metrics aid in the identification of explainability

[1] to secure their anonymity upon request, the use case presented in this paper is completely unrelated to these test cases.

[2] Note that in the literature some make a clear distinction between the terms interpretable and explainable; here we use the terms interchangeably

gaps for users and developers [33]; many are implemented in open-source toolkits, e.g. [16]–[18]. The research field of explainable AI is rapidly advancing [34], [35]. However, research in corresponding metrics is still lacking.

B. Methods for Reliability

In the context of reliability, it is essential to encompass the entire intended operational domain to ensure that an AI system will behave reliably under all possible circumstances that it may encounter. This scope can be effectively described by the Operational Design Domain (ODD), a concept from the automotive sector as defined by SAE International [36], which we extend here to other AI applications. The ODD delineates the specific conditions under which an AI system is designed to operate safely, robustly and accurately, including inherent or environmental disturbances such as electronic noise in camera sensors, lens flare in image or video processing scenarios, and environmental conditions such as illumination.

A fully covered ODD is prerequisite for the verification of the AI system's reliability. The examination of its ability to perform accurately, robustly, and in difficult situations is, however, heavily dependent on the derivation of meaningful tests. For this purpose, the proposed scheme provides a set of methods for assessing performance, robustness, and uncertainty. In this context, reliability relates to the consistent and *accurate performance* across various scenarios. *Robustness* encapsulates the system's ability to resist against perturbations, be they intentional attacks [23], such as (synthetic) adversarial examples, inherent variations in the input data , or environmental perturbations [37]. Precise *uncertainty* quantification is another critical aspect for assessing reliability of an AI system, estimating the model's limitation to encompass the deep, intrinsic complexity and dynamics of real-world environments.

For assessing performance and robustness of AI systems, confidence interval (CI) tests [38] can be conducted. They evaluate the model based on test data featuring different types of noise simulations over a range of amplitudes. Alternatively, abstract interpretation as in [39], [40] also tests robustness with the advantage that these methods need significantly less computing time compared to the CI tests.

While there are numerous metrics for evaluating performance or robustness (e.g. accuracy, recall, mAP), where the choice depends strongly on the application, the metrics specifically for uncertainty estimation are not as widely established. In the context of the certification scheme, we will focus on the following selected metrics:

- *Calibration* is measured by the so-called reliability diagram [41], [42]. The diagram plots the accuracy as a function of model confidence for different intervals. Here, accuracy measures the mean of the correct predictions for several interval thresholds, while confidence measures the mean of the predicted probabilities for these thresholds. As a metric, it assesses whether the AI system has the same confidence level as the prediction accuracy.
- *Expected Calibration Error* (ECE) [43]. It derives the average between the absolute difference of the accuracy and confidence for an ML model for different thresholds bins. An ECE score close to zero implies a correct model calibration, while larger values indicate a discrepancy.
- *(Multi-class) Brier score* [44] is another metric for the assessment of the model certainty which measures the accuracy of probabilistic predictions of the AI system, with a score of zero indicating optimal model calibration.

Alternatively, the intrinsic and extrinsic uncertainty of AI systems can be assessed using Bayesian neural networks or dropout at test-time [45] in combination with Monte-Carlo sampling and uncertainty decomposition [46].

C. Putting Certification into Practice

The assessment of the means of compliance requires suitable and efficient selection and implementation of the selected methods as described in Section IV. In the context of ML, the paradigm of MLOps aligns nicely with this approach and provides effective tools towards an implementation of such a scheme [47]. An MLOps pipeline encompassing both ML methods and data enables complete experiment tracking, and contributes to the reproducibility and auditability of the AI product. Such a pipeline can be designed with various degrees of automation. Regulatory requirements, in particular for high-risk applications, have identified pipeline orchestration, model tracking, including versioning, and data versioning as imperative key components.

While certainly not the only choices, we are addressing these requirements using the following tools:

- Orchestration of training and testing is carried out through GitHub Actions and Airflow (https://airflow.apache.org/).
- MLflow provides an analytics framework, giving insights into the (hyper)parameters of the model and the method at hand (https://mlflow.org/).
- Oxen was chosen for data version control and dataset management (https://www.oxen.ai/).

