Whitepaper | May 2022

# Towards Auditable AI Systems

## From Principles to Practice

based on the 2nd international Workshop “Towards Auditable AI Systems”, October 26th 2021, Fraunhofer Forum Digitale Technologien, Berlin, organized by the Federal Office for Information Security Germany, the TÜV-Verband and the Fraunhofer HHI

Christian Berghoff[1], Jona Böddinghaus[14], Vasilios Danos[6], Gabrielle Davelaar[13], Thomas Doms[3], Heiko Ehrich[6], Alexandru Forrai[8], Radu Grosu[9], Ronan Hamon[10], Henrik Junklewitz[10], Matthias Neu[1], Simon Romanski[11], Wojciech Samek[7,15]*, Dirk Schlesinger[4], Jan-Eve Stavesand[12], Sebastian Steinbach[2]*, Arndt von Twickel[1]*, Robert Walter[4], Johannes Weissenböck[3], Markus Wenzel[7], Thomas Wiegand[7,15] (Authors are listed in alphabetical order)

*Contact:
Arndt von Twickel (arndt.twickel@bsi.bund.de),
Wojciech Samek (wojciech.samek@hhi.fraunhofer.de) and
Marc Fliehe (marc.fliehe@tuev-verband.de)

[1]Federal Office for Information Security Germany (BSI), [2]TÜV-Verband, [3]TÜV Austria, [4]TÜV AI Lab, [5]TÜV Süd, [6]TÜV Nord / TÜVit, [7]Fraunhofer HHI, [8]Siemens Digital Industries Software, The Netherlands, [9]Technische Universität Wien, [10]European Commission, Joint Research Centre (JRC), [11]understand.ai GmbH, [12]dSPACE GmbH, [13]Microsoft, [14]Gradient Zero Deutschland GmbH, [15]TU Berlin

## Abstract

As a key technology Artificial Intelligence (AI), especially in the form of deep neural networks, is already omnipresent in many digitization applications, including security and safety relevant applications in domains such as biometrics, healthcare and automotive. Despite its undisputed benefits, the use of AI also entails qualitatively and quantitatively new risks and vulnerabilities. Together with its increasing dissemination, this calls for audit methods that allow to give guarantees concerning the trustworthiness and that allow to operationalize emerging AI standards and AI regulation efforts, e.g. the European AI act. Auditing AI systems is a complex endeavour since multiple aspects have to be considered along the AI lifecycle that require multi-disciplinary approaches. AI audit methods and tools are in many cases subject of research and not practically applicable yet. To allow for a comprehensive inventory of the auditability of AI systems in different use cases and to allow for tracking its progress over time, we here propose to employ a newly developed "Certification Readiness Matrix" (CRM) and present the initial concept. By using the CRM concept as a frame to summarize the results of a one day workshop on auditing AI systems with talks covering basic research, applied AI auditing efforts and standardisation activities we demonstrate that audit methods for some aspects are already well developed while other aspects still require more research into and development of new audit technologies and tools.

# Inhalt

## 1 Problem Statement

Due to the increase in computing power and the availability of large data sets, the performance of AI systems has significantly improved in recent years, and they are now routinely used in an ever-increasing number of applications ranging from biometrics and healthcare to the automotive domain. In spite of the large opportunities offered by AI systems, they also bring up new challenges for auditing as compared to classical software [1]. Firstly, the input and state spaces of AI systems for common tasks are enormous, rendering exhaustive testing infeasible. Second, their behaviour strongly depends on the data used to train them, and inconsistencies or deliberate manipulations of these data can engender grave consequences. Third, most AI systems currently used have a complex inner structure which does not lend itself to human interpretation. This makes finding malfunctions and attacks and mitigating them a very hard task. In order to address these challenges and to facilitate the secure, robust and transparent application of AI, especially in security and safety-critical applications, it is necessary to have available technical requirements as well as concepts, methods and tools for auditing AI systems accordingly. However, such material is largely missing so far.

This document focuses on connectionist AI systems, which are used in most complex applications (e.g. based on image processing) today. Connectionist AI systems are large data structures that contain millions of parameters. The currently most widespread examples are deep neural networks (DNNs), which consist of various layers of simple processing elements (neurons) that are highly interconnected. The life cycle of connectionist AI systems is complex and consists of different phases. After an initial planning phase, in which use case-specific requirements and ambient conditions are taken into account, data of sufficient quality and quantity are collected. These data are then used to train the parameters of a connectionist AI system to approximate the intended functionality on the data set. The performance of the AI system is then evaluated on a different data set not used for training to make sure it properly generalises. If the evaluation results are satisfactory, the AI system is then deployed and put into operation. In practice, the deployed AI system is not operating in a vacuum, but is embedded both in a complete (physical and) software system and in organisational processes and is interacting with the physical world in the context of the use case it addresses [2].

In April 2021, the European Commission published a draft regulation on AI (the AI act, AIA) [3], whose goal is to ensure that AI systems in use fulfil adequate requirements. The AIA pursues a risk-based approach, which bans some applications of AI altogether (e.g. social scoring schemes) and imposes comprehensive requirements on AI systems considered high-risk. According to the AIA, high-risk applications include, among others, the use of AI in safety-critical functions as well as in the health and justice sector and in law enforcement. The requirements in the AIA are set out on a very high level. In order for the AIA to be operationalised, these requirements need to be underpinned on the technical level. As a result, it is necessary to develop audit schemes, methods and tools for all aspects mandated by the AIA across the relevant AI life cycle phases within the next two to three years, when the AIA will start to apply.

The development of AI technology is highly dynamic, and many technical questions relating to the AIA requirements are the subject of research by academics and of new approaches developed by industry. As a consequence, the development of corresponding standards, audit schemes, methods and tools is likewise moving fast (cp. e.g. [4-13]). A periodic survey is thus needed to give an overview of the current status and progress in the auditability of AI systems, both for specific applications and in general. Based on the current status and the gaps identified, research and development must be guided and prioritised to systematically advance AI auditability until the concepts and tools are sufficiently mature to operationalise the AIA.

The present document focuses on those requirements from the AIA that are related to IT security and safety, which are technical in nature and can be objectively assessed. These criteria include security, safety, performance, robustness, interpretability and explainability, and traceability of AI systems as well as risk management procedures. Further requirements from the AIA that address user-focused fundamental ethical questions relating to bias, data privacy and human oversight are out of the scope of the present document. Since these aspects are also very important, it is necessary for them to be covered elsewhere.

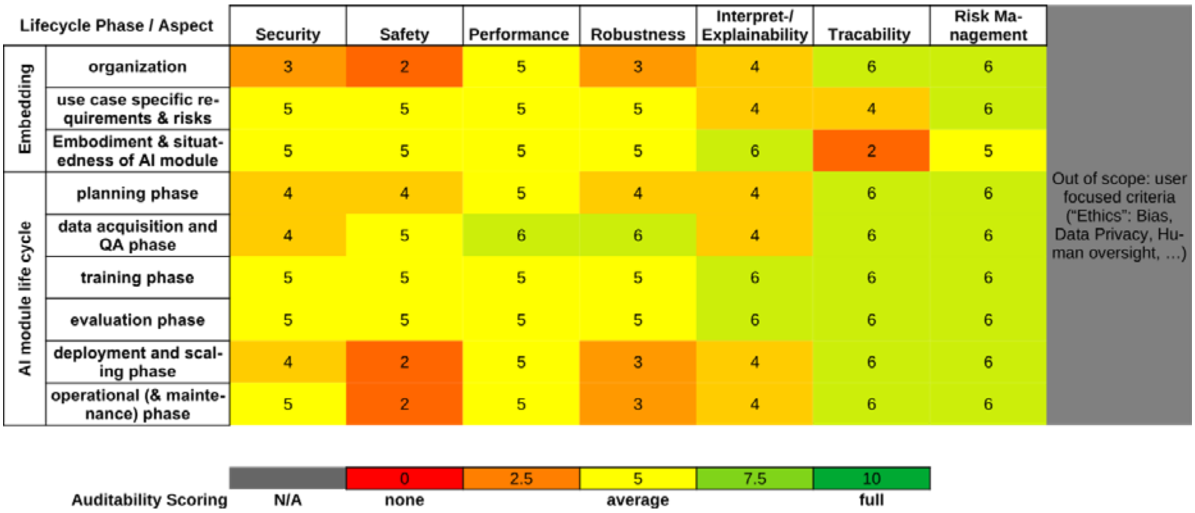
## 2 Proposal of a Certification Readiness Matrix as a Tool to Monitor the Progress of AI Auditability[1]

Assessing AI auditability is a complex problem since, on the one hand, it requires to take into account the whole life cycle of an AI system, many different technical aspects and use case-specific ambient conditions and, on the other hand, both research and development in the field are highly dynamic.

