Towards Audit Requirements for AI-Based Systems in Mobility Applications

Devi Padmavathi Alagarswamy[1], Christian Berghoff[2], Vasilios Danos[3], Fabian Langer[3], Thora Markert[3,*], Georg Schneider[4], Arndt von Twickel[2] and Fabian Woitschek[4,*]

[1]*ZF Friedrichshafen AG, System House Autonomous Mobility Systems, Friedrichshafen, Germany*

[2]*German Federal Office for Information Security (BSI), Bonn, Germany*

[3]*TÜV Informationstechnik GmbH, IT Security Hardware Evaluation, Essen, Germany*

[4]*ZF Friedrichshafen AG, Artificial Intelligence Lab, Saarbrücken, Germany*

Keywords: Artificial Intelligence, Neural Networks, Security, Safety, Trustworthiness, Regulation, AD, ADAS.

Abstract: Various mobility applications like advanced driver assistance systems increasingly utilize artificial intelligence (AI) based functionalities. Typically, deep neural networks (DNNs) are used as these provide the best performance on the challenging perception, prediction or planning tasks that occur in real driving environments. However, current regulations like UNECE R 155 or ISO 26262 do not consider AI-related aspects and are only applied to traditional algorithm-based systems. The non-existence of AI-specific standards or norms prevents the practical application and can harm the trust level of users. Hence, it is important to extend existing standardization for security and safety to consider AI-specific challenges and requirements. To take a step towards a suitable regulation we propose 50 technical requirements or best practices that extend existing regulations and address the concrete needs for DNN-based systems. We show the applicability, usefulness and meaningfulness of the proposed requirements by performing an exemplary audit of a DNN-based traffic sign recognition system using three of the proposed requirements.

1 INTRODUCTION

Artificial intelligence (AI) -based systems are increasingly used as part of mobility applications like autonomous driving (AD) or advanced driver assistance systems (ADAS). Especially, deep neural networks (DNNs) achieve an impressive performance on most tasks and are the most promising solution to achieve higher levels of automated driving. At the same time, different manufacturers already use DNN-based solutions as part of ADASs with partial automation (SAE L2 (SAE J3016, 2014)) (Karpathy, 2021) that are operating on public roads or for highly automated shuttles (SAE L4) (Waymo, 2021) operating in limited public areas. However, current DNNs introduce new and specific vulnerabilities into the systems which can impact the performance and trustworthiness of AD/ADAS systems negatively. This requires a detailed analysis of existing vulnerabilities and potential mitigation strategies. To still enable the usage of such DNN-based solutions for high-risk applications, like highly or fully automated driving (SAE L4/L5), clear guidelines and regulations are required. This assures that systems with a high degree of autonomy are trustworthy with respect to use case relevant aspects like safety, security, robustness or explainability and include mitigation strategies to known vulnerabilities.

However, currently no homologation regulations or standards exist that are tailored towards the use of AI-based systems in mobility applications and include AI-specific vulnerabilities (Radlak et al., 2020). There are no uniformly acknowledged principles and practices that the development, testing or deployment of AI-based systems must fulfill. This limits the future deployment of AI-based systems to low-risk applications. Furthermore, it represents a major challenge for industry, auditors and regulators and potentially leads to a lower level of user trust.

To provide guidance for future regulations, in this work we explore how auditing guidelines can ensure the security and safety of AI-based systems in high-risk applications. Thereby, we focus on the application of AI-based systems for mobility applications

*Main Contribution.

Alagarswamy, D., Berghoff, C., Danos, V., Langer, F., Markert, T., Schneider, G., von Twickel, A. and Woitschek, F.
Towards Audit Requirements for AI-Based Systems in Mobility Applications.
DOI: 10.5220/0011619500003405
In *Proceedings of the 9th International Conference on Information Systems Security and Privacy (ICISSP 2023)*, pages 339-348
ISBN: 978-989-758-624-8; ISSN: 2184-4356

like functionalities for AD/ADAS. Hence, we base our work on existing standards for road vehicles, like the functional safety standard ISO 26262 (ISO 26262, 2018), which are relevant for mobility applications, and provide the following contributions:

- We present an overview of existing and developing standards relevant for auditing AI-based systems in mobility applications.
- We introduce a list of generic requirements which focus on specific needs arising when auditing AI-based systems.
- We compare different use cases for AI-based systems in mobility applications to select the most suitable use case which is used for testing and refining the introduced audit requirements in practice.
- We demonstrate the applicability of the most relevant audit requirements for the selected use case.

2 RELATED WORK

During an audit the compliance to industry standards mandated by regulators is evaluated. However, traditional standards for systems in mobility applications do not contain specific guidelines in case AI-based systems are utilized instead of traditional algorithm-based systems. This includes both safety standards like (ISO 26262, 2018), (ISO 21448, 2022) or (ANSI/UL 4600, 2022) and security standards like (ISO/SAE 21434, 2021) or (UNECE R 155, 2021).