Note that there are complete, cloud-based services providing a similar feature set as the aforementioned pipeline. However, for many companies, privacy concerns prohibit the use of external facilities due to the potential loss of data ownership. The described pipeline on the other hand is containerized and can easily run on any cluster or workstation, enabling an in-house assessment.

VI. Practical Example

We illustrate the workflow of the certification scheme on a small subset of objectives with a practical example of a system for classifying dermatoscopic images of skin lesions using a convolutional neural network (CNN) model.

Regulatory requirements for systems performing medical diagnoses including an AI component will demand the definition of an appropriate degree of transparency and reliability. As a full assessment would be too extensive to be included here, we report only a few exemplary requirements for each dimension, see Table I.

TABLE I: Three examples for transparency and reliability objectives specified in the certification scheme, each applied to the skin lesion classification use case.

	Objective	Means of compliance	Practical example
Transparency	(T1) Define stakeholders.	Consider all relevant stakeholders. Document the choice.	Physicians, developer.
	(T2) Define necessary transparency criteria.	Document the criteria to be used for evaluating the transparency: - Scope, design, and degree of transparency of the procedures. - Depth and breadth of introspection considering the model outputs.	Explanations: - Local (physician, developer) - Global (developer)
	(T3) Define suitable transparency methods.	Define suitable set of transparency methods. Justify selection.	SHAP, CAM
	. . .		
Reliability	(R1) Define performance and loss metrics.	Define suitable set of performance and loss metrics. Justify selection.	Accuracy, Cross-Entropy Loss
	(R2) Perform and document verification of the trained model within the ODD.	Execute tests on ideal data. Execute tests including test cases covering: - Perturbations due to fluctuations in the input (e.g. noise on sensors). - Edge cases that can arise on the data within the ODD (e.g. light). - Combined effects.	Perturbations: - Gaussian noise - Brightness - Brightness + Gaussian noise
	(R3) Define metrics for uncertainty evaluation.	Define suitable set of metrics. Justify selection.	ECE, Brier score
	. . .		

For any use case, the data modality, model type, and its task dictate to a large extent the appropriate means of compliance for its objectives. Thus, it is important that rationales and justifications for the choice of training data and model architectures are properly documented.

For this application we used the ISIC2019 dataset [48], which consists of 25,331 dermatoscopic images from nine diagnostic categories, e.g. *melanoma and basal cell carcinoma, vascular lesions*. These images show variations in lesion characteristics such as asymmetry, margin, color and diameter, and present challenges such as class imbalance, unknown test set classes, and different image resolutions and imaging protocols.

For the AI component we used the small EfficientNetV2 (EfficientNetV2B0) architecture [49] as feature extraction backbone and extended it by a fully connected network as classification head. The model was initialized with the pre-trained ImageNet weights and it was trained for 25 transfer-learning and 10 fine-tuning epochs. Note that the objective in this practical example is not necessarily to reach state-of-the-art performance of the AI component in the task at hand, but rather to assess the certification requirements of an AI model likely to be deployed for production.

A. Transparency

As first objective in the certification scheme, all stakeholders and the form of explanation each of them need about the AI system have to be identified (T1). This dictates the selection of suitable methods for assessing compliance. This application may require the model's predictions and decisions to be explainable only to domain experts, i.e. physicians. As these decisions directly affect patients, physicians cannot rely on predictions of black-box models without proper insight into the inference process (T2).

From this objective, the certification scheme identifies the technical methods necessary to assess this requirement (cf. Table I).

Since a DL model is used, which is not intrinsically explainable, post-hoc methods are required. Here, feature attribution maps can be used to produce explanations. These methods associate input features to how they perturb the prediction. LIME [32] and SHAP [29] are well-established methods for this purpose.