To capture this complexity while achieving a high level of transparency and comparability, we propose a two-dimensional “Certification Readiness Matrix” as a tool to monitor the progress

[1] Please note that the "Certification Readiness Matrix" presented here is meant as a conceptual heuristic (e.g. it does not aspire to serve as a certification scheme specifically for BSIG §9 or the EU Cyber Security Act).

in AI auditability (for an example see Figure 1). The first dimension of the matrix covers the AI life cycle phases and the embedding of an AI system within organisational processes, in the context of its use case and in the physical world. The second dimension lists important technical and objectively assessable aspects related to IT security and safety, which are listed and defined hereafter:



| | Lifecycle Phase / Aspect | Security | Safety | Performance | Robustness | Interpret-/ Explainability | Tracability | Risk Management | |
|---|---|---|---|---|---|---|---|---|---|
| Embedding | organization | 3 | 2 | 5 | 3 | 4 | 6 | 6 | Out of scope: user focused criteria ("Ethics": Bias, Data Privacy, Human oversight, …) |
| | use case specific requirements & risks | 5 | 5 | 5 | 5 | 4 | 4 | 6 | |
| | Embodiment & situatedness of AI module | 5 | 5 | 5 | 5 | 6 | 2 | 5 | |
| AI module life cycle | planning phase | 4 | 4 | 5 | 4 | 4 | 6 | 6 | |
| | data acquisition and QA phase | 4 | 5 | 6 | 6 | 4 | 6 | 6 | |
| | training phase | 5 | 5 | 5 | 5 | 6 | 6 | 6 | |
| | evaluation phase | 5 | 5 | 5 | 5 | 6 | 6 | 6 | |
| | deployment and scaling phase | 4 | 2 | 5 | 3 | 4 | 6 | 6 | |
| | operational (& maintenance) phase | 5 | 2 | 5 | 3 | 4 | 6 | 6 | |

| Auditability Scoring | | 0 | 2.5 | 5 | 7.5 | 10 |
|---|---|---|---|---|---|---|
| | N/A | none | | average | | full |

Figure 1: An example of a certification readiness matrix is here shown for the aggregated results of the workshop, i.e. based on the workshop presentations it shows what is currently maximally possible with respect to auditing an AI system. One dimension (rows) represents the phases of the AI lifecycle and its embedding in the physical world, the use case and an organization. The other dimension (columns) represents those aspects of an AI system that were at the focus of the workshop, i.e. technical and objectively assessable aspects related to IT-security and safety. The auditability scoring scale below the matrix shows color and point scales that correspond to a scale from non-existing auditability (red, 0) to full auditability (green, 10). Since detailed scoring requirements have not been fully worked out so far, the scores shown here were derived by intuition drawing on the authors' experience and technical knowledge and only give a rough approximation. The goal is to fully work out a scoring requirement scheme within the next year.

- **security**: IT security, i.e. passive and active robustness of the AI system against attacks, especially AI-specific attacks (adversarial, poisoning and privacy attacks) and with respect to the three security goals integrity, confidentiality and availability
- **safety**: Protection of individuals, organizations and assets against (physical) harm
- **performance:** Performance of the AI system with respect to relevant performance metrics
- **robustness**: Passive and active robustness against natural variations of inputs (situations), including those, that were not covered during training
- **interpretability** and **explainability**: Ability of humans to understand the decision process of the AI system, either through inherently interpretable models or post-hoc

interpretation

- **traceability**: Traceability of the AI system throughout the life cycle, e.g. of design decisions, boundary conditions, data, models, training algorithms and processes, evaluation and operation using e.g. technical documentation and logging
- **risk management**: Identification, analysis and prioritization of risks and coordinated application of resources for a minimization of risk probability and/or impact; includes strategic and operational measures

While we restrict our attention to these aspects, the matrix may be easily extended by further aspects such as bias and data privacy. Where applicable, our goal is to assign certification readiness scores for each entry in the matrix. This way, the degree of auditability can be compared between different applications and across time.

While the matrix captures many important aspects and allows displaying the current status in a compact way, it does not cover every aspect that might be of importance for specific applications, e.g. ethical aspects are currently excluded. Since many aspects are partly overlapping and may be grouped in different ways, e.g. in a hierarchical tree-like structure, the proposed matrix is one of multiple possible views on certification readiness.

The requirements for assigning specific scores have not been worked out in detail and further work in this direction is required. The current scores in the matrix are rough estimates based on experience and intuition of the authors. They only serve to illustrate the method and their informative value should not be overrated. Our goal is to develop a systematic list of scoring requirements for all entries of the matrix within one year and to update the matrix subsequently. Once this prerequisite is achieved, the matrix should serve to track the progress in auditability for both individual applications and AI systems in general.

It is crucial to note that the technical challenges in reaching a sufficient degree of auditability vary massively between the different aspects considered and the life cycle phases. On the one hand, it is relatively straightforward to extend and audit requirements for classical risk management and for transparency and traceability. On the other hand, requirements and audit methods for IT security, safety, robustness and interpretability, especially with respect to DNNs, are the subject of current research and development efforts. Significant breakthroughs facilitating full auditability are unlikely to happen in the near future.

In Figure 1 an example of certification readiness matrix is given: it has been derived by first assigning scores to individual applications based on the talks at the 2021 workshop and then aggregating the results by taking the maximum over all talks to reflect what is already possible in principle.