Since currently no standards for the auditing of AI-based systems exist, there are approaches to develop an appropriate standardization. Best known here is the (EU AI Act, 2021) which tries to lay down uniform regulations for AI-based systems. It presents a horizontal regulatory approach with necessary requirements to address different risks and challenges when AI is used, without focusing on the needs for specific application areas. Similarly, (ISO/IEC TR 24028, 2020) focuses on the trustworthiness of AI-based systems without considering a concrete application domain. It surveys different generic threats and risks and also covers existing mitigation strategies. Additionally, (ISO/IEC TR 24029-1, 2021) provides background information on existing methods to assess the robustness of generic DNNs.

In addition to the already published drafts for standardization of AI-based systems mentioned above, there are also ongoing standardization activities. This includes (ISO/IEC DTR 5469, 2022), which covers aspects of functional safety specific for AI-based systems. In addition, (ISO/IEC PRF TS 4213, 2022) focuses on methods to assess the performance of ML-based classification systems. Furthermore, there are standardization activities on a national level which include the (DIN Roadmap AI, 2022). Here, requirements and challenges as well as standardization needs for seven topics around AI are discussed.

In contrast to these horizontal regulatory approaches, there are also vertical regulatory approaches. These aim to develop standards for concrete application areas and multiple standards are in development for the usage of AI in (high-risk) mobility applications. Here, (ISO/AWI TS 5083, 2022) gives guidance of the steps for developing and validating safe AD/ADAS systems. It covers the SAE levels L3/L4 and the impact of using AI-based systems as part of larger AD/ADAS systems. Additionally, (ISO/AWI PAS 8800, 2022) focuses specifically on the interaction between safety and AI. It defines risk factors for vulnerabilities in the behavior of AI within mobility applications.

Finally, there are only very few publications available that focus on the auditing of AI systems in practice. Here, (Raji et al., 2020) introduces a framework for auditing AI-based systems throughout the internal development lifecycle. This results in a series of documents which form an overall audit report that can be used by auditors for a formal audit.

After discussing available and developing standards for AI-based systems in general and specific to mobility applications, we now shortly present an overview of related publications which focus on specific vulnerabilities or challenges that exist for AI-based systems. In (Li et al., 2022) the authors present a summary of important aspects for trustworthy systems which they deem necessary to be audited. (Berghoff et al., 2020) focuses on discussing known vulnerabilities of current AI-based systems and how they compare with traditional algorithm-based systems. Similarly, (Mohseni et al., 2020) first discusses current vulnerabilities specific to AI-based systems and then present possible mitigation strategies.

Concretely, there are multiple new security challenges for AI-based systems. These include model extraction attacks (Papernot et al., 2017; Orekondy et al., 2019), where an adversary attempts to copy the functionality of a victim AI model. Next, evasion attacks, also known as adversarial attacks (Szegedy et al., 2014; Madry et al., 2018), are carefully perturbed input samples (adversarial examples) that change the prediction of AI-based systems according to the will of an adversary. This imposes a threat on the integrity of the system. Lastly, data poisoning attacks (Goldblum et al., 2020; Schwarzschild et al., 2021) describe the injection of poisoned data samples

in the training dataset of an AI-based system. This degrades the behavior of the resulting system depending on the specific goals of the adversary.

In addition to security related AI-specific vulnerabilities there are also challenges regarding the robustness against natural perturbations (Hendrycks and Dietterich, 2019; Geirhos et al., 2020). Concretely, out-of-domain data describes the presence of data samples that deviate from the exact training distribution used during training of an AI-based system. This effect occurs naturally when systems are deployed in the real-world outside of a completely supervised environment and presents a challenge on the generalization of such systems.

Also, the existing black-box character of AI-based systems, combined with the complexity and number of parameters of DNNs, complicates the possibility to explain the behavior of a system. It is largely unclear how a system arrives at its predictions and which features of a data sample are most important for a concrete prediction. Therefore, the need arises for methods that can explain the decision or general behavior of a system that is learned from data. These methods are published under the term of explainable artificial intelligence (Gilpin et al., 2018; Guidotti et al., 2018).

3 GENERIC AUDIT REQUIREMENTS

In the following, generic requirements are derived based on a detailed analysis of established security and safety standards. As stated in section 2 currently there are no existing AI-specific certification standards, norms or regulations for systems in mobility applications. The available standards and norms are designed for traditional algorithm-based automotive systems. We extract their AI-relevant aspects and complement them with AI-specific formulations.

3.1 Requirements Elicitation

Due to the special characteristics of AI components (e.g. high data complexity, non-linearity or lack of interpretability), some of the existing standards may not apply to such components or have to be adjusted to also ensure the safety and security of AI-based systems. Hence, the generic requirements are formulated to address the technical aspects performance, robustness, explainability, external monitoring and the documentation of the entire mobility system and its AI subsystems. Furthermore, we consider requirements along the entire lifecycle of such systems. Fairness and privacy are out of scope for this work and should be addressed in future research. In total we formulate 50 generic requirements, which are available in the project report at (AIMobilityAuditPrep, 2022). In the following, we present our general approach to derive the requirements and discuss three exemplary requirements which can be applied to most AI-based systems and are highly relevant.