The effect of each pixel in the image on the output can generally be explored by perturbing the pixels as above, or by directly computing gradients of the prediction relative to the pixels, e.g. with [28], [30]. In all cases the result is an image; often called *saliency*, *attribution*, or *gradient* map (commonly superimposed over the original input image). This map indicates the regions of the image that supported the model's prediction, specifically for this use case, what visual characteristics led to the classification of the skin lesion as malignant or benign (T3).

In Figure 2, we show results of the SHAP (SHapley Additive exPlanations) method applied on a sample of a malignant melanoma. The explanation consists of a map with *Shapely values* for each pixel: a number centered around 0, negative values detrimental to, and positive values supportive of a specified target prediction. As such, SHAP is relatively intuitive to interpret for any lay person. Concretely, SHAP thus assists physician to discern what regions and properties such as shape or coloration of the imaged skin lesion most impacted the AI component's decision, be it in favor or against malignancy.

Although a physician may be most interested in local explanations during the operation of the AI component, another requirement for transparency is a thorough, statistical evaluation of the model behavior and documentation thereof. The certification scheme offers multiple options for technical methods to comply with this requirement.

If the model grants access to its internal structure, it is possible to use class activation maximization (CAM) [50] to retrieve a global explanation of the model behavior in form of an activation map. CAM is based on the gradient

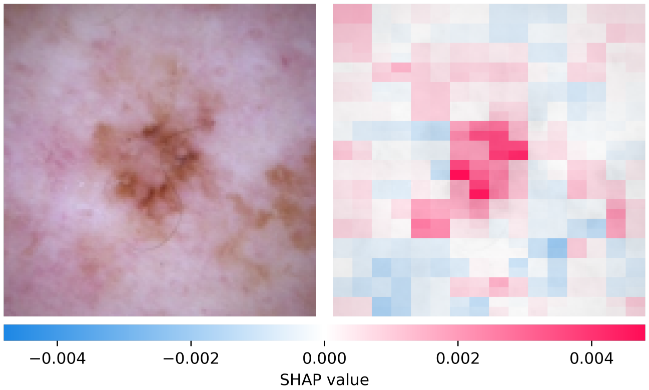


Fig. 2: The left panel shows a sample of a malignant melanoma and the corresponding Shapely value patch map on the right. A malignant lesion typically has a more asymmetric shape, a less defined border, and an irregular color profile. The Shapely values seem to support these heuristics overall.

ascent technique which is used in CNN feature visualization to maximize the response of a particular feature map. This is useful for the developer or assessor of the AI component to better understand the model's internal decision-making processes. In Figure 3, such activation maps are shown for two model variants, one properly trained (a), the other exhibiting unsatisfactory behavior (b). This highlights the practicability of this method for debugging and model assessment during the training or optimization phase of CNNs.

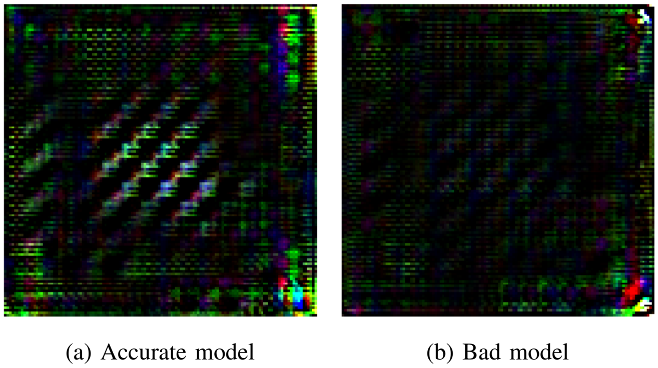
Alternatively, it could be sufficient to verify the expected behavior of a model on a small subsample of the test set which is representative of the whole data statistics. In such a context, the subsample instances are called *prototypes* and can in principle be extracted using various clustering algorithms, although ideally, prototypes are accompanied with *criticisms*, measuring how well a certain instance is represented by the prototypes; algorithms like ProtoDash [25] are more suited in this case. In combination with methods such as SHAP, the global behavior of a model can then be estimated through their evaluation on all prototypes and criticisms.

