Advances in Reliability, Safety and Security

ESREL 2024
Monograph Book Series

Certification Scheme For Artificial Intelligence Based Systems

Yann Billeter[a], Stefan Brunner[a], Ricardo Chavarriaga[a], Philipp Denzel[a], Oliver Forster[a], Carmen Mei-Ling Frischknecht-Gruber[a], Monika Reif[a], Frank-Peter Schilling[a], Joanna Weng[a], Arman Iranfar[b], Marco Repetto[b]

[a]*Zurich University of Applied Sciences ZHAW, Winterthur, Switzerland*
[b] *CertX AG, Fribourg, Switzerland*

Organiser use only: Received date; revised date; accepted date

Abstract

Artificial Intelligence (AI) and machine learning (ML) algorithms are making an impact in an increasing number of industries. AI models differ from conventional software due to their probabilistic decision-making process, with a heavy reliance on training data quantity and quality. Ensuring the trustworthiness of AI-based systems (AIS), including dimensions such as reliability and transparency, is becoming increasingly vital due to their widespread adoption. As regulatory standards are put in place, it becomes essential to have practical guidelines for certification. In this paper, we present an ongoing effort to develop a validated Certification Scheme for AIS. This scheme encompasses distinct objectives, criteria, and corresponding measures, as well as specific metrics and technical methods which support the implementation of trustworthy AI. A critical aspect of this scheme is the establishment of a clear connection between the set of requirements and the validated ML algorithms and methods used to evaluate the compliance of AIS. We provide a tangible example of the workflow for the reliability dimension on a hypothetical real-life use case: employing the Yolo5 model for the detection of construction vehicles in a diverse image dataset of construction sites. This example demonstrates the step-by-step process of the Certification Scheme from establishing initial requirements to selecting and applying technical methods for two example objectives.

Keywords: Artificial intelligence, machine learning, certification, trustworthiness, object detection

1. Introduction

Legislators and authorities are working to establish a high level of trust in Artificial Intelligence (AI). On Friday, December 8, 2023 – after months of intensive negotiations – the European Parliament and Council reached political agreement on the European Union's Artificial Intelligence Act "EU AI Act" (European Commission, 2021). Nevertheless, even with these forthcoming regulations in the next 2-3 years, there remains a notable absence of practical guidelines and translation into specific methods and practices for effectively evaluating the trustworthiness of an AI-based systems (AIS). The deployment of AI technologies that are not fully understood and reliable can cause severe harm to society, for example by excluding minorities due to biases, causing direct physical injuries, or through misdiagnosis of patients in the health sector. The goal of the ongoing project CertAInty, which is carried out by a consortium formed by Zurich University of Applied Sciences ZHAW (Winterthur, CH) and CertX AG (Fribourg, CH), is to develop a Certification Scheme as practical guide for both developers and regulatory bodies to evaluate and certify the trustworthiness of AIS.

A Certification Scheme is a framework for the certification of AIS, including specific objectives, criteria, and corresponding means of compliance. It covers the entire life cycle of an AIS encompassing data acquisition, model development and testing, deployment, and operation. Furthermore, it is designed to be inclusive, considering the perspectives of all stakeholders such as developers, end-users, auditors, and regulatory authorities.

The novelty of this certification scheme is to bridge the gap between regulations and technical standards on the one hand and concrete scientific and technical methods to verify properties of machine learning models on the other. It will cover the following certification dimensions: reliability, transparency, autonomy and control and safety. In this paper, initial results for the reliability dimension are presented.

2. Background

Numerous national and international organizations are actively engaged in initiatives aimed at fostering trust in AI. The LNE has established an AI certification program that sets impartial and objective criteria for trustworthy AI systems. This program covers essential aspects like ethics, safety, transparency, and privacy (LNE, 2023). Several organisations and initiatives such as ISO/IEC (ISO, 2023) and NIST (NIST, 2023) are currently working on developing relevant AI standards. In addition, IEEE is in the process of creating a certification program focused on evaluating transparency, accountability, bias, and privacy in AIS (IEEE, 2022).

Additionally, EASA has released a detailed guideline for the safe application of machine learning in aviation (EASA, 2023). This guide assists aviation industry players in the development and implementation of ML systems, especially those with low levels of automation. It covers the full lifecycle, from development to operational use. Moreover, DIN/DKE offers comprehensive recommendations for standardizing AI, aiming to establish a common language and principles for development, use, and certification (DIN, DKE, 2023).

The Fraunhofer Institute has contributed to increasing trust in AI by developing guidelines for designing trustworthy AIS (Poretschkin et al., 2021). Their AI catalogue assesses AIS trustworthiness across six dimensions, including fairness, autonomy and control, transparency, reliability, safety/security, and privacy.

Method toolboxes and evaluation frameworks are vital for ensuring transparent, explainable, and robust AI systems and a variety is already available. Companies like IBM and Seldon have developed toolboxes, such as AIX360 (Bellamy et al., 2018) and Alibi (Klaise et al., 2021), featuring methods for explainability and uncertainty quantification. The Adversarial Robustness Toolbox (ART) is another framework focused on evaluating the adversarial robustness of neural networks, incorporating different types of attacks (Nicolae et al., 2018).