The (ISO 26262, 2018) introduces the automotive safety level integrity (ASIL), which is a risk level based classification of recommendations for automotive systems. In this classification scheme the system's risk level is categorized in four risk levels, through the possible exposure to hazards, the controllability of a hazard and the severity of possible injuries to the driver and passengers stemming from the hazard. The four ASIL levels range from "ASIL A" associated with the lowest degree of risk to "ASIL D" associated with the highest degree of risk. The ISO 26262 associates safety requirements to recommendations for each risk level. These recommendations are described as "highly recommended" (++), "recommended" (+) and "not recommended" (o), where "highly recommended" indicates a need for implementation of the associated requirement for an application associated with the corresponding risk level. Since the ASIL is a well-defined classification scheme, we follow this risk level based categorization approach to classify each of the requirements according to their risk level definition. This allows for a risk based selection of a set of requirements for each individual mobility application.

During a homologation process, the integration of vehicle components is evaluated at each integration step. Therefore, the functional safety and security of the entire mobility system must be addressed during the requirements elicitation. Accordingly, we categorize the requirements whether they apply to the entire mobility system or the AI subsystem.

3.2 Entire System Requirements

To provide more insight into the requirements elicitation process, we show an example requirement catered towards the entire system and its ASIL classification. REQ. 7 from Table 1 ensures that the performance of the entire system is not affected under worst-case conditions. This can either encompass natural phenomena such as weather or lighting conditions, but also security threats for example by adversarial attacks or side-channel attacks. It is derived from the ASIL recommendation which states that the system shall be tested against worst-case errors. An example definition of a worst-case error is provided in subsection 5.3.

Table 1: Three exemplary generic requirements and their ASIL risk level classification.

Identifier	Requirement	ASIL A	ASIL B	ASIL C	ASIL D
Req. 7	The performance shall be compliant to the allowed worst-case error.	++	++	++	++
Req. 30	The training, test and evaluation datasets shall be independent from each other.	++	++	++	++
Req. 33	The model's decisions shall be explained to aid the comparison between the modelling of the system and the trained model.	++	++	++	++

Table 2: Detailed analysis of the exemplary generic requirements from Table 1.

Identifier	Applicability	Concretization	Testability	Test Procedure
Req. 7	Complex	Major	High	Metric-based
Req. 30	Simple	Minor	High	Evidence-based
Req. 33	Complex	Minor	Medium	Metric-based & Evidence-based

3.3 AI Subsystem Requirements

REQ. 33 from Table 1 is targeted towards the AI subsystem within the mobility system. This requirement is derived from an ASIL recommendation that states that the modelling of the system shall be compared to the resulting system. It is modified to fit the AI subsystem by stating that the model's decisions shall be explained. This is because AI models are similar to black-boxes and there is low to no insight into how the model's decisions are made. Therefore, specific explainability methods shall be used to gain insight into the correct functionality.

Analogously, REQ.30 states that the training, evaluation and testing datasets used during the development shall be independent from each other. This ensures that performance or training issues can be detected during the training phase. The testing procedure for this requirement depends on the size and format of the datasets at hand. An example evaluation of this requirement is explained in subsection 5.3.

3.4 Testability and Applicability

The above-formulated requirements must be specified for each individual use case. To determine the effort needed to transfer a requirement between different mobility use cases we perform an analysis to determine the applicability and testability of each requirement. Additionally, we also provide indicators on the concretization effort between use cases and the type of test procedure. Table 2 presents the categorization for the 3 example requirements from subsection 3.2 and subsection 3.3.

4 AI-BASED SYSTEMS IN MOBILITY APPLICATIONS

To assess the suitability of the generic audit requirements proposed in section 3, we aim to perform practical tests for a concrete AI-based system. To achieve universally applicable results, this system should be representative for different use cases in mobility applications and at the same time enable efficient initial tests. Thus, to select the most suitable exemplary use case we first introduce categories, which help to assess the suitability of different use cases. Following, we present a summary of possible AI-based use cases in mobility applications. Based on the introduced categories we then analyze all collected use cases to decide which use case is best suited for the initial practical tests of the audit requirements.

4.1 Analysis Categories

To assess the suitability of different AI-based use cases in mobility applications for the practical audit requirements tests we choose five categories. These cover important aspects regarding the feasibility of the tests and the meaningfulness of the achieved results. In Table 3 these are later applied to individual use cases in mobility applications.

First, the relevance of each use case for the safety of the entire mobility system is rated as high, medium, low or none. To achieve results for the auditing of the most critical use cases it is preferable to select a use case where a certain amount of safety relevance is given and a high relevance is ideal. This allows to

develop audit criteria for critical tasks where an audit is important and required.