B. Reliability

In this example, there are numerous perturbations that pertain to the ODD, which should be covered according to the objectives defined in Table I (R2). We report two of them: one arising from the camera hardware (Gaussian Noise, GN) and another caused by environmental factors (Brightness Change, BC). Table II lists the parameter intervals for each perturbation that can be expected in real-world scenarios. These intervals were used for perturbation simulations on the ISIC2019 test set and all subsequent tests.

To illustrate the effect of the perturbations and their combination, Figure 4 shows samples with increasing noise or brightness from left to right.

For each type of perturbation, we applied various perturbation levels represented by different parameter values (e.g., with a step size of 0.025 for GN) in order to cover a wide range of potential scenarios within the ODD. Regarding BC, we randomized the BCs for each pixel within a five percent deviation from the chosen level to create a more realistic perturbation. In the CI tests, we conducted 1000 simulations for each of these perturbation levels on every sample in the test set, employing the specified noise and/or brightness parameter value. Based on the simulated perturbation statistics, the mean accuracy and the 2-sided 95% CI were computed. Table III exemplifies the results for two different perturbation levels for each perturbation type.



(a) Accurate model (b) Bad model

Fig. 3: Panel (a) shows the CAM map for a malignant classification of a properly trained model, and (b) of an improperly configured model. These maps explain the internal features of the model with arbitrary input. For this specific use case, it is evident that certain color channels in the central region of the image have the most impact in a malignant classification.

TABLE II: Definition of perturbations relevant to the ODD with specific parameter intervals for each perturbation type, taking into account combined perturbations.

Description	**Perturbation type**	**Parameter interval**
White noise on camera	GN $\mathcal{N}(\mu = 0.0, \sigma)$	$\sigma \in [0; 0.1]$
Different illumination	BC	$b \in [-0.3; 0.3]$
Perturbations combined	Both combined	$\sigma \wedge b$

This CI serves as an indicator of the robustness of the AI system. According to the objectives (R1), specific appropriate performance metrics (and their goals) are to be defined. Thus, the performance of the model can be evaluated by comparing the performance goals (e.g., for accuracy in this example) with the test results in Table III.

These tests also provide the basis for evaluating the uncertainty of the AI system. They allow both a graphical evaluation using the reliability diagram and a metric-based evaluation. The reliability diagram is derived based on the unperturbed test set, while the applied metrics are also evaluated based on the different perturbation levels.

The reliability diagram (a) in Fig. 5 illustrates the model calibration accuracy. The proximity of the results to the green dotted line indicates good calibration, while being below it reflects overconfidence compared to actual accuracy. In addition, the confidence diagram (b) gives further information about the distribution of samples for different confidence levels within the test set. Thereby, it becomes evident that the evaluated

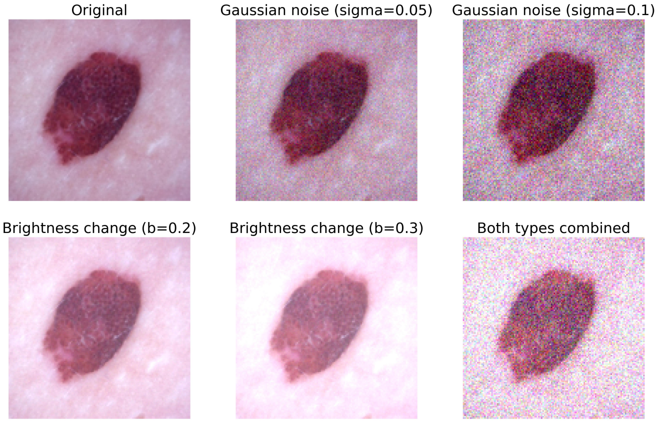

Original
Gaussian noise (sigma=0.05)
Gaussian noise (sigma=0.1)
Brightness change (b=0.2)
Brightness change (b=0.3)
Both types combined

Fig. 4: For skin lesion detection, a dermatology camera is typically used, where perturbations such as the GN (caused by electronic circuits) or BC (caused by environmental factors) occur. All relevant types of perturbations can arise at the same time.