3. Certification Scheme

The proposed Certification Scheme for AIS will cover the following four trustworthiness dimensions: **(1) Reliability:** Assesses the AIS ability to perform consistently under varying conditions, and be robust against errors, biases, or potential security threats; **(2) Transparency:** Essential for allowing different stakeholders to comprehend the AI system's decision-making processes, thereby facilitating informed trust and reliance on the system; **(3) Safety:** Particularly crucial in safety critical domains like healthcare or autonomous vehicles, to prevent any adverse outcomes or unintended consequences of AIS's operations; **(4) Autonomy and Control:** Involves understanding the level of the AIS's independence, ranging from systems requiring human oversight ('Human-in/on-the-Loop') to those operating autonomously.

The Certification Scheme uses a risk-based approach and consists of two parts: (a) the framework itself that summarizes objectives, criteria, and various means of compliance (MOC) needed to assess the trustworthiness of an AIS; and (b) guidance linking these requirements to a set of technical and scientific methods for assessing relevant characteristics of AIS. It provides a complete workflow to identify and apply methods and processes for assessing compliance with the emerging AI regulations.

To define the Certification Scheme, an iterative approach is employed. Initially, a draft Certification Scheme is created, outlining the main objectives for achieving conformity with EU legislation. Then, the associated means of compliance are defined, distinguishing between metrics, processes, documentation, and methods. The next step is to identify, test and refine these technical and algorithmic methods in order to achieve sufficient compliance with the certification objectives. The choice of appropriate methods can vary based on several factors, including the data type, model type, the life cycle phase, and the stakeholders involved. The iterative approach ensures the certification scheme is agile, reliable, and effective. In the first version of the Certification Scheme, we included two dimensions, namely Transparency (encompassing 29 objectives and 100 MOC) and Reliability (44 objectives and 156 MOC). In the following, we focus on the reliability dimension and show the path from the objectives to the technical methods for two example objectives and MOCs.

Objective 1:
O1: The applicant should define performance metrics to evaluate the AIS performance and reliability for the regular case.
MOC: The applicant should define a suitable set of performance metrics for each high-level task to evaluate the AIS performance and reliability.
MOC: The applicant should define the expected performance with training, validation, and test data sets.
MOC: The applicant should provide a comprehensive justification for the selection of metrics.

Objective 2:
O2: The coverage of the application boundary must be formalized and quantified, if possible, and application-specific target ranges for the coverage of the application boundary must be defined.
MOC: The applicant should formalize and quantify the coverage of the Operational Design Domain (ODD, see below), especially when data point perturbations are influenced or described by factors not fully addressed for the ODD.
MOC: The application boundary may be graded differently depending on the nature and severity of the perturbation.

4. Reliability assessment of AI systems

Reliability in AI systems, as discussed in this paper, is the capacity of an AIS to consistently execute its intended functions. Robustness, on the other hand, is about maintaining performance and functionality in the face of disturbances. The requirements discussed in the following refer to different AI development lifecycle phases such as requirements definition, data acquisition, model training and verification, inference model verification and integrated AIS verification.

Operational Design Domain (ODD): While the term ODD is deeply rooted in autonomous vehicle technology (SAE International, 2021) its utility may be extended to a broader range of AIS applications. Essentially, an ODD defines the specific conditions under which a given AIS should operate safely and effectively. This concept is essential to ensure that the AIS operates within its designed capabilities and environmental constraints.

To clarify this concept in the context of autonomous driving, the ODD includes several operational parameters. These may include the types of roads the vehicle can navigate, the geographic areas in which it can operate, the static and dynamic objects the system may encounter, the weather conditions it can handle, and even the times of day or night during which it will function. For example, an autonomous vehicle's ODD may specify that it can only operate on clear days, on highways, and not on one-way streets.

The level of detail within an ODD varies depending on the intended audience. For developers and engineers, an ODD may need to include highly specific technical details. Thus, the ODD is the basis for defining the input space for an AIS, which can then be categorized into regular, robustness, and out-of-domain (OOD) cases. In the regular case, the system should be able to reliably handle small disturbances, while in the robustness case it should be able to robustly cope with large disturbances. However, in OOD cases, where data falls outside the application domain, the system may not perform adequately, leading to potential errors.

4.1. Reliability key areas

In the reliability dimension of the certification scheme for AIS, which is firmly grounded in the definition of the ODD, we consequently distinguish between 4 different key areas, as detailed in Table 1.

In the first area, the **regular case**, which focuses on the standard application domain, several essential actions are required to ensure reliable system performance. A thorough data coverage assessment is vital to verify that the data set accurately represents the application domain. Data augmentation is often necessary to increase the diversity of the dataset and prepare the system for a wider range of situations during training or testing. The adaptability and accuracy of the system must be evaluated using performance metrics with acceptable target values. At the same time, the application of loss metrics during the training phase is essential to improve model accuracy and minimize error rates. Conditioning the system to effectively handle common and environmental perturbations is another critical step to improve operational resilience. Finally, it's important to develop strategies that ensure reliable generalization on unseen data within this standard application domain and to implement mitigation measures for misjudgement.

The **robustness case** extends these principles by maintaining data coverage evaluation and augmentation while emphasizing robustness, especially under challenging conditions such as edge and corner cases. Performance metrics are evaluated under these stringent conditions, and a systematic vulnerability assessment is performed to identify and address potential weaknesses. The robustness case also addresses common perturbations and extreme environmental scenarios, incorporating strategies to mitigate adversarial attacks and exploring generalization beyond typical operations.