Next, the complexity and auditability of each use case is categorized as complex, medium or simple. This category covers the test effort required to derive the residual risk of an AI-based system implementing a use case. For the initial practical tests, it is preferable to select a use case with rather low complexity. This enables the most feasible development and to perform more extensive tests.

Third, the applicability of potential (adversarial) attacks on the AI component of each use case is rated as unrealistic, complex, medium or simple. The most important factors that influence the categorization are the scalability of an attack, the availability of literature or demonstrations of an attack and the required access interface of an adversary. Here, it is important that a direct attack interface to the AI component exists and it is best when attacks are comparatively simple to execute in practice.

Additionally, the required resources for implementing an exemplary system of each use case are rated as high, medium or low. Multiple factors like the availability of open-source datasets or representative implementations, the model size of involved AI components and the required computational resources for training or inference are considered. It is most suitable when datasets and implementations are publicly available and only low resources are required.

Lastly, the generalizability of the results of the practical audit requirements tests to other use cases is categorized as high, medium or low. Most importantly, this categorization considers which sensors are used, which perception components are involved and whether the use case impacts the planning or control of a vehicle. Use cases are preferable when some of these characteristics are shared with other use cases.

4.2 Potential Use Cases

Selecting a representative use case from the sheer number of potential use cases in AI-based mobility applications (Yurtsever et al., 2020; Ziebinski et al., 2016) is a difficult task. We tackle this by first collecting a list of ten high-level use cases which commonly occur in the specific areas of AD/ADAS. For such use cases the need for audit requirements and procedures is highest, as associated systems are safety-relevant and are increasingly tested on public roads.

Concretely, we start by considering AD/ADAS use cases which have a direct impact on the control of a vehicle. For all use cases we include the required perception of the respective road users or objects, meaning we do not only consider the final control algorithms. Here, the first use case is termed *collision avoidance*, which includes all functionalities that react to potential obstacles in the driving path of a vehicle, by initiating deceleration and/or steering motions. Notably, collision avoidance also contains (automatic) emergency breaking. Next, we consider the *lane keeping* use case, which includes all functionalities that keep a vehicle in the current driving lane. Here, mainly steering motions are performed to tackle the given task. Further, we consider the *lane changing* use case, which includes all functionalities that lead to a change in the driving lane of the vehicle. Like lane keeping, steering motions are most important but also acceleration or deceleration motions are required to be able to merge in between two vehicles. Fourth, the *adaptive cruise control* use case includes functionalities that manage the distance to a vehicle driving in front of the ego vehicle. Here, deceleration and acceleration motions are important to control the distance to the leading vehicle adaptively based on its driving maneuvers and speed.

After presenting use cases that directly affect the control of a vehicle, we now discuss additional use cases which have no direct control impact but are important to obtain a list of the most important AI-based AD/ADAS use cases. Therefore, the fifth use case is *global path planning*. This includes functionalities to plan the global route/path of a vehicle, which consists of the rough path from the starting location to the target location. This route is updated online during driving depending on the current occupancy of roads or the probability of traffic jams. Next, the *traffic sign assistant* use case includes all functionalities that show currently relevant traffic signs to the driver. However, this purely acts as an assistance feature and for example does not adapt the speed of a vehicle to the detected speed limit automatically. Additionally, we consider the *driver monitoring* use case. Here, the goal is to detect drowsiness or distraction of a driver and provide a warning to the driver. This can reduce the number and criticality of accidents and is important when the driver must monitor assistance functionalities and be able to intervene rapidly.

In addition to the use cases that describe a specific functionality, we also consider basic use cases which provide information for various AD/ADAS-related functionalities. Here, the *map-based localization* use case includes functionalities to determine the current position of a vehicle as it navigates through the environment. Typically, a map of the environment is used which can either be created dynamically or is created a priori. Next, the *road user detection* use case is considered, which includes functionalities to detect dynamic traffic participants like pedestrians, vehicles

Table 3: Overview of considered AI-based use cases in mobility applications and their suitability for the practical testing of audit requirements. For each parameter, the symbol in brackets indicates whether this parameter value is suitable (↑), partially suitable (o) or unsuitable (↓) for the initial testing of the proposed audit requirements.