TABLE III: Results of the CI tests (simulated 1000 times for each sample of the test set) summarized by their mean accuracy and respective 2-sided 95% CI, along with the ECE and multi-class Brier score for uncertainty evaluation.

Perturbation type	**Mean accuracy (95% CI)**	**ECE**	**Brier score**
GN ($\sigma = 0.05$)*	0.713 ± 0.13	0.10	0.29
GN ($\sigma = 0.1$)**	0.697 ± 0.21	0.16	0.36
BC ($b = 0.1$)*	0.726 ± 0.07	0.09	0.31
BC ($b = 0.2$)**	0.711 ± 0.15	0.14	0.45
Both combined*	0.691 ± 0.29	0.18	0.48
Both combined**	0.675 ± 0.35	0.21	0.57

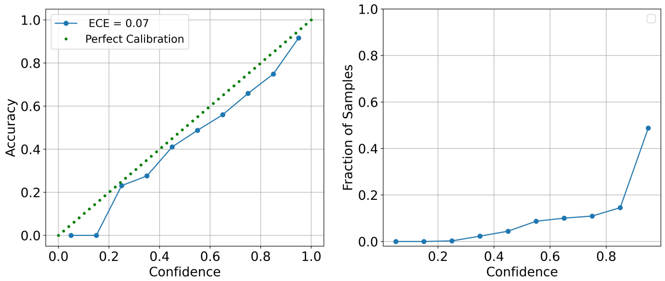
predictions of the AI system consist mostly of predictions with a confidence level above 0.7.

Beyond the graphical evaluation, the uncertainty metrics ECE and multi-class Brier score, introduced in Section V are evaluated on the perturbed test set data for the different perturbation levels used within the CI tests (cf. Table III), as required in the last objective of the certification scheme (R3). These results indicate that the CI system under test has increasing model uncertainty (increasing ECE and Brier scores) when processing perturbed data.

VII. Conclusions

Certifying the trustworthiness of AI systems is crucial to ensure their safety and for enabling compliance with regulatory requirements. AI-related ethical guidelines and frameworks for regulation provide high-level objectives for various dimensions such as transparency, reliability, safety and others. However, they often lack detailed information on the metrics, criteria, processes, and methods that should be used to assess compliance.

We present here our ongoing work on the development of a certification scheme, specifically tailored to the proposed requirements of the imminent EU AI regulation and based on international standards and guidance documents. The certification scheme extends previous work by explicitly bridging the high-level regulatory requirements to state-of-the-art technical methods for reliably assessing AI trustworthiness.



(a) Reliability diagram

(b) Confidence diagram

Fig. 5: The reliability diagram (a) visualizes the model calibration capability, while the confidence diagram (b) depicts the fraction of samples for the confidence levels.

This paper demonstrates the certification scheme using an example case for a small set of objectives for transparency and reliability. The entire certification scheme thoroughly covers the full set of objectives and trustworthiness dimensions. In the next stages of our research, we will extend the scheme to four dimensions of trustworthiness and validate it on more real-world use cases in the limited and high-risk categories. Yet, several challenges are still ahead: as research in the topic continues, emergent technical methods need to be tested for applicability in real-world scenarios. We advice researchers the use of datasets reflecting real-world conditions and integration of trustworthiness objectives in the validation of these methods. Moreover, the scarcity of appropriate metrics in these cases is concerning and has to be addressed.

As organizations will be required to comply with the coming AI-related regulations, the certification scheme can support developers, compliance officers, and certification bodies in appropriately evaluating AI-based systems, taking informed decisions, and enabling certification of innovative AI-based products.

Acknowledgment

This work was co-financed by Innosuisse (101.650 IP-ICT).