The **OOD case** further expands the scope to include scenarios outside of the regular and robust cases. Data augmentation is adapted for OOD data, and special attention is given to catching errors at the input. Vulnerability assessment remains vital, along with protection against adversarial attacks and exploration of generalization outside the expected operational domain.

Finally, the **Uncertainty Estimation** aspect involves a detailed uncertainty assessment, where appropriate uncertainty metrics need to be defined and applied. This stage focuses on quantifying both intrinsic and extrinsic uncertainties and developing mitigation measures to address uncertainties in decision processes.

Table 1. Reliability key areas with their main actions

Regular Case	Robustness Case	OOD Case	Uncertainty Estimation
Data coverage assessment	Data coverage assessment	Data coverage assessment	Uncertainty assessment
Data augmentation	Data augmentation	Data augmentation	Uncertainty metrics
Performance metrics	Performance metrics	Catching errors at input	Uncertainty estimation
Loss metrics	Vulnerability assessment	Vulnerability assessment	
Common perturbations	Common perturbations	Common perturbations	
Environmental perturbations	Environmental perturbations	Environmental perturbations	
	Adversarial attacks	Adversarial attacks	
Generalization within regular case	Generalization and exploration within robustness case	Generalization and exploration outside expected operation	
Mitigation measures for misjudgement	Mitigation measures for misjudgement	Mitigation measures for detected errors	Mitigation measures for uncertainties

In addition, process steps such as evaluating the model architecture, implementing optimization techniques such as pruning or quantization which are often required in order to deploy the AI model on an edge device, and ensuring reproducibility are important, as are conducting regular assessments and meticulously documenting all activities.

4.2. Metrics and technical methods for reliability assessment

We identified more than 55 metrics and 95 methods relevant for the certification of the reliability dimension based on a comprehensive review of the state of the art. After further analysis, we selected a subset of 35 metrics and 50 methods for empirical tests using as criteria: (i) their relevance to the intended applications; (ii) their applicability for certification regarding execution time, reliance on available information and computational costs.

The selection of the appropriate **evaluation metrics** is essential for an AIS and varies significantly across application domains and the objectives of the model. For example, in regression tasks, metrics such as (Mean) Squared Error (MSE) and (Mean) Absolute Error (MAE) are often used to evaluate prediction accuracy. In classification scenarios, a variety of metrics such as (Mean) Accuracy, Precision, Recall, F1 Score, and Area Under the Curve (AUC) value are used to measure how effectively the model categorizes the data. Ranking algorithms typically use Mean Reciprocal Rank (MRR) and Discounted Cumulative Gain (DCG) to determine the relevance of ranked outputs, while clustering models use measures such as Silhouette Value and Adjusted Mutual Information (AMI) Score to evaluate data groupings. In specialized areas such as computer vision, metrics such as Mean Intersection over Union (mIoU) and Mean Average Precision (mAP) are used, while Natural Language Processing (NLP) often relies on Perplexity Score and Bilingual Evaluation Understudy (BLEU) Score for language models and translation quality.

To assess and improve the reliability of AIS, **technical methods** are provided that are tailored to specific challenges such as common perturbations, environmental factors, adversarial inputs, and uncertainty. For common perturbations, techniques such as injecting Gaussian noise into data sets or applying geometric transformations (e.g., rotations, scaling) test the stability of the system against routine data variations. Environmental perturbations are addressed through methods such as varying lighting conditions in image processing tasks or simulating different climatic conditions through scenario-based testing and real-world data augmentation, enriching the training dataset with diverse environmental examples to increase adaptability and resilience. To counter adversarial threats, approaches such as the Fast Gradient Sign Method (FGSM) or Generative Adversarial Networks (GANs) are used to create challenging adversarial examples that enhance the system's ability to withstand malicious input. Uncertainty assessment incorporates Bayesian methods for probabilistic modelling and Monte Carlo simulations to evaluate the model's response to a wide range of inputs, thereby assessing predictive uncertainty. Coverage assessment is addressed through techniques such as formal verification, which uses logical and mathematical proofs to confirm that the system meets specified criteria, and model coverage analysis, which assesses whether all aspects of the system have been thoroughly tested against a comprehensive range of scenarios, including edge cases.

It is important to note that the above mentioned are examples, and a wider range of other metrics and methods may be more appropriate depending on the context and nuances of the specific problem. Therefore, a thorough evaluation is required to select the most appropriate metric for the certification, considering the objectives of the model, the nature of the data, and the desired outcomes.

5. Real-life use case: Vehicle detection on construction sites

In the following, we consider a hypothetical real-life use case in which a computer vision model in an AIS is required to detect specific kinds of vehicles and equipment on construction sites, as depicted in specific image datasets from the Roboflow 100 (RF100) benchmark[1]. RF100 is a crowdsourced, open-source benchmark which contains over 90’000 image datasets spanning a wide range of domains. This initiative, sponsored by Intel, is designed to provide a comprehensive and diverse benchmark for machine learning object detection models, allowing them to be tested across a wide range of real-world scenarios and increase their generalizability. We selected three of them, suited for the training of model: the *Excavators*[2], *Heavy Equipment*[3], and *Construct.AI.v2.BB*[4] datasets. The compiled dataset contains 12963 images, featuring 7 classes of vehicles, and over 15’000 class instances (see Table 2 for the full list). The train/validation/test split is 85/10/5%, respectively, preserving the class distributions in each partition.