Use Case	Safety Relevance	Complexity/ Auditability	Attack Applicability	Required Resources	Generalizability
Collision Avoidance	High (↑)	Complex(o)	Medium (o)	High (↓)	High (↑)
Lane Keeping	High (↑)	Medium (o)	Simple (↑)	Medium (o)	Medium (o)
Lane Changing	High (↑)	Complex(o)	Medium (o)	High (↓)	High (↑)
Adaptive Cruise Control	High (↑)	Medium (o)	Complex(o)	High (↓)	Medium (o)
Global Path Planning	None (↓)	Simple (↑)	Unrealistic (↓)	High (↓)	Low (o)
Traffic Sign Assistant	Low (o)	Simple (↑)	Simple (↑)	Low (↑)	Medium (o)
Driver Monitoring	Medium (o)	Medium (o)	Unrealistic (↓)	Medium (o)	Low (o)
Map-based Localization	High (↑)	Medium (o)	Complex(o)	High (↓)	Low (o)
Road User Detection	High (↑)	Complex(o)	Medium (o)	Medium (o)	Medium (o)
Behavior Prediction	High (↑)	Complex(o)	Unrealistic (↓)	Medium (o)	Low (o)

or cyclists. Lastly, the *behavior prediction* use case includes functionalities to identify the behavior and subsequently the trajectory of traffic participants. All three use cases have no direct impact on the control of a vehicle since they only provide information to functionalities for specialized use cases like collision avoidance or lane changing.

4.3 Use Case Selection

After collecting a list of use cases in AI-based mobility applications we use the categories from subsection 4.1 to assess the suitability of each presented use case in Table 3. It is important to note that the assigned values in each category must be seen relative to each other. For example, the value low only indicates that the use case is on the lower end when compared to all other presented use cases.

For the final selection of a use case, we start by dropping all use cases which do dot not fulfill the basic prerequisite for any category. This means that *global path planning* is no longer considered since it has no direct safety relevance and it is unrealistic to apply attacks which target the AI component directly. Similarly, *driver monitoring* and *behavior prediction* are also no longer considered as it is very challenging or unrealistic for an adversary to apply attacks. Next, the use cases *collision avoidance*, *lane changing*, *adaptive cruise control* and *map-based localization* are considered as unsuitable because they typically require larger model sizes and less use case specific datasets are available for training and testing.

After dropping the seven unsuitable use cases only the three use cases *lane keeping*, *traffic sign assistant* and *road user detection* are considered for further analysis. All three are in principle suitable and fulfill the basic prerequisites for all categories from subsection 4.1. The remaining use cases differ with respect to their safety relevance (higher for *lane keeping* and *road user detection*) and their resource requirements/complexity (lower for *traffic sign assistant*). To be able to test more audit requirements with the available resources we value the feasibility, i.e. the lower complexity, higher and select the *traffic sign assistant* use case for further practical investigations. More complex and safety-relevant use cases can be explored later, once it is shown that the audit requirements can be applied and provide useful results. Details on the implementation of a system representing the selected use case are given later in subsection 5.1.

5 PRACTICAL IMPLEMENTATION

To test the applicability and expressiveness for real applications we implement the selected generic audit requirements from section 3 for the traffic sign assistant use case. Thus, in the following we first introduce the detailed architecture of the exemplary ADAS system which represents the traffic sign assistant functionality. Then, we discuss the application of some interesting audit requirements and describe results and challenges.

5.1 Experimental Setup

For the traffic sign assistant use case, a single outside facing forward RGB camera sensor is used. Based on the data of this sensor, the classification (and preceding detection) of traffic signs is performed using a DNN. This mimics the approach used for real traffic sign assistants (Lim et al., 2017). For our initial practical experiments, we only consider a DNN that performs a pure classification of traffic signs. The reason

is that typically there exists a common detector for all kinds of road users and elements, like traffic signs, vehicles, pedestrian, etc. Based on the detected objects, the content of the detected bounding boxes is fed to special classification modules that specialize in concretely classifying the object in a box. For the case of a traffic sign assistant this means that a preceding road elements detector exists which outputs bounding boxes around detected signs. Then, a classifier that focusses on traffic signs is used to determine the exact sign type based on the given subset of the entire image selected by the preceding detector. The output of this system is therefore the detected traffic sign in the given image. In this work we use the classifier to test the proposed audit requirements in practice.

To train the DNN that classifies traffic signs we use the German Traffic Sign Recognition Benchmark (GTSRB) (Stallkamp et al., 2011) as training dataset. This dataset features 43 classes of German traffic signs. Using this dataset, we train a ResNet-18 (He et al., 2016) to represent the traffic sign assistant. ResNet-18 is selected because this architecture is one of the most successful architectures in DNN history and is often used in literature as a sensible baseline independent of the concrete task and use case. The training is performed without any augmentations and using only the standard GTSRB training dataset.

5.2 Generic Toolbox

To allow for an easy expansion of our tests to further mobility use cases, we design a toolbox in a modular way. In this work the toolbox is implemented in an exemplary fashion for the traffic sign assistant use case and some audit requirements discussed in subsection 5.3. The goal is to continuously expand this toolbox[1] and incorporate more audit requirements and use cases over time.