# 3 Summary of the Workshop

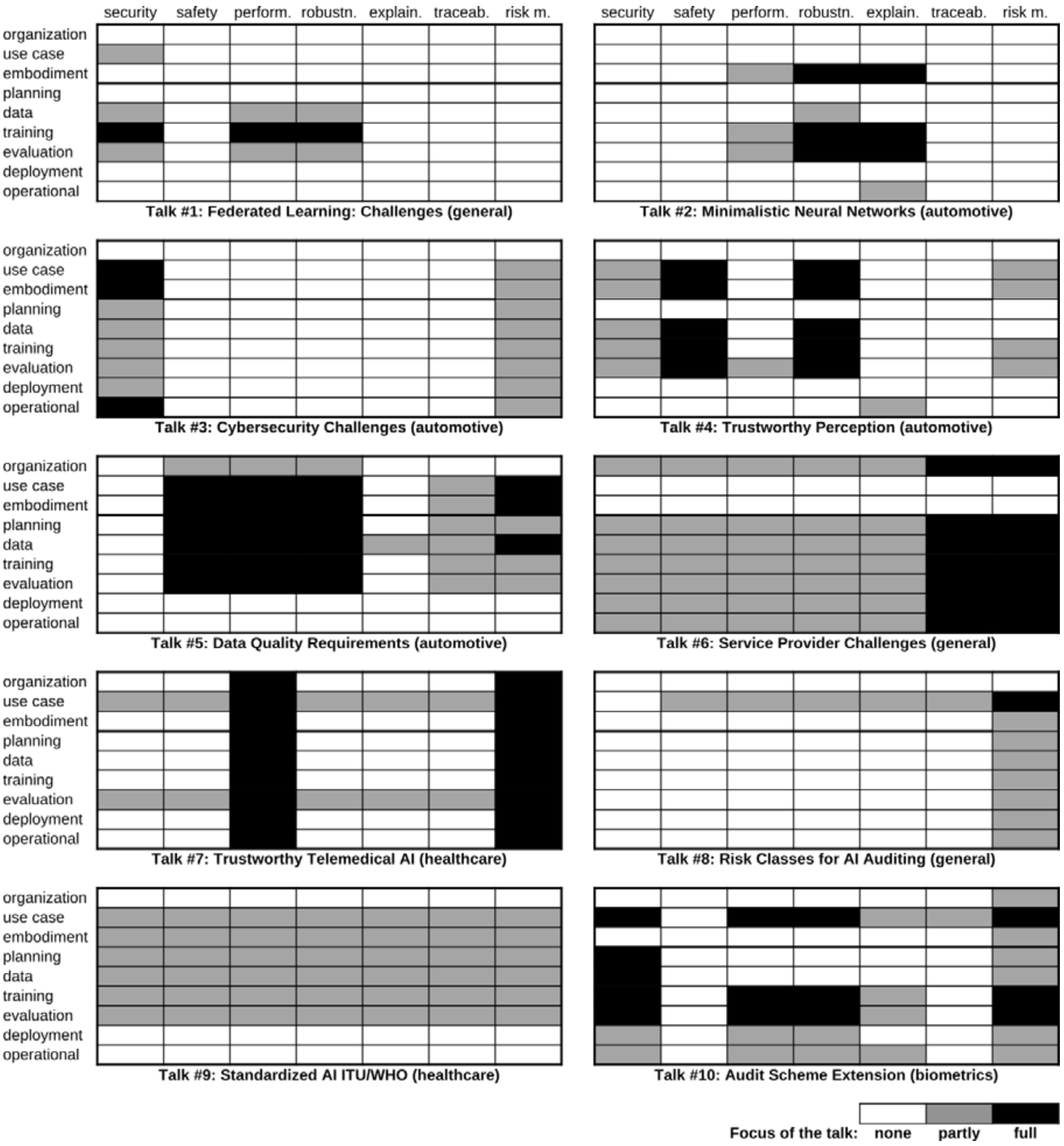

security
safety
perform.
robustn.
explain.
traceab.
risk m.
organization
use case
embodiment
planning
data
training
evaluation
deployment
operational
Talk #1: Federated Learning: Challenges (general)
Talk #2: Minimalistic Neural Networks (automotive)
Talk #3: Cybersecurity Challenges (automotive)
Talk #4: Trustworthy Perception (automotive)
Talk #5: Data Quality Requirements (automotive)
Talk #6: Service Provider Challenges (general)
Talk #7: Trustworthy Telemedical AI (healthcare)
Talk #8: Risk Classes for AI Auditing (general)
Talk #9: Standardized AI ITU/WHO (healthcare)
Talk #10: Audit Scheme Extension (biometrics)
Focus of the talk:
none
partly
full

Figure 2: For each talk of the workshop, the coverage of aspects and life cycle phases are schematically indicated using the same matrix layout as in Fig. 1. Black cells indicate a focus aspect of the talk with good coverage, grey indicates an incomplete coverage and white indicates that the aspect was not covered by the talk. No single talk covered all aspects but the talks complemented each other well and together covered a wide range of aspects and many important aspects in detail. Note: The matrix layout was developed after the workshop. Neither did the speakers have access to it nor were they requested to cover as many aspects as possible.

During the workshop many AI auditability aspects of different AI lifecycle phases have been covered by 10 talks in a complementary way (cp. Figure 2), ranging from scientific challenges and new approaches (Talks 1-3) to experiences gained in industry use cases (Talks 4-7) and finally to standardization of AI (Talks 8-10). The certification readiness results aggregated across all talks are shown in Figure 1.

Since the scoring requirements have not been completely developed yet, the results derive from estimations based on experience and have to be treated with care. To avoid possible misunderstandings, individual score matrices for the talks / use cases are not shown here. As expected, the currently highest degree of auditability was visible with regard to the aspects of traceability, risk management and partly performance, as these may largely be derived from classical audit approaches.

Most work is still open with respect to the aspects of security, safety, robustness and interpretability. Here the technical challenges are, in the authors' opinion, the most demanding.

# 4 Short Summary of the Talks and Contextualization with Regard to the Matrix[2]

## 4.1 Talk #1: Federated Learning: Challenges for Security and Safety

The first talk of the workshop discussed the security- and safety-related challenges of federated learning (FL), which is a newly established training paradigm for privacy-preserving collaborative optimization of AI models. Federated learning allows multiple parties to jointly train an AI model (e.g., a deep neural network) on their combined data, without any of the participants having to reveal their local data to a centralized server. Instead of exchanging data, in FL participants exchange "knowledge" derived from their local data in form of parameter updates. The local data remains at the clients at all times, which is advantageous wrt. privacy and ownership (other participants do not get access to local data), security (since learning is decentralized, there is no single point of failure), and efficiency (no need to move data around).

However, despite these great advantages the distributed optimization of AI models leads to a set of new challenges, e.g., privacy, security, robustness, model personalization and data heterogeneity. The talk formulated four exemplary questions, which relate to important security and safety challenges:

[2] Please note that the talk summaries reflect the opinions of the respective speakers and not necessarily those of all authors and the organizers (BSI, TÜV-Verband, Fraunhofer HHI).

1. What if a (group of) client is an adversary?
2. What if we do not trust the server?
3. Can we deploy FL on the blockchain?
4. Can we ensure privacy of local data?

New developments in the FL research were presented which (to some extend) address these four questions. First, the Clustered FL approach [14] was presented, a newly developed variant of FL which measures the similarities in the clients' data distributions and thus allows to train personalized FL models. This approach has been showed to be able to identify adversarial clients, even in Byzantine settings where a subset of clients behaves unpredictably or tries to disturb the joint training effort. Then, very briefly homomorphic encryption has been presented as a potential technology to secure the connections between clients and the server. Since operations such as averaging (aggregation of model updates) and scalar product (cosine similarity) can be performed in the encrypted space, homomorphic encryption can be easily integrated within (Clustered) FL. Third, the recently proposed approach [15] to efficiently implement FL on a blockchain was discussed. The approach is based on communication-efficient federated distillation [16], which allows it to aggregate the heavily compressed (1-bit encoding) soft-label updates on a smart contract. A key feature of the work is also the incentive mechanism, which allows it to reward honest participation based on peer consistency in an incentive compatible fashion. Finally, a provably differentially private FL approach was presented and its performance was compared to other privatized and non-private baselines. The talk concluded with a recap and a short discussion of other open challenges in FL.

### 4.2 Talk #2: How may Minimalistic Neural Networks Improve Interpretability

In this talk the results of an international research team from TU Wien (Vienna), IST Austria and MIT (USA) were presented: a novel biologically inspired AI-system that can control a vehicle with just a few artificial neurons [17]. The system has decisive advantages over previous deep learning models: It copes much better with noisy inputs, and, because of its simplicity, its mode of operation can be explained in detail. It does not have to be regarded as a complex "black box", but it can be understood by humans.

Biological inspiration comes from the brain of C. elegans, a threadworm, that possesses 302 neurons and 8000 synapses and for which the connectome, i.e. the connectivity between all neurons, is completely mapped. From the connectome of C. elegans specific (feedback) policy motifs (or sensori-motor-circuit blueprints) are identified and used as building blocks for an artificial neural network. In contrast to the currently extremely popular deep neural networks

(DNNs) the networks presented here firstly consist of neurons which are based on slightly more complex biophysical models, secondly consist of only a small number of neurons and their connectivity is thirdly highly sparse, i.e. not every neuron is connected to every other neuron. Overall this massively reduces complexity and enhances interpretability of the neural network models.

These models for example may be used for parallel parking or Autonomous Lane Keeping systems in vehicles: for autonomous lane keeping a modular system consisting of pre-processing convolutional neural networks (CNNs) and a biologically inspired control network called neural circuit policy (NCP) is employed. Camera images of the road are fed into the CNNs which output 32 virtual sensor neurons which are in turn used as inputs for the NCP lane keeping network which automatically decides whether to steer to the right or left. The NCP only consists of 19 neurons. Both subsystems are stacked together and are trained simultaneously using many hours of traffic videos of human driving in the greater Boston area, together with the desired steering command. After learning the system can independently handle new situations. In comparison to todays state of the art deep learning lane keeping models with many millions of parameters the new modular approach allows to reduce the size of the control network by three orders of magnitude.