Table 2. The dataset’s class instance distribution in numbers and percentages of the total. Note that an image may feature multiple instances.

Class	Instances	Percentage (%)
dump truck	3470	23
loader	2799	19
excavator	3793	25
roller	2843	19
mobile crane	456	3
bulldozer	918	6
grader	787	5

The YOLOv5 architecture was chosen for the object detection component of the AIS, as it presents a solid choice for a wide variety of object detection tasks with (nearly) state-of-the-art performance. It was designed by Ultralytics to be fast, accurate, and easy to use, with flexible open-source licensing. It is the fifth iteration of the original YOLO network (You Only Look Once; Redmon et al. 2015, 2016, 2018), which, as its name suggests, connects the procedure of predicting class labels (classification) with bounding boxes (detection) in an end-to-end differentiable network. Thus, it approaches the object detection task as a regression problem by spatially separating bounding boxes and associating scores to each of them. This allows the model to detect multiple instances in a single image, and in fact often predicts multiple box candidates for a single instance in which case a thresholding algorithm filters out the relevant predictions.

6. Technical assessment of the use case

The following section addresses objectives **O1 (Definition of performance metrics and evaluation of the AIS’s performance and reliability)** and **O2 (Formalization and quantification of application coverage)** from the certification scheme in section 3, specifically in the context of Operational Design Domain (ODD) coverage. This is demonstrated based on the use case described in the previous section.

[1] https://www.rf100.org/
[2] https://universe.roboflow.com/mohamed-sabek-6zmr6/excavators-cwlh0
[3] https://universe.roboflow.com/kfu-ye4kz/heavy_equipment-ifaqm
[4] https://universe.roboflow.com/andrew-hannell/construct.ai.v2.bb

6.1. Objective 1: Definition of metrics and performance of the model

During the operation of the AIS, the object detection model may be required to identify a specific class, such as excavators or loaders, with a given minimum performance score. For this task there are several suitable metrics to evaluate the model's performance.

Intersection over Union (IoU) is a metric used in object detection to quantify the accuracy of a predicted bounding box compared to the ground truth (actual) bounding box. IoU is calculated by dividing the area of overlap between the predicted and ground truth bounding boxes by the area of the union of these two boxes. This ratio ranges from 0 to 1, where 1 indicates perfect overlap and 0 means no overlap. IoU is widely used in evaluating the performance of object detection models like YOLO.

Another metric widely established for object detection models is the **average precision (AP)** which generally measures the area under the precision-recall curve. Following the COCO (T.Y. Lin, 2014) standard, average precisions are evaluated for each class separately for a 101-point interpolated curve with an IOU threshold between 0.5 and 0.95 (in steps of 0.05) and afterwards averaged over all classes, yielding the mean average precision (mAP@50-95); analogously for mAP@50 at an IOU threshold of 0.5 corresponding to the VOC standard[5]. Both metrics incorporate the trade-off between precision and recall and take false positives and false negatives into account which makes them especially suited for object detection applications.

Based on the discussion above, we used precision, recall, mAP@50, and mAP@50-95 as metrics for evaluating the model performance on the test dataset. As the main objective of this work is the assessment of an AIS rather than the development of a highly tuned, optimized model, the performance goal was arbitrarily set to >90% precision on class average. Consequently, the model training was run using mostly default hyperparameter settings for almost 100 epochs until the mean precision for the validation dataset converged above this threshold. The final test set performance per class label is listed in Table 3.

Table 3. Test set performance of the model evaluated using the chosen metrics.

Class Metric	dump truck	loader	excavator	roller	mobile crane	bulldozer	grader
mAP@50	0.53	0.80	0.75	0.96	0.38	0.76	0.78
mAP@50-95	0.77	0.93	0.93	0.93	0.62	0.96	0.97
Recall	0.62	0.85	0.85	0.95	0.59	0.97	0.96
Precision	0.85	0.95	0.94	0.85	0.75	0.89	0.87

6.2. Objective 2: ODD Assessment

The reliability assessment focused on a subset of methods to cover both "regular case" and "robustness case" scenarios based on the defined ODD and expected hardware effects, covering normal operating conditions as well as less common and more challenging, but still expected conditions. The simulated perturbations are described in the following and the range of parameters for the different perturbations is summarized in Table 4.

6.2.1 ODD Simulation

Common perturbations: In order to evaluate the robustness of the ML model, several common perturbations are applied to the images of the dataset, in accordance with the objective O2. These perturbations, which address real-world challenges such as camera noise, motion effects, or other visual distortions, are simulated by various linear transformations and noise, and are designed to cover the scenarios within the ODD that the algorithms may encounter. The following common perturbations were applied to the data set:

- **Homogeneous noising:** involves adding a consistent layer of noise across the entire image, affecting its clarity.
- **Gaussian noise,** a statistical noise having a probability density function equal to that of the normal distribution, is often added to images to mimic random variations in pixel values.
- **Brightening** is another perturbation where the image's luminance is uniformly increased, which can lead to loss of detail in brighter areas.