5.3 Application of Requirements

The generic catalogue of requirements that is elicited in section 3 enables a simple selection of requirements based on the risk level of the specific use case. We assume the traffic sign assistant to be an assistance system which is not able to gain automated control of the vehicle on its own. However, as input to an AD system that for example regulates the vehicle's speed in a certain range based on the detected traffic signs and additional parameters, the use case might be classified as ASIL A during homologation. Due to this assumption, we select and specify requirements that are "highly recommended" (++) for the ASIL A risk level from our requirement catalogue (AIMobilityAuditPrep, 2022).

After the requirements are selected, the evaluation process consists of the following three steps for each requirement:

1. Parameter selection: If the requirement requires the specification of parameters, the parameters are chosen according to the use case (if necessary by domain experts). Moreover, a rationale/justification how these parameters are derived is provided.
2. Description of the audit procedure: The audit procedure for the requirement is described. For "metric-based" requirements the technical evaluation/tests that are performed shall be described. For "evidence-based" requirements the procedural evaluation of evidence is described.
3. Verdict: The test results of "metric-based" tests or findings of "evidence-based" evaluations are assessed and a "pass" or "fail" verdict is given describing whether the requirement is fulfilled.

In the following, we schematically show these steps for the exemplary requirements introduced in section 3 with the traffic sign assistant use case described in subsection 5.1.

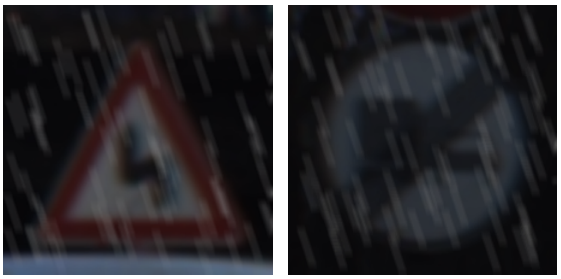
5.3.1 Requirement 7

The performance shall be compliant to the allowed worst-case error.

To fulfill this requirement the "performance" and "allowed worst-case error" have to be specified. In the case of the traffic sign assistant use case, we choose to measure the "performance" through the accuracy of the system. The "allowed worst-case error" is chosen as an accuracy greater than 90 % and we take heavy rain as an example of a worst-case error. In the scope of this work, we schematically assume the vehicle running the traffic sign assistant is operated in Germany, where heavy rain is common and a 90 % accuracy still offers sufficient reliability of the assistance system. This selection for the two parameters only serves as an example to demonstrate how the requirement can be tested in practice. Depending on the boundary conditions, operational design domain, mitigation strategies or level of automation of the overall system it could be necessary to adapt the value for the required accuracy. The rationale for the selection of these two parameters in a real-world use case requires a rationale from domain experts.

The audit procedure for this requirement is "metric-based", where a dataset (that the model was

[1] An overview of the toolbox is available at www.bsi.bund.de/dok/1079914.



(a) GTSRB class 21. (b) GTSRB class 42.

Figure 1: Examples of the heavy rain transformation.

not trained on) containing data samples of all classes in heavy rain conditions is evaluated by the model. It is possible to use data samples captured in heavy rain conditions or transform data samples from clear weather conditions with a heavy rain simulation. If the accuracy of this evaluation is greater than 90 %, the requirement is fulfilled.

Our toolbox implements a heavy rain transformation using the *albumentations* library (Buslaev et al., 2020). This allows to test the worst-case error on any suitable traffic sign dataset under heavy rain conditions. We transform images from GTSRB using their heavy rain transformation, which for example results in images depicted in Figure 1. On 2580 data samples the system reaches an accuracy of $\sim$ 79 %. Since the accuracy from the evaluation under heavy rain transformation is 79 %, which is not greater than 90 %, the requirement is not fulfilled and fails this evaluation. For real-world use cases an assessment of domain experts is required regarding the representativity of different parameters of the used heavy rain transformation like the strength or structure of the rain drops.

Alternatively, a worst case could be represented by an adversarial attack to the system. As an example we take a PGD attack (Madry et al., 2018) with a perturbation budget of 0.3. We repeat the outlined audit process but execute a PGD attack instead of applying a heavy rain transformation. Against this attack the system reaches an accuracy of $\sim$ 21 % which means REQ. 7 is also not fulfilled using this second specification. Note that PGD represents an attack in the digital domain. In reality physical attacks which are applied in the environment itself pose a larger threat and testing against such attacks is more important.

5.3.2 Requirement 30

The training, test and evaluation datasets shall be independent from each other.

Since REQ. 30 has no parameters to be set, this step is skipped. The testing procedure of this requirement is classified as "evidence-based". Hence, the dataset documentation, code and contents of each of the datasets shall be consulted. The documentation and code of the training procedure gives insight on how these datasets are generated. In this example, the evidence showed that the datasets were split before training the model into three disjoint datasets. Also, the datasets follow the same underlying distribution and are independent. Therefore, the requirement is fulfilled. It is important to use independent splits of the data to get a fair assessment of the quality of a model. For example, images from a video recording of a single scene should not be used in different splits. Instead, images from a different recording like a different scene or in different weather must be used.