The model was evaluated with respect to its attention focus during driving. Hereby, the role of every single cell in any driving decision could be identified and the function of individual cells and their behaviour understood. Achieving this degree of interpretability is impossible for larger deep learning models. It was also tested how robust NCPs are compared to previous deep learning models. The input images were perturbed and it was evaluated how well the agents can deal with the noise. The NCPs demonstrated strong resistance to input artifacts as a direct consequence of the novel neural model and the architecture. Using this new method training time is reduced and interpretability improved while retaining the power of employing AI technology.

### 4.3 Talk #3: Cybersecurity Challenges in the Uptake of AI in Autonomous Driving

In the automotive domain, AI systems, in particular deep neural networks, have largely contributed to advancing the automation of vehicles in recent years. The most widely used standard [18] in this context defines six different levels of increasing automation, ranging from no automation (level 0) up to full automation under certain (level 4) or all (level 5) environmental conditions. Typically, these automated vehicles include a perception module that is gathering input from various types of sensors (cameras, lidars, radars. etc.), upon which a planning module returns a sequence of actions to follow a calculated trajectory, which results in an adequate adjustment of the actuators. All of these modules make an increasing use of AI, in particular in the perception stage that heavily rely on computer vision techniques.

Even if AI led to significant advances in terms of automation, systems, in particular those based on machine learning (ML) models, are susceptible to novel attacks with potentially severe consequences [19]. This presentation is based on a report evaluating the new cybersecurity challenges that come with this uptake of AI in autonomous driving [20]. Attackers may for instance insert manipulated data in datasets to alter the integrity of the training phase of the models (poisoning attacks), or try to rebuild the model based on targeted queries in order to extract the logical mechanisms at play in the decision-making and steal company's intellectual property (model stealing). Nowadays, adversarial attacks have received the most attention for their ability to deceive AI models at inference time into returning wrong outputs through the alteration of input data. Many methods to efficiently

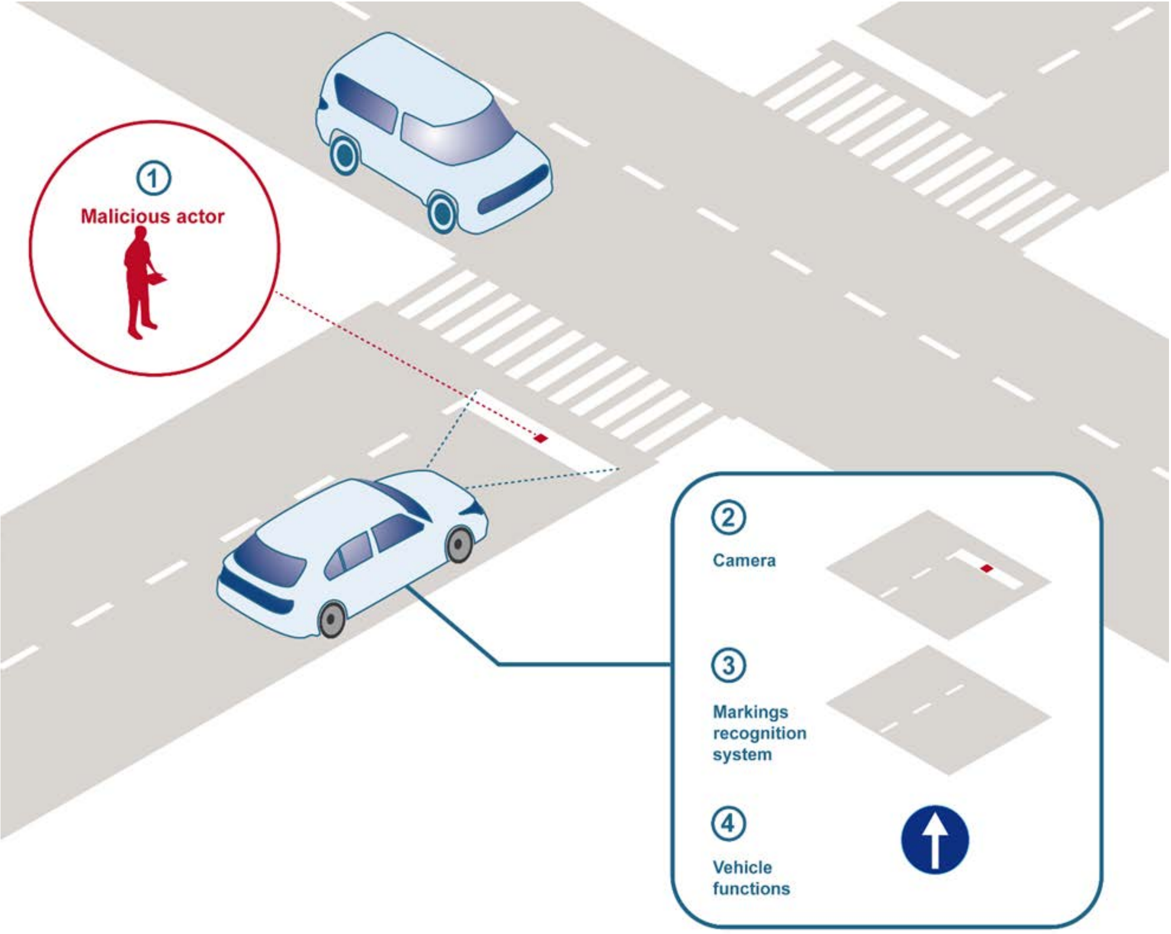


Figure 3: Attack scenario for autonomous vehicles: adversarial attacks on street markings. 1. A malicious actor applies an adversarial sticker with physical perturbations onto a stop marking. 2. The camera of the AV sees the stop markings and the adversarial sticker. 3. The markings recognition system is deceived into perceiving the stop marking. 4. The planning and control systems handle the situation as if there was no stop creating considerable safety risks for traffic actors.

perform such attacks have been proposed in the literature, but depending on the specific constraints (e.g. digital vs. physical domain, white-box vs. black-box access to model under attack, etc.), the challenges for running successful attacks can vary considerably in practice. The same is true for devising mitigation measures and defences, which usually can only secure models in a limited way.

Both development of attacks and defences are part of the active research field of adversarial ML. Demonstration in laboratory and real-world conditions of successful attacks on perception systems in the context of autonomous vehicles suggests the relevance of the cyber threats targeting AI components. On this basis, the development of a range of typical attack scenarios constitutes a starting step for a more in-depth threat modelling of AI models in autonomous driving (see Figure 3).

General recommendations to tackle the security challenges faced by AI systems include a systematic validation of the AI model and the associated data throughout the entire life cycle while properly integrating these measures with traditional cybersecurity. On a high level, it is also important to increase the capacity and the expertise on the cybersecurity aspects of AI within the automotive industry.

### 4.4 Talk #4: Towards Trustworthy Camera-Based Sensing and Perception Systems

The control of an automated driving system relies on a hierarchical and distributed control system which is embedded in a sense-plan-act loop and is subject to external disturbances and noise. The development of highly automated driving systems faces three overarching challenges that require proper tradeoffs: a) the technological challenge to build a safe car that perceives the environment and takes decision at least as good as a human driver, b) the regulatory challenge to build a functional car, accepted by society that fits into defined regulatory framework and c) the business challenge to build a cost-effective car that consumers are willing to use.

Furthermore, the trustworthiness of the system hereby is defined as: the users can trust such a system if a) it has been certified by a certification authority after development and before release and b) any of its behavior can be well explained throughout its lifetime. Two relevant standards – supporting the development and the certification process are: the ISO 26262 [21] functional safety standard, which deals with the question of how the system should detect and respond to failures, and the ISO 21448 [22] complementing standard “SOTIF safety of the intended functionality”, which covers the situations like how the system should detect and respond to functional insufficiencies of the intended functionality or by reasonably foreseeable misuse by persons. Further relevant standards in this context are e.g. ISO/TR 4804 [23] and ISO/DIS 22737 [24].

The objectives of these standards are to support the validation and certification process of the automated function in all relevant scenarios, especially in difficult conditions for both sensors and algorithms. Since even the best sensing and perceptions system will eventually fail a sensing architecture with fault detection or even fault tolerance is highly recommended. Therefore, robust and reliable sensing and perception systems are often based on the fusion of the information of multiple sensors, e.g. camera, radar and lidar. Such a system is expected to operate properly in the system's operational design domain (ODD) which includes different environmental conditions (e.g. city vs. highway and daytime vs. nighttime).