References

[1] A. Jobin, M. Ienca, and E. Vayena, “The Global Landscape of AI Ethics Guidelines,” *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019. [Online]. Available: https://doi.org/10.1038/s42256-019-0088-2

[2] Council of European Union. (2021) Council Regulation (EU) No Com/2021/206 Final. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[3] ——. (2023) Artificial Intelligence Act: Council and Parliament Strike a Deal on the First Rules for AI in the World. [Online]. Available: https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/

[4] The White House. (2023) Fact Sheet: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. [Online]. Available: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

[5] ISO. (2023) ISO/IEC JTC 1/SC 42 Artificial Intelligence. [Online]. Available: https://www.iso.org/committee/6794475.html

[6] IEEE Standards Association. (2023) IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. [Online]. Available: https://standards.ieee.org/industry-connections/ec/autonomous-systems/

[7] NIST. (2023) NIST Technical AI Standards. [Online]. Available: https://www.nist.gov/artificial-intelligence/technical-ai-standards

[8] DIN, DKE. (2023) Artificial Intelligence Standardization Roadmap. [Online]. Available: https://www.dke.de/en/areas-of-work/core-safety/standardization-roadmap-ai

[9] M. Poretschkin *et al.*, "Leitfaden Zur Gestaltung Vertrauenswürdiger Künstlicher Intelligenz," Fraunhofer IAIS, Tech. Rep., 2021. [Online]. Available: https://www.iais.fraunhofer.de/de/forschung/kuenstliche-intelligenz/ki-pruefkatalog.html

[10] LNE. (2023) Certification of Processes for AI. [Online]. Available: https://www.lne.fr/en/service/certification/certification-processes-ai

[11] IEEE. (2022) IEEE CertifAIEd. [Online]. Available: https://engagestandards.ieee.org/ieeecertifaied.html

[12] G. Soudain, "First Usable Guidance for Level 1 Machine Learning Applications: A Deliverable of the EASA AI Roadmap," 2021. [Online]. Available: https://www.easa.europa.eu/en/downloads/134357/en

[13] Trusted-AI LF AI Foundation. AI Fairness 360 (AIF360). [Online]. Available: https://github.com/Trusted-AI/AIF360

[14] ——. Adversarial Robustness Toolbox (ART). [Online]. Available: https://github.com/Trusted-AI/adversarial-robustness-toolbox

[15] ——. AI Uncertainty Quantification 360 (UQ360). [Online]. Available: https://github.com/Trusted-AI/UQ360

[16] ——. AI Explainability 360 (AIX360). [Online]. Available: https://github.com/Trusted-AI/AIX360

[17] N. Kokhlikyan, V. Miglani, M. Martin *et al.*, "Captum: A Unified and Generic Model Interpretability Library for Pytorch."

[18] Seldon. Alibi Explain. [Online]. Available: https://github.com/SeldonIO/alibi

[19] Microsoft. InterpretML. [Online]. Available: https://github.com/interpretml/interpret

[20] PAIR-code. What-If Tool. [Online]. Available: https://github.com/pair-code/what-if-tool

[21] W. D. Nothwang, M. J. McCourt, R. M. Robinson *et al.*, "The Human Should Be Part of the Control Loop?" in *2016 Resilience Week (RWS)*. IEEE, 2016. [Online]. Available: https://doi.org/10.1109/RWEEK.2016.7573336

[22] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable Machine Learning – a Brief History, State-Of-The-Art and Challenges," *Communications in Computer and Information Science*, pp. 417–431, 2020. [Online]. Available: https://dx.doi.org/10.1007/978-3-030-65965-3_28

[23] W. Zhao, S. Alwidian, and Q. H. Mahmoud, "Adversarial Training Methods for Deep Learning: A Systematic Review," *Algorithms*, vol. 15, no. 8, p. 283, 2022. [Online]. Available: https://doi.org/10.3390/a15080283

[24] V. Arya, R. K. E. Bellamy, P.-Y. Chen *et al.*, "One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques," 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1909.03012

[25] K. S. Gurumoorthy, A. Dhurandhar, G. Cecchi, and C. Aggarwal, "Efficient Data Representation by Selecting Prototypes with Importance Weights," 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1707.01212