5.3.3 Requirement 33

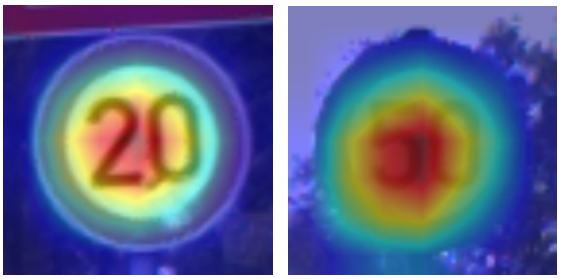
The model's decisions shall be explained to aid the comparison between the modelling of the system and the trained model.

For this requirement, the method used for explaining the decision and the system modelling the decisions have to be determined. In the schematic traffic sign assistant use case, we choose the following exemplary functional system requirement: *The model decision on a traffic sign image shall depend on the figure displayed by the traffic sign, the signs coloration and/or the shape of the sign.* In real-world applications depending on the chosen modelling, it would also be possible to implement automatic testing detecting whether a certain amount of background information is considered for the model's decisions. In our case, we choose the *GradCam* explainability method (Selvaraju et al., 2017) to explain a random set of 60 images per GTSRB class. Figure 2 presents some examples of a GradCam explanation on some images of the GTSRB dataset. It clearly shows that the most important information for the decision (highlighted in red) of the model is based on the center of the image. We analyze all 60 images for each class and they show similar results. Hence, this evaluation is passed and the requirement is fulfilled.

6 CONCLUSION

6.1 Summary

We introduce a list of generic audit requirements, which are technically relevant to assure the trustworthiness, security, safety, robustness, explainability, etc. of AI-based systems in mobility applications. These requirements evolved under attention



(a) GTSRB class 0. (b) GTSRB class 2.

Figure 2: Examples of the GradCam explanation. The information with the highest influence on the model's decision is highlighted in red.

to existing regulations, norms, guidelines and an extensive literature review. Additionally, we implement tools for exemplary audit requirements to demonstrate the applicability using a selected mobility application. For this, we perform a comparison of different AD/ADAS use cases, based on various categories like complexity, auditability, available resources. Using this analysis, we determined the traffic sign assistant use case to be best suited for the initial practical testing of the audit requirements. Thus, we examine two exemplary DNNs trained on German traffic signs using the implemented audit requirements. We find that the generic audit requirements can be specified to provide meaningful results on the DNN-based traffic sign assistants for different AI-specific properties.

6.2 Outlook

As discussed in subsection 5.3 we only use a subset of all proposed audit requirements for the initial practical tests. A natural next step is to extend the practical tests to include all proposed requirements. Additionally, one can expand the extent of the already implemented requirements. Some of these requirements are quite extensive and can be implemented for practical tests in different ways. In a follow-up work the exemplary implementation can be expanded to cover further aspects of the associated audit requirements. This enables more extensive audits and increases the meaningfulness of the obtained results.

Furthermore, it is especially interesting to test some audit requirements using actual hardware and test facilities. Instead of performing all tests in a simulation environment, the most interesting audit requirements should also be tested in reality. Only these tests enable to properly assess the feasibility and expressiveness of the proposed audit requirements.

Additionally, the complexity of the audited system should be increased. Instead of using only a DNN-based classifier, the system should be extended to be more representative of systems used in reality. Ideally, this is complemented by the application of the audit requirements to industry systems operating in practice. This allows judging the applicability under real-world conditions and limitations.

Our goal is to continue this work and to consider at least one additional use case in addition to the traffic sign assistant. We are working actively on the outlined next steps to further increase the meaningfulness of our results and refine the proposed requirements and best practices based on practical insights and limitations. We want to move towards applying the audit requirements in practice and create a formal technical guideline. The obtained results could then be used as a blueprint for standardization activities and should be introduced to the relevant committees.

ACKNOWLEDGEMENTS

This work was supported by the Federal Office for Information Security (BSI), Germany, in project P538 *AIMobilityAuditPrep*.

REFERENCES

AIMobilityAuditPrep (2022). AIMobilityAuditPrep: Final Results - Documentation. Technical report, German Federal Office for Information Security. www.bsi.bund.de/dok/1079912.

ANSI/UL 4600 (2022). Standard for Safety for the Evaluation of Autonomous Products. Standard, American National Standards Institute, Underwriters Laboratories.

Berghoff, C., Neu, M., and von Twickel, A. (2020). Vulnerabilities of Connectionist AI Applications: Evaluation and Defence. *Frontiers in Big Data*, 3.

Buslaev, A., Parinov, A., Khvedchenya, E., Iglovikov, V., and Kalinin, A. (2020). Albumentations: Fast and Flexible Image Augmentations. *Information*, 11.

DIN Roadmap AI (2022). German Standardization Roadmap on Artificial Inteligence. Draft, German Institute for Standardization.

EU AI Act (2021). Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts. Draft, European Commission.