One of the technical challenges is how to develop and certify robust and reliable sensing and perception systems? Robustness is the ability of a system to resist change without adapting its initial configuration. Its relevance becomes obvious with the increasing prevalence of neural networks in the sensing and perception systems. Especially in safety critical applications it is of high importance to verify their behavior, but there is no generally agreed methodologies and procedures to assess robustness. Often, definitions of robustness assessment focus on the worst-case (adversarial inputs), which might be too conservative. One of the key questions is: how big is the tolerable uncertainty? Uncertainty may be divided in aleatoric uncertainty, which is due to randomness, and epistemic uncertainty which can be divided in model uncertainty and parametric uncertainty. All types of uncertainty may be observed along the image processing pipeline and they may be quantified using the p-norm. For example, aleatoric uncertainty is caused by hardware faults and targeted as well as untargeted adversarial attacks. Parametric uncertainty is mainly due to parameters variations due to external variations (e.g. temperature, humidity, etc.) as well as aging.

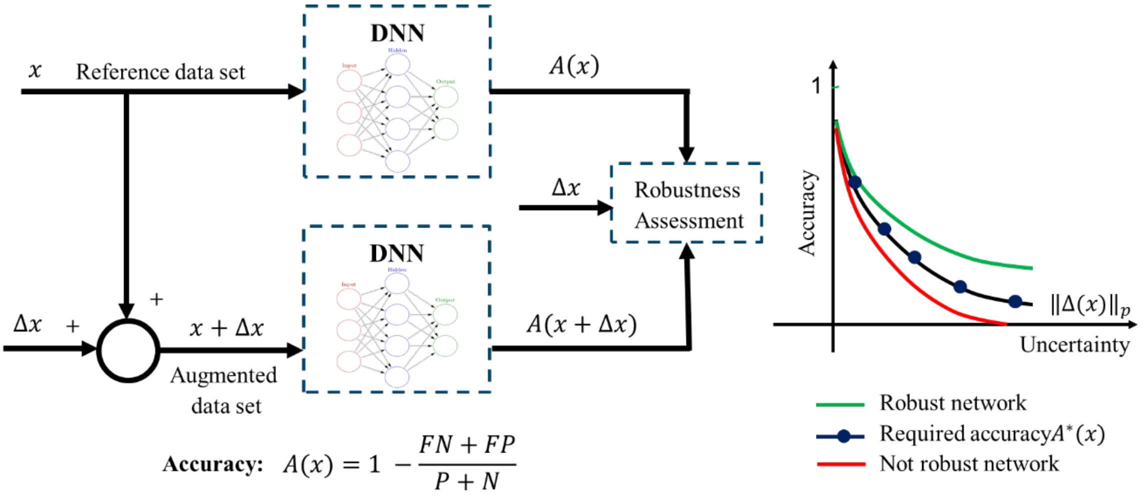


Figure 4: Methodology to assess the robustness of deep neural networks

To improve robustness, including robustness against adversarial attacks, during the training phase an inclusive dataset design, data normalization and data augmentation (e.g. flipping & rotating) are used. To assess efficiently the robustness of Deep Neural Networks (DNNs), data sets with variable uncertainty need to be efficiently generated and, metrics to measure the uncertainty and network performance shall be in place – the methodology is briefly illustrated in Figure 4.

As conclusion, defining and widely accepting a unified methodology to assess the robustness of deep neural network could significantly accelerate the development and certification of highly automated driving systems.

### 4.5 Talk #5: Data Quality Requirements for the Development and Validation of Automated Vehicles

Automated vehicles contain a large number of sensors and cameras monitoring the environment of the vehicle and providing the input data for decision processes of the control software (e.g. steering commands, speed control etc., Figure 5). The majority of the sensor and camera modules provide object detection and classification and rely on ML based techniques. Therefore, validation of these functionalities and evaluation of the underlying data quality are crucial aspects for safety and relevant for homologation processes in the automotive industry.

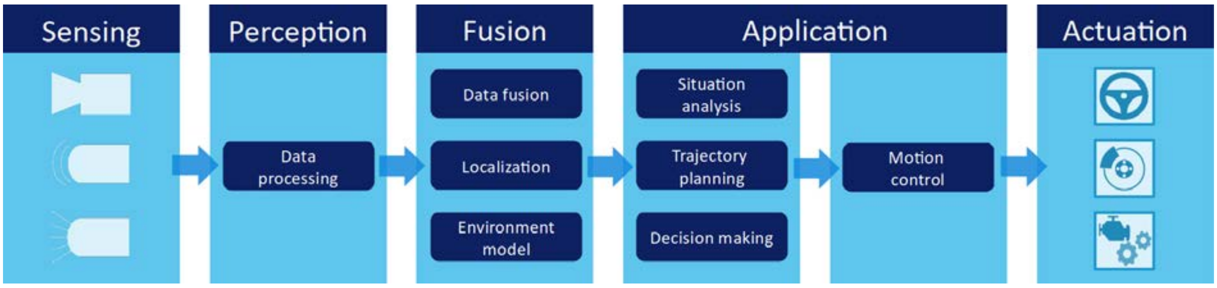


Figure 5: Basic principle of information processing in automated vehicles

The effort for validation and requirements to the used data are strongly depend on the autonomous level. The shift of liability from driver to the manufacturer (from level 3) demands a high effort which increases disproportionately up to the level 5. In principle, the quality of the validation data can be defined in three dimensions:

- **Volume:** The amount of data used for validation. E.g. the amount of driven kilometers within a variety of traffic situations and acquired objects (other cars, pedestrians, traffic signs etc.)
- **Accuracy:** The accuracy of the labels of the detected objects surrounding the vehicle. As the labeling process is based on the sensor and camera inputs of the vehicle, the accuracy of the labels is strongly depended on the setup of these systems.

Consequently, there will be always some ambiguity at the decision boundaries.

- **Right Data:** Selection of relevant objects and labels. Distinguish between relevant and not relevant objects for the current traffic scene. Typically, not so relevant objects are hard to label.

All these aspects are crucial for safety considerations and therefore subject of evaluation within homologation processes (Figure 6). To assess the safety aspects of automated vehicles, the above-mentioned validation data can be embedded in homologation frameworks, which comprise several steps and strategies to create suitable test scenarios for automated vehicles.

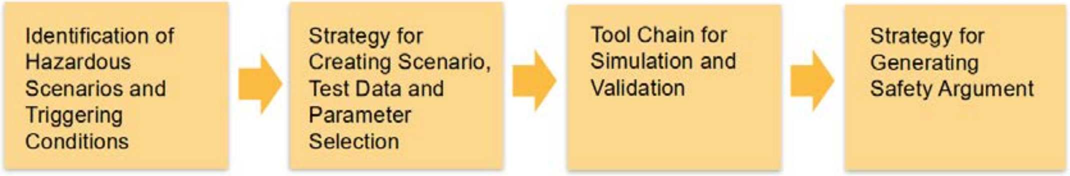
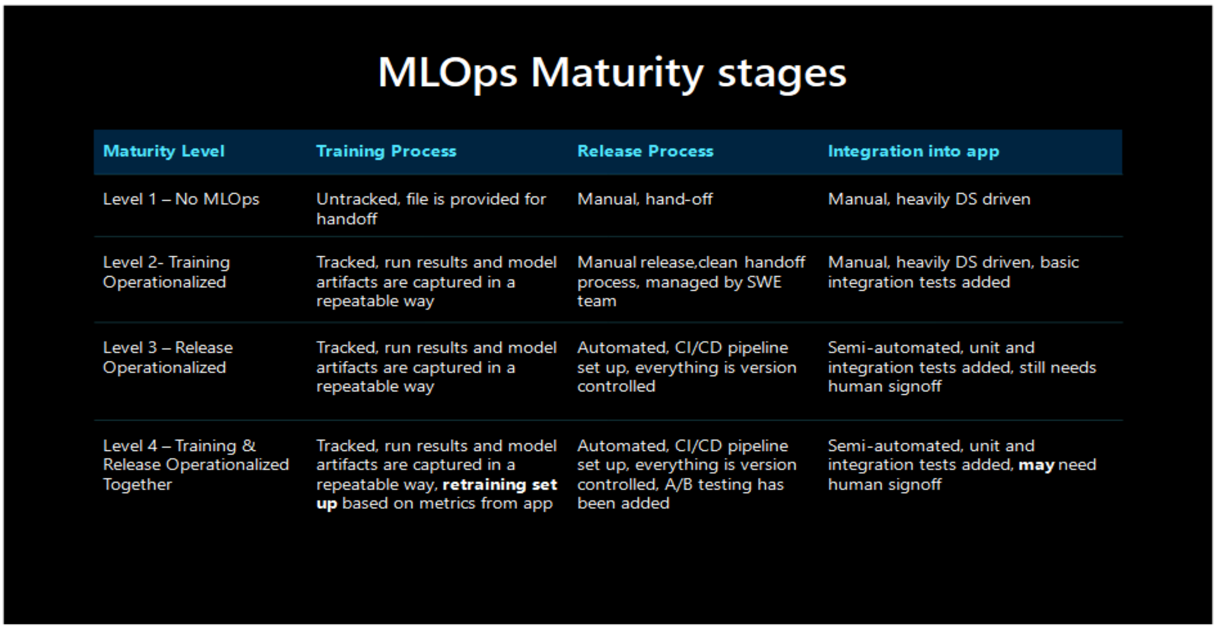