[5] Note that in the literature, mAP and AP are often used interchangeably, since the context usually implies the appropriate interpretation.

Table 4. Parameters of common and environmental perturbations with their relation to the ODD

Perturbation type	Name	Parameter	ODD reference
Common	Homogenous noise (ISO, 2021)	$k = [-0.2, 0.2]$	Camera type
Common	Gaussian noise (Cattin, P., 2016)	$\mu = 0.0; \sigma = [0.0, 0.2]$	Camera type
Common/ Environmental	Brightening (ISO, 2021)	$b = [-0.2, 0.2]$	Illumination (day, artificial)
Common/ Environmental	Contrast (Peli, E. 1990)	$c = [0.8, 1.2]$	Illumination (day, artificial)
Common	Rotation (ISO, 2021)	$\text{angle}[°] = [-10, 10]$	Camera type
Common	Blurring (ISO, 2021)	$\sigma = [0.8, 1.5]$	Camera type
Common	Motion blur (vertical, horizontal) (Navarro, F., 2011)	$\text{Kernel size} = [5, 20]$	Camera type
Common/ Environmental	Blooming (ISO, 2021)	$\text{Value threshold} = [200, 240]$	Illumination (day, artificial).
Environmental	Snow (Von Bernuth, A., 2019)	Light to heavy (5 levels)	Snowfall
Environmental	Fog (Von Bernuth, A., 2019)	Light to heavy (5 levels)	Non-precipitating smaller water droplets
Environmental	Sand (Si, Y., 2022)	Light to heavy (5 levels)	Larger airborne particles
Environmental	Dust (Si, Y., 2022)	Light to heavy (5 levels)	Smaller airborne particles
Environmental	Rain (Tremblay, M., 2021)	Light to heavy (5 levels)	Rainfall

- **Contrast** specifically targets the range and intensity of the light-to-dark spectrum in an image. By increasing or decreasing contrast, it can either exaggerate or diminish the distinctiveness of features.
- **Rotation** involves turning the image around a central point, testing the algorithm's ability to recognize objects from different angles.
- **Blurring** is a common perturbation that simulates out-of-focus images, reducing their sharpness and detail.
- **Motion blur** replicates the effect of movement during image capture, creating streaks or blurs in the direction of motion.
- **Blooming** occurs when intense light sources in an image cause bleeding of light, affecting the visibility of adjacent areas.

Each perturbation introduced presents its own challenges, and evaluating its impact is critical to assessing the reliability of the ML model. This assessment is particularly important for determining the model's ability to effectively handle a range of real-world perturbations, as required by the objective O2. In Fig. 1, a selection of implemented common perturbations is shown.

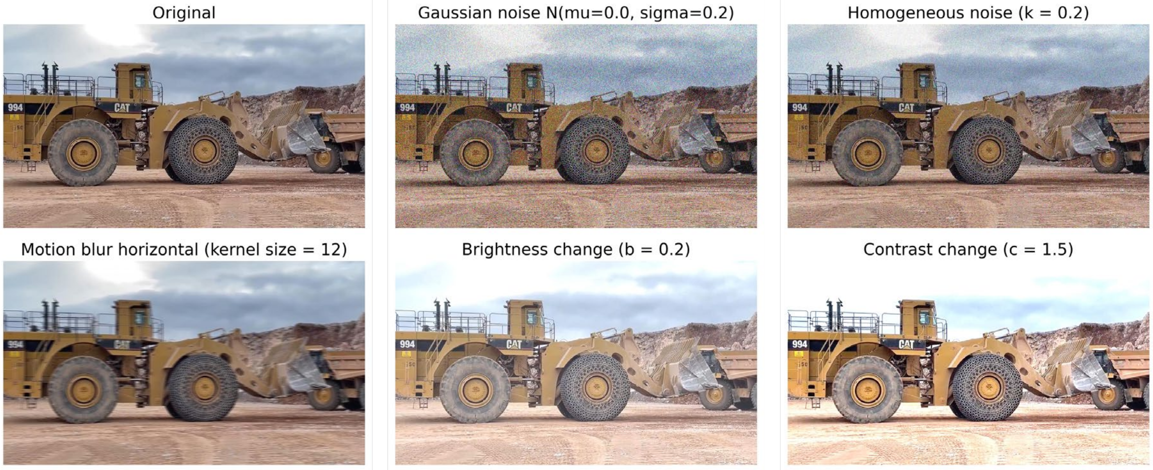
Original
Gaussian noise N(mu=0.0, sigma=0.2)
Homogeneous noise (k = 0.2)
Motion blur horizontal (kernel size = 12)
Brightness change (b = 0.2)
Contrast change (c = 1.5)

Fig. 1. Examples of common perturbations

Environmental perturbations: In order to cover the environmental aspects within the objective O2, ODD simulation with a focus on environmental perturbations is necessary to evaluate the reliability of object detection algorithms under all relevant conditions as listed in Table 4. The following environmental perturbations were applied to the data set:

- **Fog,** creating a uniform, diffuse layer of moisture that can significantly reduce visibility. It tends to scatter light in all directions, resulting in a general loss of contrast in visual scenes and a white or grayish appearance.
- **Snow:** Assessing the impact of snow is vital as it can obscure road markings, signs, and even the roadway itself. Snow can lead to reduced visibility and can also interfere with the sensors and cameras.
- **Rain** can affect visibility and also interfere with sensors. Heavy rain can cause reflections, distortions, and other optical effects that challenge the perception systems of autonomous vehicles.
- **Sand:** Airborne sand can cause more localized and irregular visibility problems. Sand particles scatter light in a less uniform manner, resulting in heterogeneous visual effects. Visibility can be severely reduced, but the effect is more one of blocking and distorting light rather than uniformly scattering it.
- **Dust:** Finer dust particles can create a widespread haze that reduces overall visibility and image sharpness by uniformly scattering light. In addition, dust can give the air a brownish or yellowish hue and cause a heterogeneous effect on visibility, with some areas appearing clearer than others.