Geirhos, R., Jacobsen, J., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020). Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2:665–673.

Gilpin, L., Bau, D., Yuan, B., and Bajwa, A. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. In *International Conference on Data Science and Advanced Analytics*, Turin, Italy.

Goldblum, M., Tsipras, D., Xie, C., and Chen, X. (2020). Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses. *IEEE Transactions on Pattern Analysis & Machine Intelligence*.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A Survey Of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51:1–42.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA.

Hendrycks, D. and Dietterich, T. (2019). Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*, New Orleans, USA.

ISO 21448 (2022). Road vehicles - Safety of the intended funtionality. Standard, International Organization for Standardization.

ISO 26262 (2018). Road vehicles - Functional safety. Standard, International Organization for Standardization.

ISO/AWI PAS 8800 (2022). Road vehicles - Safety and artificial intelligence. Standard, International Organization for Standardization.

ISO/AWI TS 5083 (2022). Road vehicles - Safety for automated driving systems - Design, verification and validation. Standard, International Organization for Standardization.

ISO/IEC DTR 5469 (2022). Artificial intelligence - Functional safety and AI systems. Standard, International Organization for Standardization, International Electrotechnical Commission.

ISO/IEC PRF TS 4213 (2022). Artificial intelligence - Assessment of machine learning classification performance. Standard, International Organization for Standardization, International Electrotechnical Commission.

ISO/IEC TR 24028 (2020). Artificial intelligence - Overview of trustworthiness in artificial intelligence. Standard, International Organization for Standardization, International Electrotechnical Commission.

ISO/IEC TR 24029-1 (2021). Artificial intelligence - Assessment of the robustness of neural networks - Part 1: Overview. Standard, International Organization for Standardization, International Electrotechnical Commission.

ISO/SAE 21434 (2021). Road vehicles - Cybersecurity engineering. Standard, International Organization for Standardization, SAE International.

Karpathy, A. (2021). Workshop on Autonomous Driving - Tesla Keynote. In *Conference on Computer Vision and Pattern Recognition*, Nashville, USA.

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., and Zhou, B. (2022). Trustworthy AI: From Principles to Practices. *ACM Computing Surveys*.

Lim, K., Hong, Y., Choi, Y., and Byun, H. (2017). Real-time Traffic Sign Recognition based on a General Purpose GPU and Deep Learning. *PLOS ONE*, 12:1–22.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*, Vancouver, Canada.

Mohseni, S., Pitale, M., Singh, V., and Wang, Z. (2020). Practical Solutions for Machine Learning Safety in Autonomous Vehicles. In *Conference on Artificial Intelligence: Workshop on Safe AI*, New York, USA.

Orekondy, T., Schiele, B., and Fritz, M. (2019). Knockoff Nets: Stealing Functionality of Black-Box Models. In *Conference on Computer Vision and Pattern Recognition*, Long Beach, USA.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, B., and Swami, A. (2017). Practical Black-Box Attacks against Machine Learning. In *ACM Asia Conference on Computer and Communications Security*, Abu Dhabi, United Arab Emirates.

Radlak, K., Szczepankiewicz, M., Jones, T., and Serwa, P. (2020). Organization of Machine Learning based Product Development as per ISO 26262 and ISO/PAS 21448. In *Pacific Rim International Symposium on Dependable Computing*, Perth, Australia.

Raji, I. D., Smart, A., White, R., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *ACM Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain.

SAE J3016 (2014). Levels of Driving Automation. Standard, SAE International.

Schwarzschild, A., Goldblum, M., Gupta, A., Dickerson, J., and Goldstein, T. (2021). Just How Toxic is Data Poisoning? A Unified Benchmark for Backdoor and Data Poisoning Attacks. In *International Conference on Machine Learning*, Vienna, Austria.

Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *International Conference on Computer Vision*, Venice, Italy.

Stallkamp, J., Schlipsing, M., Salmen, J., and Igel, C. (2011). Man vs. Computer: Benchmarking Machine Learning Algorithms for Traffic Sign Recognition. In *International Joint Conference on Neural Networks*, San Jose, USA.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing Properties of Neural Networks. In *International Conference on Learning Representations*, Banff, Canada.

UNECE R 155 (2021). Uniform provisions concerning the approval of vehicles with regards to cyber security and cyber security management system. Standard, United Nations Economic Commission for Europe.

Waymo (2021). How we've built the World's Most Experienced Urban Driver. *Waypoint*.

Yurtsever, E., Lambert, J., Carballo, A., and Takeda, K. (2020). A Survey of Autonomous Driving: Common Practices and Emerging Technologies. *IEEE Access*, 8:58443–58469.

Ziebinski, A., Cupek, R., Erdogan, H., and Waechter, S. (2016). A Survey of ADAS Technologies for the Future Perspective of Sensor Fusion. In *International Conference on Computational Collective Intelligence*, Halkidiki, Greece.