Identification of Hazardous Scenarios and Triggering Conditions
Strategy for Creating Scenario, Test Data and Parameter Selection
Tool Chain for Simulation and Validation
Strategy for Generating Safety Argument

Figure 6: Most important steps of the assessment within homologation processes

### 4.6 Talk #6: Challenges for AI Service Providers

Service providers for AI either develop or operate AI systems on behalf of their customers. This relationship is independent from the organizational affiliation of the service provider, i.e. an AI- department inside the same company providing AI-services to a market going business unit of the same corporation has to be treated the same as an external customer and hence, the same quality and documentation standards shall apply. Naturally and depending on the use-case, such services can be different, as can the underlying technology. Therefore, on top of mastering the technology, important quality criteria for an AI service provider are to properly manage the development, deployment and operations process of AI-systems, as well as the risks associated with it in a repeatable and replicable fashion.

Quality management and risk management are nothing new in software development and the associated best practices have been widely published and embraced. However, the quality of AI-systems is determined at least as much by the quality of the data used for training and validation as by the quality of the software per se. Hence, the challenge for ML or AI is how to fuse the well understood software-devops process, which manages the software lifecycle with the new data-lifecycle, which manages everything from data provenience, wrangling, labelling, re-purposing the data as well as meta-data but also possible re-use application specific data, e.g. weights of a trained model. Therefore, a new term has been minted – ML Ops for ML Ops processes, which are not limited to ML, but also expand to other domains of AI, such as neural networks. A proper ML Ops process does not only allow better reproducibility and auditability, but, if done right, also increases velocity of development/deployment and security of AI applications.



MLOps Maturity stages

| Maturity Level | Training Process | Release Process | Integration into app |
|---|---|---|---|
| Level 1 – No MLOps | Untracked, file is provided for handoff | Manual, hand-off | Manual, heavily DS driven |
| Level 2- Training Operationalized | Tracked, run results and model artifacts are captured in a repeatable way | Manual release,clean handoff process, managed by SWE team | Manual, heavily DS driven, basic integration tests added |
| Level 3 – Release Operationalized | Tracked, run results and model artifacts are captured in a repeatable way | Automated, CI/CD pipeline set up, everything is version controlled | Semi-automated, unit and integration tests added, still needs human signoff |
| Level 4 – Training & Release Operationalized Together | Tracked, run results and model artifacts are captured in a repeatable way, **retraining set up** based on metrics from app | Automated, CI/CD pipeline set up, everything is version controlled, A/B testing has been added | Semi-automated, unit and integration tests added, **may** need human signoff |

Figure 7: Microsoft MLOps maturity model

Microsoft responsible AI Principles embrace the management of audit trails (tracking of end to end lineage) & datasheets (document metadata) of a ML model along its entire lifecycle as a key component. It also introduces MLOps Maturity stages, which reflect an organizations ability to tightly and replicable manage AI-development and operation, including retraining and versioning/release management (Figure 7).

How does this approach stack up against the requirements of the proposed certification readiness matrix? Obviously, only a subset of categories can be addressed directly by an ML Ops process, namely the dimensions traceability and risk management. Others, e.g. Security, Robustness or Explainability can be indirectly addressed via specific parameters or acceptance criteria, though. In the current definition of MLOps maturity stages, mainly procedural parameters are tracked, i.e. if run results and model artifacts are captured and how they are tracked. The requirements to do this in a repeatable and standardized way and the appropriate metrics for model retraining are a precondition to reach maturity level on the ML Ops process. This implies that such performance indicators are there and can be quantified. Guidance, exactly which results and artifacts have to be documented and in which granularity this needs to happen still needs to be made explicit, bound to the industry and the severity of influence it can put a human's life. This type of severity can be standardized through an MLOps approach to make sure that local regulation can be audited in the correct way. Furthermore, an agreed upon taxonomy and ways to describe datasets needs to be used as the basis for documentation. The foundation and possible pathway is outlined in “Datasheets for Datasets', which now needs to be operationalized, refined and subsequently translated into generally accepted best practice [25].

### 4.7 Talk #7: Developing and Certifying Trustworthy and Reliable Telemedical Systems

AI has become a strong trend in IT in recent years, both at the customer and enterprise level. The AI sub-discipline ML in particular contributes to the lasting success of AI systems. With ML techniques and sufficiently large data sets, it is possible to find correlations and hidden patterns that humans would not be able to detect on their own. ML prediction models based on object recognition systems in the field of computer vision, for example, are already autonomously making decisions with, in cases like autonomous driving or medical decision support systems, far-reaching effects.

Compared to previous technologies, however, due to its complexity the decision-making process of AI, and specifically (deep) ML models, is often difficult or even impossible to comprehend.

In order to increase trust in AI systems, questions such as "can an AI system be responsible?", "how can or should AI developers take responsibility?" and, "what ethical principles should guide responsible AI?" need to be clarified. Answers to these questions could be derived from, for example, biomedical ethics, which provides a starting point for ethical AI with requirements like respect for autonomy, beneficence, nonmaleficence, justice, and "explicability". These principles, their implementation in the development process and the verification of the chosen answers must be worked on in a multidisciplinary manner. Experts from Computer Science, Engineering, Sociology, Philosophy, Psychology, and Law must be involved as well as those stakeholders who use and are affected by the AI systems to be developed.

TÜV certification plays a central role as a trustworthy assessment. In the context of the use case "Skin Lesion App", a mobile application for skin cancer detection based on ML technology and secure telemedicine platform for dermatology, the TÜV AUSTRIA Trusted AI Audit Catalog will be presented, which was developed together with Prof. Sepp Hochreiter from the Machine Learning Institute of the Johannes Kepler University Linz. The Trusted AI Audit Catalog and the audit process itself, are based on three pillars, which support the areas of "Secure Software Development", "Functional Requirements" for ML and "Ethics and Privacy". In addition, the catalog assesses the risk potential of an AI application and adapts the different sets of questions accordingly.

Currently, the catalog allows supervised ML application with low risk to be assessed and certified. For example, with the help of the catalog, critical issues of the "Skin Lesion App" related to privacy-related data and data bias were identified and highlighted at multiple levels in the audit catalog modules "Functional Requirements" and "Ethics and Privacy". As a result, the Trusted AI Audit Catalog for the certification of ML applications was able to demonstrate that ethics by design were consistently applied in the development of the "Skin Lesion App" and that it is responsibly embedded in the telemedical context.

### 4.8 Talk #8: Risk Classes as the Basis for AI Auditing

The AI Act proposal of April 2021 is a pioneering attempt of defining a legal framework for AI regulation. In its essence, it follows a risk-based approach with four AI risk classes ranging from minimal over low and high to unacceptable. The risk class assesses what level of risk an AI application poses to people, organizations or society as a whole. Depending on the risk classification, different regulatory requirements for AI operators apply.

