role in assessing the dependence on the decision are control, selection and correction [123].

→ Decisions and actions of an AI system that are additionally filtered by human interaction (e.g. the purchase of recommended items in an online store) imply a lower need for regulation than machines that act without human intermediaries (e.g. the emergency shutdown of a nuclear power plant). This aspect is summarized under control.
→ The ability to exchange the AI system for another one (e.g. by changing a provider) or to avoid being exposed to an algorithmic decision at all is called **selection**. A one-sided dependency relationship between producers or operators and users, as well as monopolistic structures lead to dependence on one or a few systems. In the worst case, the user does not have the possibility to turn away from using certain services without being confronted with personal or social consequences (e.g. lack of access to health care, financial market).
→ The importance of the possibility to challenge or have corrected an automatically generated decision, as well as the time needed for an adequate follow-up of the relevant application should not be underestimated. This is summarized by the term **correction**. Machine decisions that cannot be challenged at all increase the dependency on the decision. Repairing significant individual harm requires more time and effort than many cases with less harm. This aspect concerns the compensation for damage/liability, which is addressed in the dependence on the decision (y-axis).

4.2.2.4.4 Risk classes

For systems that fall into Class 1, no transparency obligations would be required and no control processes would be permanently installed. In cases of doubt, a post-hoc analysis could be used to check for relevant damage. If the suspicion is confirmed, a new evaluation into a higher class would be conceivable.

In Class 2, the first transparency obligations would be required. To enable a "black-box analysis" [139], an appropriate interface must be provided for the system so that a controlling instance can check the input-output behaviour of the system. A description of how the system is embedded in the social decision-making process would also be necessary.

For systems in class 3, the input data should be described completely to a controlling instance. The stated quality (in the sense of numerical values describing the quality) of the decision system should be verifiable.

In class 4, all information about and decisions made by the software must be traceable and verifiable within a reasonable time, at least for a controlling instance. The demand for traceability generally excludes many learning processes (e.g. artificial neural networks), since they cannot fulfil this demand at the current state of research. All necessary interfaces would have to be provided.

Systems in class 5 should not be implemented. This class is justified by systems that are not compatible with the principles of democracy, such as evaluation systems based on continuous monitoring of the population, systems that override the presumption of innocence, or systems that have an approvingly lethal effect without human influence. Furthermore, systems that exceed a certain potential for harm and can only be implemented with a high error rate due to the difficult data situation (e.g. incomplete or faulty) would be in this class (e.g. identification systems for terrorists). This class does not exclude statistical methods that search for patterns in large amounts of data, but the finding of such patterns should not lead to unreflected decisions.

4.2.3 Standardization needs

NEED 1:
Design initial criticality checks of AI systems quickly and easily

Unintended ethical problems and conflicts occur primarily in ADM systems with learning components that make decisions about people, their belongings or access to scarce resources, and have the potential to damage individual basic rights and/or basic democratic values. An initial criticality check as to whether a system can trigger such conflicts at all or whether it is an application far removed from any ethical issue must be made quick and easy by standardization. This horizontal, for all areas low-threshold check must quickly and legally clarify whether the system must meet transparency and traceability requirements at all. Especially with regard to the wide fields of application of artificial intelligence, such a risk-based criticality check in critical areas offers the possibility to make adequate demands and at the same time to counter the accusation of "ethical red taping" by developing completely uncritical fields of application free of additional requirements.

NEED 2:
Operationalization of ethical values
It is currently unclear how organizations that develop and use AI systems can measure and operationalize abstract ethical values. There are a number of promising approaches that have the potential to meet the challenge (such as the WKIO model), but the practical application of such approaches is still in its infancy. Open questions, problems and challenges can currently only be addressed to a limited extent, which is why standards offer the opportunity to transfer theoretical concepts for the operationalization of ethics into practice, to accompany them and to shape them consensually in dialogue with companies.

NEED 3:
Standardization of a concept for privacy ethical design
The principle of privacy protection is an expression of human dignity, autonomy and individual freedom, and an essential criterion for the acceptance of new systems. For this reason, standardization should promote the design of technology which safeguards the personal interests of users and affected parties in the sense of a "privacy ethical design". This should take up and shape the previous approaches from the fields of medicine and occupational safety in a cross-divisional concept. This can be done within the framework of the project currently initiated in ISO/IEC JTC 1/SC 42 on an MSS for AI (4.1.3, Need 1 "Support for international standardization work on an MSS for AI") by including the explainability of AI systems in the catalogue of requirements of the resulting document, and by extending the concept of risk to include ethical risks, as already done in the ISO/IEC 23894 Risk Management project.