Fig. 2 shows a selection of implemented environmental perturbations. Different methods are used to augment the existing image dataset, including physics-based effects and a hybrid approach combining physics-based and learning-based methods as described in (Tremblay, M., 2021).

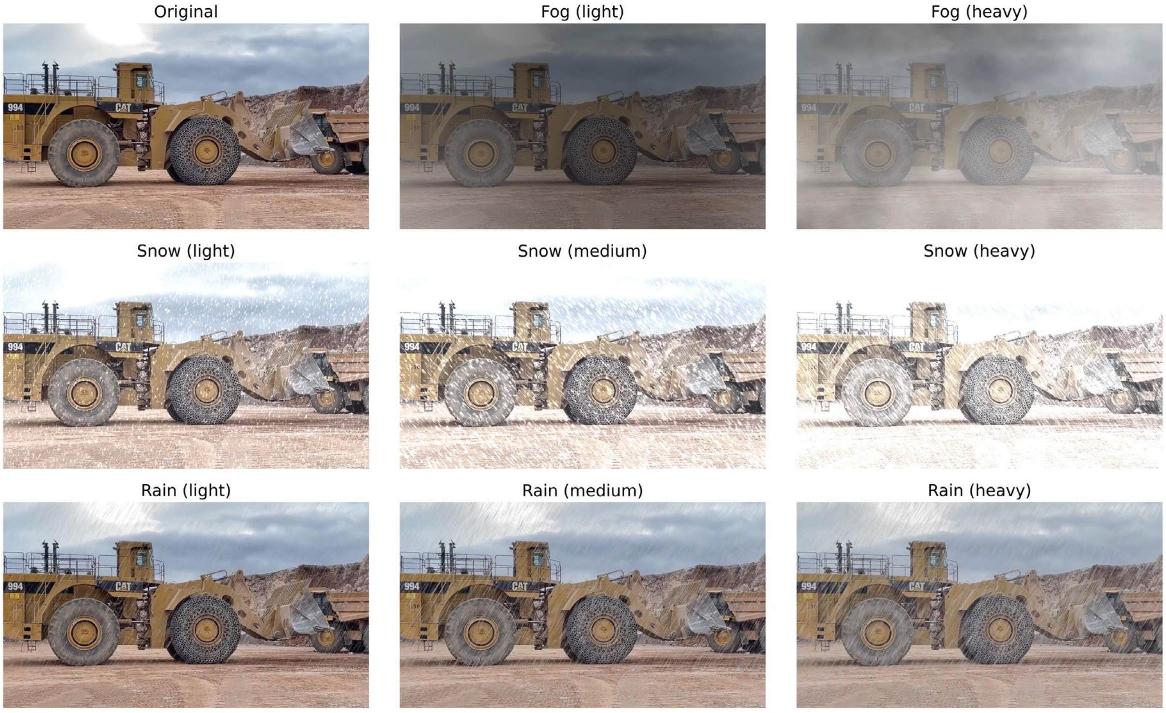

Fig. 2. Examples of environmental perturbations

Combined perturbations: The simulation of combined environmental and common perturbations offer a comprehensive evaluation of its robustness against multiple simultaneous challenges. Fig. 3 shows a selection of implemented combined perturbations. This approach provides a more accurate representation of the system's performance in diverse real-world scenarios.