It is assumed that most AI applications on the market will be classified as minimal risk. This space is intended to be entirely unregulated with no legally binding obligations for operators. An example application that falls into this category would be a common product recommender system as applied in various forms by many e-commerce platforms. Unacceptable risks applications are conflicting with European values and define the other extreme of the spectrum. The operation of these applications is prohibited entirely. An example for such an application would be a social scoring system that rates citizens based on their beliefs and actions in order to control their access to aspects of society.

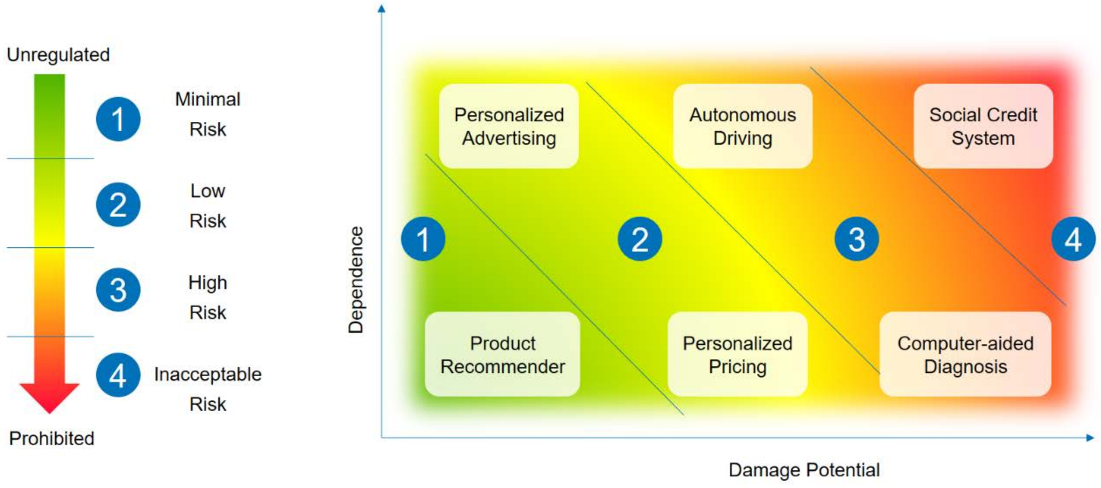


Figure 8: Two-dimensional AI risk space with subdivision in four risk classes and example applications for each class.

The AI Act proposal distinguished between two categories of AI systems, component and stand-alone. An AI component is defined as a part or module of a system that is already subject to existing regulations. These components are automatically classified as high risk. AI systems that are deployed outside of an existing regulatory framework are called stand-alone AI systems.

One question that arises might be the following. What is AI risk and how is the risk class of a given application determined? [26] proposes a two-dimensional scheme for assessing AI risk. One axis denotes the damage potential of an AI system, or how impactful bad AI decision making can be. The other axis denotes dependence, or to what extend a human depends or subordinates on AI decisions. Fig. 8 shows a visualization of that space along with some example applications.

For unacceptable risks, the definition in the AI Act is quite clear leaving little room for interpretation. The class of high risk systems is defined via a list of concrete applications in the appendix, where an actual risk has materialized in the past. The AI Act appears to follow a reactive approach for classifying a system as high risk rather than a proactive one. With every new application that poses a high risk, that list needs to be updated. On the other hand, applications can slightly be altered specifically for avoiding a high-risk classification, which may lead to a cat-and-mouse game between AI operators and legislation. To facilitate the operationalization of the AI Act, a clear methodology of set of rules for systematically determining the risk class of an AI application is required.

### 4.9 Talk #9: Standardized AI/ML Model Validation in Healthcare: The ITU/WHO Focus Group on "AI for health"

AI and ML offer a large potential for clinical and public health, with a wide range of use cases that can benefit from AI/ML-powered automation. Typical application examples include medical image classification, segmentation, or reconstruction, as well as systems for lab data analysis, decision support, or disease risk assessment. Prior to and during the usage of a given AI/ML tool in practice, it must be assured that the method and its implementation are safe and effective, and can be applied with good faith for the intended use. Forthcoming standards for health AI are expected to play a crucial role in facilitating the responsible use of the promising AI/ML technology. Standards will need to cover both qualitative as well as quantitative quality control procedures along the entire AI life cycle that can assure that the tools are accurate, robust, fair, and overall trustworthy. In 2018, the International Telecommunication Union (ITU) and the World Health Organization (WHO) have established a joint focus group on AI for health (FG-AI4H), as collaborative initiative to explore standardization needs and possibilities. FG-AI4H members and leadership include experts from different disciplines from around the globe that are engaging in a steady and well documented dialogue. Numerous topic groups tackle specific health use cases, e.g., to fight Tuberculosis, Malaria or cardiovascular diseases with the help of AI/ML, and aim at proposing standardized procedures to benchmark AI models for a given task, on undisclosed, reproducible test data, as technical validation step prior to the eventual clinical evaluation following best practices. The regulatory, ethical and technical perspectives are considered by several FG-AI4H working groups that are dedicated to overarching themes, which affect all topic groups. FG-AI4H establishes processes and related policies, and creates reference documents with best practices. The processes and documents are being mirrored in open-source reference implementations, that will support the entire assessment process, including, e.g., ground truth annotation tools and a benchmarking platform for health AI, which makes possible the auditing of AI technology taking both quantitative and qualitative criteria into account.

### 4.10 Talk #10: ML Based Biometrics: Challenges and Extension of Existing Audit Schemes as a Basis for Standardization

Biometric systems are increasingly used in a wide range of applications (Figure 9). Whether smartphones, banking applications or border control within airports, the use of unique features of individuals (e.g. face, fingerprint or retina) replaces classic authentication techniques like passwords and provides a higher usability. ML has become an indispensable part of Biometric systems extracting and recognizing the features of the raw image data.

Despite the relatively high fault tolerance and performance of ML techniques, every ML based biometric application suffers from specific vulnerabilities evoked from the robustness issues of ML. E.g. Adversarial Attacks can be applied to spoof the system. As biometric systems are commonly embedded in security critical areas and applications, the authentication process shall comprise a zero or relatively small False Acceptance Rate (FAR) to avoid unauthorized access to the system. Thus, current standards and schemes (e.g. Fido, NIST/ISO, Common Criteria) for biometric systems define certain countermeasures, FAR values and mitigations to keep the required security levels.

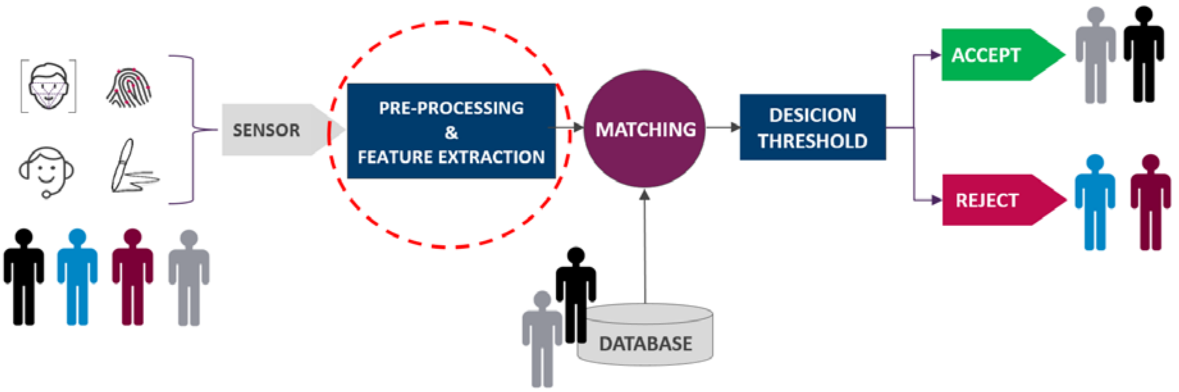


Figure 9: Basic functionality of a biometric system

Unfortunately, the ML specific issues are not completely in scope of the existing standards and schemes. Even if low FAR values are measured by performance testing and anti-spoofing techniques are implemented, the system can be still vulnerable against adversarial attacks or similar attack vectors targeting the underlying ML Model. To close this gap, current or future schemes shall extend or add requirements in order to address such ML specific vulnerabilities (Figure 10). Relevant aspects could be comprising the evaluation regarding adversarial robustness of the model and mitigation strategies against adversarial examples. Other aspects may cover requirements for "feature checks" and "model stealing". Feature checks address the need of checking the features taken into account for the classification process and conducting plausibility checks to avoid the influence of non-relevant features (e.g. background, noise etc.). In case of model stealing, an attacker might try to extract and re-build the used ML Model (victim) in order analyze the extracted model within an unprotected environment and eventually to craft attacks "white box" transferable to the victim model.