NEED 4:
Design of the value system
Intelligent decisions based on general ethical principles require an examination of ethical values. If the machine knows the relation of meaning of different values and objects by means of an ontology, this is helpful. Autonomous systems must also be able to process unplanned situations. If, for example, the internal representation of objects is enriched by knowledge from an ontology when autonomous machines recognize the environment, this is a possibility to make a value system accessible to the machines. Ontologies allow the machine to create contexts without having to specify case patterns beforehand (as in W3C [140]). The research and subsequent standardization of the interfaces of ontologies to consider ethical principles in concrete scenarios promises to meet the potential of the challenge.

NEED 5:
Design earmarking of data
Standardization should further shape the existing earmarking of data. This can ensure that there is documentation of the purpose for which the data was collected and can allow regulation of the conditions under which the data may or may not be used for other purposes.

NEED 6:
Design interfaces for the AI development process
The long development process of AI systems should be shaped by standardized interfaces. Here standardization can make an important contribution. These interfaces could include, for example, access to relevant training data sets and models of an AI system as a basis for external review. Primarily International Standards would promote the interchangeability of components and provide access for verifiers, and would ensure that requirements are met without much effort, thereby increasing confidence in the system.

NEED 7:
Include quality backward chain in the AI life cycle
It is recommended to include a quality backward chain with field data collection in the AI life cycle to identify and correct unethical behaviour during the application (see 4.3.2.3.2.4 Process checks: Quality assurance after delivery by product monitoring).

NEED 8:
Design re-evaluation of AI systems
AI systems should be widely used in a complex social context. A systematic process of ethical reflection and participation should therefore be initiated in AI development. Depending on the complexity of the AI and potential risks, several evaluation steps and a continuous involvement of interested parties, as well as ethics experts and ethically trained staff are recommended.

4.3 Quality, conformity assessment and certification

AI is increasingly being applied in different areas of everyday life (see also the chapters 4.5 to 4.7). Based on the assumption that AI can only unfold its full application potential if it is used according to high quality criteria, the following chapter deals with the resulting need for standardization with regard to quality criteria and their verification by a corresponding conformity assessment (based on the ISO/IEC 17000 series of standards [38]–[44]). A number of ideas discussed in this chapter can also be found in the Impulspapier and White Paper "Certification of AI Systems" issued by the Platform Learning Systems.

When testing AI systems, two levels can be distinguished (see Figure 18): On the one hand, assured properties of an AI system can be confirmed by a technical test. For example, the accuracy of a classification can be determined by precision and recall (technical level of testing). The second level is the evaluation level, which checks whether a system is suitable for a certain application (is the tested accuracy sufficient for the application?) or whether it meets certain ethical, legal or social requirements. A seal of approval [123] has been proposed for ethical considerations, which represents an interesting approach for the ethical evaluation of AI systems and is based on a value analysis procedure using a combination of target criteria, indicators and measurable variables. All tests of the second type should always be based on technical tests. It is to be expected that standards and specifications can be formulated primarily for the first level of testing, but that questions of the second level of testing are often the subject of regulation or social discourse.

Such conformity assessments can be carried out by the manufacturer itself, the buyer or an accredited third party body. In the course of conformity assessment, products, systems and processes may be subject to testing, calibration, validation, verification and certification or inspection. In certain areas (such as in accordance with the EU Medical Devices Regulation [141]), certification by a Notified Body is even mandatory prior to placing the product on the market.

Certification is carried out within the framework of conformity assessment by a third party according to the applied conformity assessment programme.

According to international expert commissions, such as AI HLEG, proof of conformity for AI products and processes is based on the following normative, legal and technical quality criteria (cf. also IAIS White Paper) [45].