[26] A. Kumar, P. Sattigeri, and A. Balakrishnan, "Variational Inference of Disentangled Latent Concepts from Unlabeled Observations," 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1711.00848

[27] B. M. Greenwell, B. C. Boehmke, and A. J. McCarthy, "A Simple and Effective Model-Based Variable Importance Measure," 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1805.04755

[28] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," 2013. [Online]. Available: https://doi.org/10.48550/arXiv.1312.6034

[29] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio *et al.*, Eds. Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[30] A. Ali, T. Schnake, O. Eberle *et al.*, "XAI for Transformers: Better Explanations through Conservative Propagation," 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2202.07304

[31] D. K. I. Weidele, H. Strobelt, and M. Martino, "Deepling: A Visual Interpretability System for Convolutional Neural Networks," 2019. [Online]. Available: https://mlsys.org/Conferences/2019/doc/2019/demo_22.pdf

[32] M. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. Association for Computational Linguistics, 2016. [Online]. Available: https://arxiv.org/abs/1602.04938

[33] A.-p. Nguyen and M. R. Martínez, "On Quantitative Aspects of Model Interpretability," 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2007.07584

[34] F. Bodria, F. Giannotti, R. Guidotti *et al.*, "Benchmarking and Survey of Explanation Methods for Black Box Models," *Data Min. Knowl. Discov.*, vol. 37, no. 5, pp. 1719–1778, Jun. 2023. [Online]. Available: https://doi.org/10.1007/s10618-023-00933-9

[35] N. Akhtar, "A Survey of Explainable AI in Deep Visual Modeling: Methods and Metrics," 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2301.13445

[36] SAE. (2021) Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. SAE J3016. [Online]. Available: https://www.sae.org/standards/content/j3016_202104/

[37] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations."

[38] D. R. Cox and D. V. Hinkley, *Theoretical statistics*. CRC Press, 1979.

[39] T. Gehr, M. Mirman, D. Drachsler-Cohen *et al.*, "AI2: Safety and Robustness Certification of Neural Networks with Abstract Interpretation," in *2018 IEEE Symposium on Security and Privacy (SP)*, 2018, pp. 3–18. [Online]. Available: https://doi.org/10.1109/sp.2018.00058

[40] G. Singh, T. Gehr, M. Püschel, and M. Vechev, "An Abstract Domain for Certifying Neural Networks," *Proceedings of the ACM on Programming Languages*, vol. 3, no. POPL, pp. 1–30, 2019.

[41] M. H. DeGroot and S. E. Fienberg, "The Comparison and Evaluation of Forecasters," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 32, no. 1/2, p. 12, 1983.

[42] A. Niculescu-Mizil and R. Caruana, "Predicting Good Probabilities with Supervised Learning," in *Proceedings of the 22nd international conference on Machine learning*, ser. ICML '05. ACM Press, 2005, pp. 625–632.

[43] M. Pakdaman Naeini, G. Cooper, and M. Hauskrecht, "Obtaining Well Calibrated Probabilities Using Bayesian Binning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015.

[44] G. W. Brier, "Verification of Forecasts expressed in Terms of Probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.

[45] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1050–1059. [Online]. Available: https://proceedings.mlr.press/v48/gal16.html

[46] S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft, "Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-sensitive Learning," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 1184–1193. [Online]. Available: https://proceedings.mlr.press/v80/depeweg18a.html

[47] C. Huyen, *Designing Machine Learning Systems*. USA: O'Reilly Media, 2022.

[48] P. Tschandl, N. Codella, B. N. Akay *et al.*, "Comparison of the Accuracy of Human Readers Versus Machine-Learning Algorithms for Pigmented Skin Lesion Classification: An Open, Web-Based, International, Diagnostic Study," vol. 20, no. 7, pp. 938–947.

[49] M. Tan and Q. Le, "EfficientnetV2: Smaller Models and Faster Training," in *International conference on machine learning*. PMLR, 2021, pp. 10 096–10 106. [Online]. Available: https://proceedings.mlr.press/v139/tan21a.html

[50] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2921–2929.