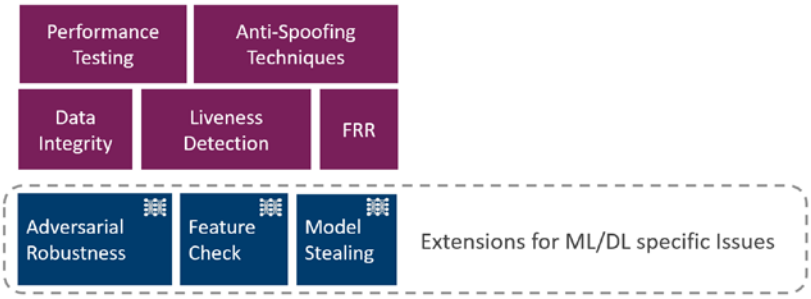


Figure 10: Possible extension of an existing scheme: ML specific issues.

## 5 Conclusion, Future Challenges and Next Steps

Due to the complex lifecycle of connectionist AI systems embedded in applications, physical systems and organizations, due to the multitude of interdependent aspects that need to be considered for security and safety relevant AI systems [2] and as a prerequisite for the operationalization of AI regulations such as the EU AI Act, requirements, audit methods and audit tools have to be available for practical use for all relevant aspects and for all relevant lifecycle phases of AI systems. As of now the requirements, audit methods and audit tools are not available to a sufficient degree but the development in the field is highly dynamic as illustrated by the strong increase in practically oriented approaches to auditing AI systems (cf. e.g. the approaches presented during this workshop). Strong efforts in research and development are required but these efforts have to be focused and prioritized. As a basis, the state of the art and the progress of auditability for relevant applications and for AI systems in general have to be periodically assessed and presented in a comprehensive way.

In this whitepaper we propose a “Certification Readiness Matrix” (Figure 1) as a tool to assess and present the state of the art in a comprehensive, comparable and dynamically traceable way. Based on the results from 10 talks throughout the workshop that each focused on a unique combination of aspects, life cycle phases and use cases (Figure 2), we aggregated the results and estimate that, despite substantial progress, a lot of work has yet to be done. This is especially the case for the technically challenging aspects IT-security, safety, robustness and interpretability while aspects such as traceability and risk management may be, to a large extent, dealt with using variations of existing methods and tools. In our view it is desirable to further develop the certification readiness matrix presented above: in order to make it an reliable basis for tracking developments and for prioritizing and coordinating future work in research and development, score requirements have to be developed and formulated in a comprehensive way for all cells of the matrix. Subsequently, the matrix should be applied to important use cases and updated over time until the AI Act may finally go into operation.

## 6 References

[1] Berghoff C, Neu M and von Twickel A: Vulnerabilities of Connectionist AI Applications: Evaluation and Defense. Front. Big Data 3:23, 2020.

[2] BSI, TÜV-Verband and Fraunhofer HHI: Towards Auditable AI Systems - Current status and future directions, Whitepaper, 2021, https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Towards_Auditable_AI_Systems.pdf (Accessed May 18, 2022).

[3] European Commission: Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), 2021, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (Accessed May 18, 2022).

[4] DIN SPEC 92001-1 Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 1: Quality Metamodel, 2019.

[5] ISO/IEC 22989 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology (under development as of May 18, 2022).

[6] ISO/IEC CD 5338 Information technology — Artificial intelligence — AI system life cycle processes (under development as of May 18, 2022).

[7] ISO/IEC DIS 25059 Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems (under development as of May 18, 2022).

[8] ENISA: Securing Machine Learning Algorithms, 2021, https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms (Accessed May 18, 2022).

[9] ETSI GR SAI 002 V1.1.1 (2021-08) Securing Artificial Intelligence (SAI); Data Supply Chain Security, http://www.etsi.org/deliver/etsi_gr/SAI/001_099/002/01.01.01_60/gr_SAI002v010101p.pdf (Accessed May 18, 2022).

[10] ETSI GR SAI 004 V1.1.1 (2020-12) Securing Artificial Intelligence (SAI); Problem Statement, http://www.etsi.org/deliver/etsi_gr/SAI/001_099/004/01.01.01_60/gr_SAI004v010101p.pdf (Accessed May 18, 2022).

[11] ETSI GR SAI 005 V1.1.1 (2021-03) Securing Artificial Intelligence (SAI); Mitigation Strategy Report,

http://www.etsi.org/deliver/etsi_gr/SAI/001_099/005/01.01.01_60/gr_SAI005v010101p.pdf (Accessed May 18, 2022).

[12] NIST: AI Risk Management Framework: Initial Draft, 2022, https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf (Accessed May 18, 2022).

[13] Stanton, B. and Jensen, T.: Trust and Artificial Intelligence, NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology, Gaithersburg, MD, 2021, https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=931087 (Accessed May 18, 2022).

[14] Felix Sattler, Klaus-Robert Müller and Wojciech Samek: Clustered Federated Learning: Model-Agnostic Distributed Multi-Task Optimization under Privacy Constraints, IEEE Transactions on Neural Networks and Learning Systems, 32(8):3710-3722, 2021.

[15] Leon Witt, Usama Zafar, KuoYeh Shen, Felix Sattler, Dan Li and Wojciech Samek: Reward-Based 1-bit Compressed Federated Distillation on Blockchain, arXiv:2106.14265, 2021.

[16] Felix Sattler, Arturo Marban, Roman Rischke and Wojciech Samek: CFD: Communication-Efficient Federated Distillation via Soft-Label Quantization and Delta Coding, IEEE Transactions on Network Science and Engineering, 2021.

[17] Lechner, M., Hasani, R., Amini, A. et al.: Neural circuit policies enabling auditable autonomy. Nat Mach Intell 2, 642–652 (2020).

[18] SAE J3016B: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles’. 2018.

[19] ENISA: Artificial Intelligence Cybersecurity Challenges - Threat Landscape for Artificial Intelligence’, 2020, https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges/@@download/fullReport (Accessed May 18, 2022).

[20] Dede, G., Hamon, R., Junklewitz, H., Naydenov, R., Malatras, A. and Sanchez Martin, J.I., Cybersecurity challenges in the uptake of Artificial Intelligence in Autonomous Driving, EUR 30568 EN, Publications Office of the European Union, Luxembourg, 2021.

[21] ISO 26262 Road vehicles – functional safety, edition 2018.

[22] ISO/DIS 21448: Road vehicles – Safety of the intended functionality, edition 2021.

[23] ISO/TR 4804: Road vehicles - Safety and cybersecurity for automated driving systems - design, verification and validation, first edition, 2020-12.

[24] ISO/DIS 22737: Intelligent transport systems – Low-speed automated driving (LSAD)

systems for predefined routes – Performance requirements, system requirements and performance test procedures, edition 2020.

[25] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé and Kate Crawford: Datasheets for Datasets, arXiv:1803.09010v3, 2018, https://www.microsoft.com/en-us/research/uploads/prod/2019/01/1803.09010.pdf (Accessed May 18, 2022).

[26] Tobias D. Krafft and Katharina A. Zweig: Transparenz und Nachvollziehbarkeit algorithmenbasierter Entscheidungsprozesse - Ein Regulierungsvorschlag aus sozioinformatischer Perspektive, Verbraucherzentrale, 2019, https://www.vzbv.de/sites/default/files/downloads/2019/05/02/19-01-22_zweig_krafft_transparenz_adm-neu.pdf (Accessed May 18, 2022).

## Published by

**Bundesamt für Sicherheit in der Informationstechnik**
Godesberger Allee 185-189
53175 Bonn
Deutschland

**Fraunhofer-Institut für Nachrichtentechnik**
Heinrich-Hertz-Institut
Einsteinufer 37
10587 Berlin
Deutschland

**TÜV-Verband e. V.**
Friedrichstraße 136
10117 Berlin
Deutschland