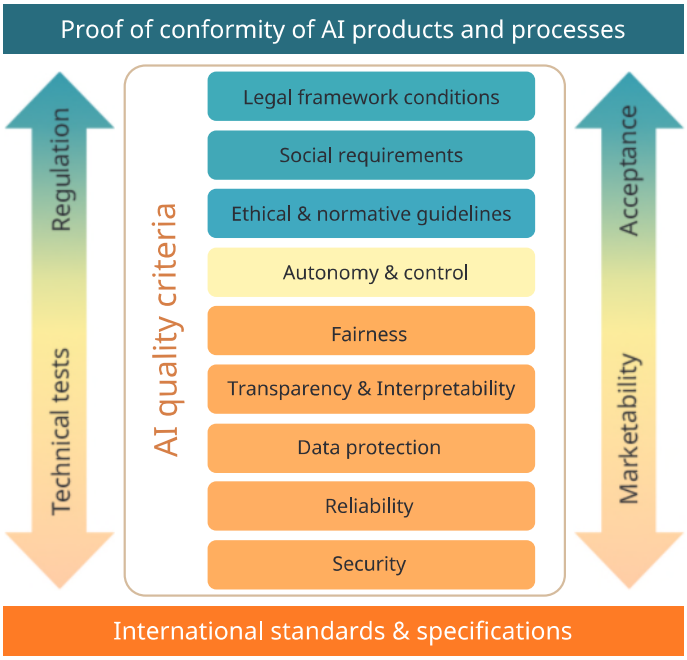

Figure 18: Classification of the categories of AI quality criteria in conformity assessment [45]

Law, society and ethics

AI applications have a disruptive potential. Conformity with social, ethical and legal frameworks mainly serves the protection of legal or ethical fundamental interests of persons (4.2.2.3). The AI conformity checks in these categories are intended to prevent and help to avoid impairments of groups and individuals, injustice or ethically unjustified conditions of society.

Autonomy and control

AI applications increasingly work autonomously, i.e. they pursue a given goal while freely choosing the means to achieve it. The AI system is free to choose the means, but not the actual objective. In this context one speaks misleadingly of the "autonomy of action" of the system, although the objective is not changed. This analogy gives rise to an area of conflict regarding the autonomy of humans, since such AI applications can in turn influence humans in their choice of goals and means. AI conformity tests must be able to make statements about autonomy and control at the interface to the technical AI system if the AI application interacts with human decision-making, for example, by generating decision proposals, generating and possibly executing control commands, communicating with humans or being integrated into work processes.

The following quality categories are part of the technical testing of AI systems.

Fairness and non-discrimination
AI applications learn from historical data, which is not necessarily unprejudiced. In order to avoid unjustified unequal treatment in an AI application and to exclude undue discrimination, AI applications must be verifiable to ensure that individuals are not discriminated against in social outcomes because they belong to a marginalized or discriminated group (see 4.2.2.2).

Transparency and interpretability
The transparency of an AI application can significantly contribute to its acceptance. For this purpose, information on the correct use of the AI application must be available. Essentially, requirements for interpretability, traceability and reproducibility of results must be checked, requiring insights into the inner processes of the AI application. There is still a considerable need for research into the colloquially associated demand for the explainability of an AI application, even if the explainability of the effects of AI-specific technological features is limited.

Data protection
The technical examination of the data protection regulations, in particular the GDPR [95], the BDSG [142] and the requirements of the Hambach Declaration [143], must be observed for AI conformity tests.

Reliability
From a technical point of view, testing the reliability of an AI system includes requirements for correctness, traceability, assessment of the uncertainty of results, and of the robustness against attacks, errors, and unexpected situations and thus overlaps with the concept of security in the narrower sense. Tests of the reliability and security of AI applications are essential basic requirements to make statements about their trustworthiness.

Security/safety
The security of AI applications includes security against threats and attacks and functional safety in the broadest sense. The security/safety of AI systems is discussed in detail in Chapter 4.4. Reliability, data protection and data security are also taken into account. In terms of testing methods, it should be noted that the technical test bases for AI systems must be developed and related to existing test procedures.

4.3.1 Status quo

In the following the essential terms of objects and activities of conformity assessment are listed.

4.3.1.1 Conformity assessment

Demonstration that specified requirements are met (ISO/IEC 17000 [38]). Defined requirements (i.e. needs or expectations) can be detailed (e.g. concrete technical specifications) or general (e.g. safe, robust, transparent, fair).

To differentiate the objects of conformity assessment:
1. Product (e.g. hardware, software)
2. Process
3. System
4. Service
5. Management system:
6. Person
7. Information (e.g. declarations, assertions, predictions)

Objects of a conformity assessment can also be combinations of these individual objects (e.g. development process + product, product + service, system + assertion). The specified requirements must be clearly assigned to the object (e.g. technical specification for the hardware, fairness criteria for the process, robustness of a system, competence requirements for a person, plausibility conditions for an assertion).