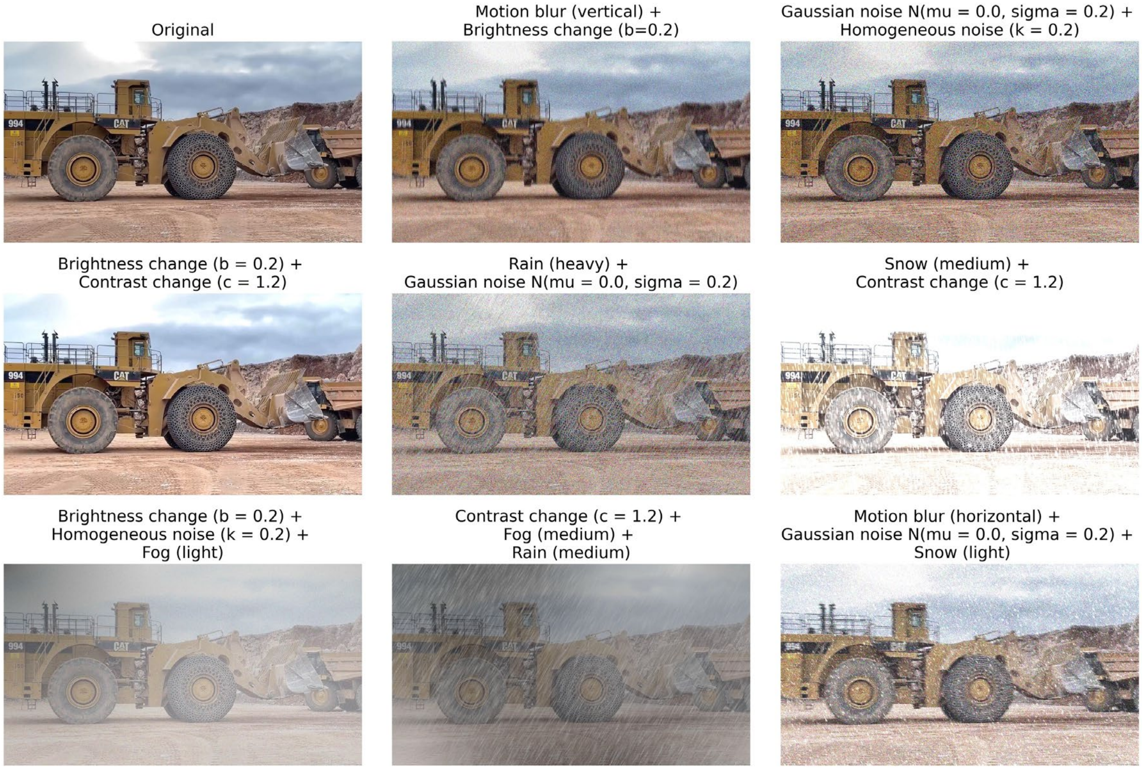
Original
Motion blur (vertical) + Brightness change (b=0.2)
Gaussian noise N(mu = 0.0, sigma = 0.2) + Homogeneous noise (k = 0.2)
Brightness change (b = 0.2) + Contrast change (c = 1.2)
Rain (heavy) + Gaussian noise N(mu = 0.0, sigma = 0.2)
Snow (medium) + Contrast change (c = 1.2)
Brightness change (b = 0.2) + Homogeneous noise (k = 0.2) + Fog (light)
Contrast change (c = 1.2) + Fog (medium) + Rain (medium)
Motion blur (horizontal) + Gaussian noise N(mu = 0.0, sigma = 0.2) + Snow (light)

Fig. 3. Examples of combined perturbations

6.2.2 ODD Assessment - Performance of the model

The AI model's performance was assessed using the previously defined metrics as function of the different simulated perturbations to measure how effectively the computer vision model can detect the specified vehicles and equipment under the formalized ODD conditions. Results show significant impact on the model performance, especially in the case of combined perturbations, and emphasize the need for rigorous testing using simulations to enhance the model's robustness in real-world applications. Based on the results, as a next step, specific target ranges for the chosen performance metrics need to be established. These targets set the expected performance levels for the model across diverse scenarios within the ODD, for instance different noise levels, and serve as benchmarks for future evaluations and enhancements. Fig. 4 summarizes the described certification workflow from the initial objectives to the robustness assessment. A complete evaluation of the application boundary would include a broader variety of technical methods, for example also adversarial attacks, assessments of concept and model drifts etc. not included in this analysis.