To differentiate the activities:
By clearly assigning the specified requirements to defined objects (see above), the activities for “selection” and “determination” (see process of conformity assessment) can be determined. Their results may be sufficient for the given situation (e.g. for analysis or characterization) or may subsequently be subject to “assessment” with a view to a “decision” on conformity of the object.

4.3.1.1.1 Types of conformity assessment

In the following, the types of conformity assessment (see Figure 19) are described.

Testing

Determination of one or more characteristics of an object of conformity assessment by a procedure. The procedure may be intended to control variables within the test as a contribution to the accuracy or reliability of the results. The results of the test can be presented in the form of specified units or objective comparisons with agreed references. The result of the test may include comments (e.g. opinions and interpretations) on the test results and compliance with the specified requirements.

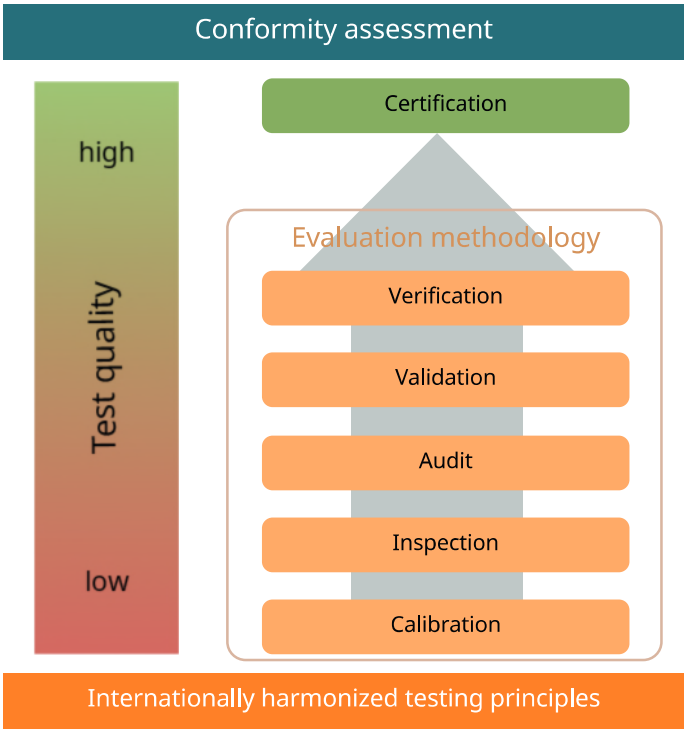

Figure 19: Evaluation methodology and test quality

Calibration

Activity which, under specified conditions, in a first step establishes a relationship between the quantity values provided by standards with their measurement uncertainties and the corresponding displays with their associated measurement uncertainties, and in a second step uses this information to establish a relationship with the aid of which a measurement result is obtained from a display.

The result of a calibration can be expressed in the form of a specification, a calibration function, a calibration diagram, a calibration curve or a calibration table. In some cases it can consist of an additive or multiplicative correction of the display with the assigned measurement uncertainty. Calibration should not be confused with adjustment of a measuring system, which is often wrongly called "self-calibration", nor with verification of the calibration. Often only the first step in this definition is considered as calibration [144].

Inspection

Examination of an object of conformity assessment and determination of its conformity with detailed requirements or, on the basis of expert assessment, with general requirements. An examination may include direct or indirect observations, which may involve measurements or reading of measuring instruments. Inspections can be limited to examinations in conformity assessment programs or contracts.

Audit

Check that an organization's processes, practices and procedures meet certain requirements formulated in a standard (e.g. an MSS, see 4.1.2.2.3). This check is usually based on a list of criteria derived from the underlying standard, which describes how requirements are checked. Audits include the inspection of documentation provided by the organization to be audited, interviews by the auditor, but also on-site inspections.

ISO differentiates between three levels of the audit:

→ Audit by the organization to which the audit refers (self-disclosure);
→ Audit by a customer, supplier or partner of the organization to be audited;
→ Audit by an independent third party. Such an audit can lead to certification.

ISO 19011 [145] provides guidelines for audit planning, audit execution and audit follow-up.

Validation

Confirmation of the plausibility of a specific use or application purpose by providing objective evidence that specified requirements have been met. Validation can be applied to assertions to confirm the information provided by an assertion in relation to its intended future use.

Verification

Confirmation of truthfulness by providing objective evidence that specified requirements have been met. Verification can

be applied to assertions in order to confirm the information provided by an assertion that relates to events that have already occurred or that relates to results that have already been obtained.

Certification

Confirmation by a third party relating to an object of conformity assessment (accreditation excluded). A "third party" is independent of the supplier of the object of the conformity assessment activity and has no interest as a user. Testing, inspection and validation/verification activities may also be performed by the supplier (first party) of the object to be evaluated or by a person/organization with an interest as a user of that object (second party). Certifications are only offered by independent bodies.

4.3.1.1.2 Conformity assessment process

Conformity assessment is divided into five phases:

- → **Selection** = Selection of applicable requirements, choice of methods, planning, sampling
- → **Determination** = Activities to collect evidence of conformity with regard to the specified requirements, i.e. analyses, tests, evaluations, investigations, audits, tests, inspections, validations, verifications, etc.
- → **Review** = Conclusion regarding suitability, adequacy and the sufficient amount of evidence collected
- → **Decision** = Deciding whether or not the assessed object has been shown to conform to the specified requirements
- → **Attestation** = Formal issue of the statement of conformity, e.g. test report (test passed/failed) or certificates

4.3.1.1.3 Types of conformity assessment bodies:

Depending on the type of conformity assessment, the ISO/IEC 17000 series distinguishes between different types of assessment bodies which, according to the activities listed above, inspect, analyze, test or measure product safety and quality and objects of protection:

- → **Testing laboratory** (ISO/IEC 17025 [42])
- → **Inspection body** (ISO/IEC 17020 [39])
- → **Validation/Verification body** (ISO/IEC 17029 [43])
- → **Certification body** (ISO/IEC 17021-1 [40] for management systems, ISO/IEC 17024 [41] for persons and ISO/IEC 17065 [44] for products, processes and services)

4.3.1.2 Existing standards and specifications from other areas with relevance for AI quality and conformity assessment

AI applications are usually implemented as components of larger IT systems. These AI components can be realized by a variety of different technologies. These AI applications are used in many industrial and everyday applications, where the actual AI component often interacts with other software, information technology, mechanical and electronic modules of the overall system.

The first step in standardization is therefore to identify existing standards and specifications that are relevant to the quality (and checking) of these systems. Standards from the areas of software (AI component), IT security (overall IT system), data quality and functional safety (application context) are particularly worthy of consideration.

Table 13 in Chapter 6.4 shows national and global standardization committees. Working groups relevant for quality, conformity assessment and certification are marked in the column "Relevance for quality, conformity assessment and certification (4.3)".

In principle, any standard that formulates requirements for a software application is also relevant for AI components as a special software component, regardless of the technology used. It must first be checked which standards already sufficiently cover the AI-specific properties and whether additions or changes are necessary.

Some prominent examples from the fields of software development and functional safety are listed below. However, this list makes no claim to completeness. In addition, there are very many relevant standards on IT security which are discussed in detail in Chapter 4.4. In addition, standards focusing on AI are being revised in this and other areas to address AI-relevant aspects. Table 11 in Chapter 6.2 gives an overview of standards and specifications in different thematic areas that do not yet provide detailed information on the application of AI components. The standards that formulate relevant requirements and quality criteria for software are marked in the column "Relevance for quality, conformity assessment and certification (4.3)".

4.3.1.2.1 Software development

AI processes can be integrated into existing software development standards such as ISO/IEC/IEEE 12207 [58] (Software life cycle processes), ISO/IEC 27034 [71]–[78] (Application Security) and ISO/IEC 25010 [146] (System and software quality models). For example, a test of trained AI-based software systems for "functional safety, efficiency, transferability, maintainability and reliability" can be carried out according to ISO/IEC 25010 [147].