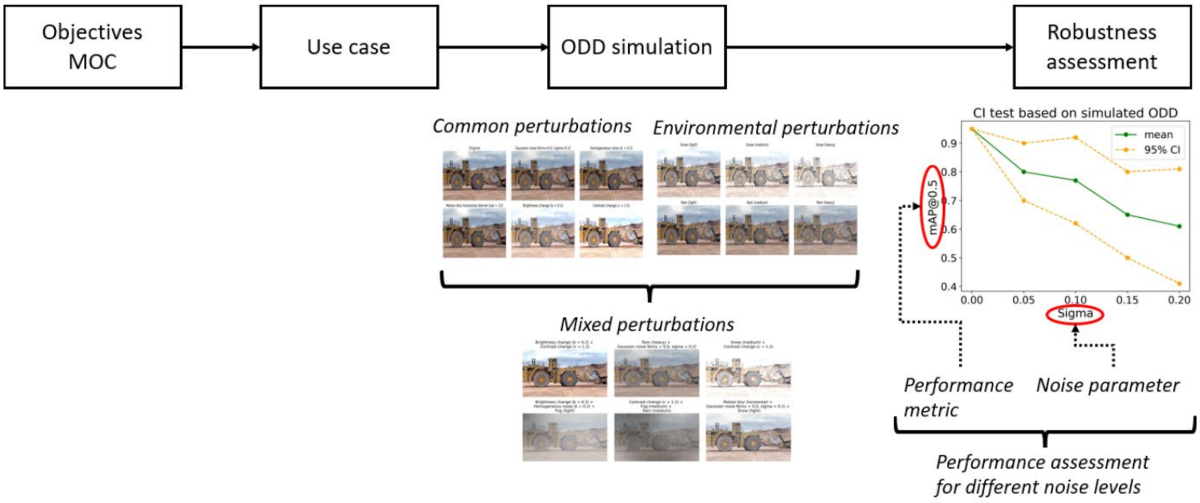

Fig. 4. Certification Workflow

7. Conclusion

Upcoming laws and regulations drive the need for established procedures to demonstrate and verify the trustworthiness of AI systems. In this paper we presented the development of an AI Certification Scheme which will cover the AI trustworthiness dimensions of reliability, transparency, autonomy and control, and safety. The scheme consists of two parts, namely the framework itself that summarizes objectives, criteria, and various means of compliance needed to assess the trustworthiness of an AIS, as well as guidance linking these requirements to a set of technical methods and metrics for assessing relevant characteristics of AIS. Objectives are defined based on the current legislation and the state of the art. Then, means of compliance to achieve these objectives are specified, distinguishing between criteria and metrics, processes, documentation, and methods to comply with the objectives. We demonstrated the application of this newly developed AIS Certification Scheme to a real-life use case for the reliability dimension, namely an AIS using a computer vision system to detect vehicles on construction sites. Starting with clear objectives, the workflow shows how technical methods and metrics are selected and applied to this use case. Performance metrics are defined and investigated as a function of various simulated perturbations, as defined in the Operational Design Domain (ODD). Besides the reliability dimension discussed in this paper, the final Certification Scheme will cover other dimensions such as transparency, autonomy and control and safety, and be applicable to a wide range of AIS, providing a guideline towards future certification of AI systems.

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

References

Bellamy, R. K. E., et al. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. https://arxiv.org/abs/1810.01943.

Cattin, P. 2016. Image Restoration: Introduction to Signal and Image Processing. MIAC, University of Basel.

DIN, DKE. 2023. German Standardization Roadmap on Artificial Intelligence (2nd edition). www.din.de/go/roadmap-ai.

EASA. 2023. EASA Concept Paper: First usable guidance for Level 1 & 2 machine learning applications, Issue 02. https://www.easa.europa.eu/en/document-library/general-publications/easa-artificial-intelligence-concept-paper-proposed-issue-2.

European Commission. 2021. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. COM/2021/206 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206.

IEEE Standards Association. 2023. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. https://standards.ieee.org/industry-connections/ec/autonomous-systems.

ISO (International Organization for Standardization). 2021. Artificial Intelligence (AI) - Assessment of the robustness of neural networks - Part 1: Overview. ISO/IEC TR 24029-1:2021. https://webstore.iec.ch/publication/68715.

ISO (International Organization for Standardization). 2023. Artificial intelligence. ISO/IEC JTC 1/SC 42. https://www.iso.org/committee/6794475.html.

Klaise, J., Van Looveren, A., Vacanti, G. and Coca, A. (2021). Alibi Explain: Algorithms for Explaining Machine Learning Models. Journal of Machine Learning Research, 22 (181), 1–7. http://jmlr.org/papers/v22/21-0017.html.

Lin, T. Y., et al. 2014. Microsoft COCO: Common Objects in Context. ECCV 2014: 13th European Conference. https://arxiv.org/abs/1405.0312.

LNE. 2023. Certification of Processes for AI. https://www.lne.fr/en/service/certification/certification-processes-ai.

Navarro, F., Serón, F. J., Gutierrez, D. 2011. Motion Blur Rendering: State of the Art. Computer Graphics Forum, 30, 3-26. https://doi.org/10.1111/j.1467-8659.2010.01840.x.

Nicolae, M.-I., et al. 2018. Adversarial Robustness Toolbox v1.0.0. https://arxiv.org/abs/1807.01069.

NIST. 2023. Technical AI Standards. https://www.nist.gov/artificial-intelligence/technical-ai-standards.

Peli, E. 1990. Contrast in complex images. Journal of the Optical Society of America A, 7 (10), 2032-2040. https://doi.org/10.1364/josaa.7.002032.

Poretschkin, M., et al. 2021. Leitfaden zur Gestaltung vertrauenswurdiger Kunstlicher Intelligenz (KI-Prufkatalog). Technical Report, Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS. https://www.iais.fraunhofer.de/de/forschung/kuenstliche-intelligenz/ki-pruefkatalog.html.

Redmon, J., et al. 2015. You Only Look Once: Unified, Real-Time Object Detection. https://arxiv.org/abs/1506.02640.

Redmon, J., et al. 2016. YOLO9000: Better, Faster, Stronger. https://arxiv.org/abs/1612.08242.

Redmon, J., et al. 2018. YOLOv3: An Incremental Improvement. https://arxiv.org/abs/1804.02767.

SAE International. 2021. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. SAE J3016_202104. https://www.sae.org/standards/content/j3016_202104/.

Si, Y., Yang, F., Guo, Y., Zhang, W., Yang, Y. 2022. A comprehensive benchmark analysis for sand dust image reconstruction. Journal of Visual Communication and Image Representation, 89, 103638. https://doi.org/10.1016/j.jvcir.2022.103638.

Tremblay, M., Halder, S. S., De Charette, R., Lalonde, J.-F. 2021. Rain rendering for evaluating and improving robustness to bad weather. International Journal of Computer Vision, 129, 341-360. https://doi.org/10.1007/s11263-020-01366-3.

Von Bernuth, A., Volk, G., Bringmann, O. 2019. Simulating photo-realistic snow and fog on existing images for enhanced CNN training and evaluation. 2019 IEEE Intelligent Transportation Systems Conference (ITSC), 41-46. https://doi.org/10.1109/ITSC.2019.8917367.