4.3.1.2.2 Functional safety

The IEC 61508 [79]–[86] series of standards defines requirements for the various life cycle phases of electrical, electronic and programmable electronic (E/E/PE) systems that perform safety-related functions. IEC 61508-3 places a special focus on the requirements for the development of safety-relevant software. This also includes requirements for the tools used in the development process. Four safety integrity levels (SIL) are defined as a measure for the necessary risk-reducing effectiveness of safety functions and the resulting requirements on the safety-relevant system. Until now, the use of AI functionality is not recommended but also not excluded by IEC 61508. The responsible committee IEC/SC 65A, however, is considering the topic of AI for an update of IEC 61508 and is working together with ISO/IEC JTC1 SC42. There, a technical report on functional safety and AI systems is under development. IEC 61508 has a broad acceptance and application in industry and is the basis for several application-specific standards, e.g. for the process industry, mechanical engineering, control technology in nuclear power plants and railway signalling technology. The specification ISO/PAS 21448 [148] describes the safety of the target function and also includes performance restrictions that have their origin in environmental influences or communication. The standard ISO 12100 [124], [125] defines general principles and methods of machine safety as a basic safety standard, but is not a functional safety standard in the narrower sense.

4.3.1.2.3 Data quality

Since the quality of an AI component is closely linked to data quality, standards on data quality and big data are also listed in Table 11. DIN ISO/IEC 25012 [88] introduces a model of data quality. ISO/IEC 20546 [34] and ISO/IEC TR 20547-2 [149] and -5 [150] deal with big data, its terminology and reference architectures.

4.3.1.3 Existing standards and specifications on AI quality and conformity assessment

Table 10 in Chapter 6.1 lists existing standards and specifications that deal explicitly with AI applications. The standards that formulate relevant requirements are marked in the column "Relevance for quality, conformity assessment and certification (4.3)". This list is not exhaustive, but from today's perspective it represents the majority of the relevant standards and specifications.

In Germany, DIN has published two DIN SPECS in which a quality meta model for AI (DIN SPEC 92001 [87]) and a guide for deep learning image recognition systems (DIN SPEC 13266 [151]) are presented. At European level, ETSI addresses artificial intelligence in technical specifications relating to emotion recognition (ETSI TS 103 296 [152]) and autonomous networks (ETSI TS 103 195-2 [153]). At international level, the ITU-T focuses, within the published standards on requirements (Y.3170 [154]) and AI capabilities (Y.3173 [155]) with regard to AI in future networks. Within published documents, the consortia IEEE and UL deal with the assessment of autonomous systems (IEEE 7010-2020 [156] and UL 4600 [157]).

Apart from DIN SPEC 92001-1 [87] all specifications mentioned and listed in the table deal with AI components related to a concrete application. Work on a number of standards dealing with the quality of AI systems is in progress at international level, for example in ISO/IEC JTC 1/SC 42. In addition, IEEE standards are also in preparation or available, as are DIN SPEC 92001-1 and SPEC 92001-2. In further standardization activities, the quality criteria mentioned there would have to be compared with the quality criteria of ISO/IEC 25010 [146] and the AI-specific requirements would have to be highlighted.

4.3.1.4 Standardization activities with relevance for AI quality and conformity assessment

Table 12 in Chapter 6.3 lists standardization activities relevant to AI quality and conformity assessment in the column "Relevance for quality, conformity assessment and certification (4.3)". This list is not exhaustive, but from today's perspective it represents the majority of the relevant standardization projects.

Currently, numerous activities for AI standardization are taking place on all levels of standardization. Especially the work in ISO/IEC NP 5059 is to be emphasized, because here work is done on quality requirements for AI following the software quality requirements of ISO/IEC 25010 [146].

4.3.2 Requirements, challenges

4.3.2.1 Need for testing and marketability

In April 2019 the AI HLEG published ethical, legal and technical key requirements for trustworthy AI-based systems [22]. In most cases, these are hybrid applications, i.e. they consist of AI components and non AI-based software and hardware, and are basically understood as special IT. The user industry in Europe expects the market-driven development of criteria and methods for the technical testing of I systems. There follows a discussion of the scope of such testing.

4.3.2.2 Scope of a test

In this chapter we discuss which aspects should be considered in the context of a test of an AI system. This includes the components of an AI system, as well as AI-specific challenges that arise when testing these systems.

Further quality requirements result from the fact that AI systems are often also components of a larger product (e.g. the Platform Economy) for whose interoperability standards must also be set to ensure additional connectivity and interchangeability in the end product. Without such guarantees, global interaction is almost impossible, and this ultimately also prevents the scalability of solutions.

With regard to the aspect of the examination of quality criteria there is partly a great affinity to test procedures of functional safety, software development and IT security, which can be attributed to the fact that AI applications are hybrid IT systems. The marketability of a potential test method that addresses the above-mentioned aspects therefore requires an integrated approach that extends existing test methods to AI-specific criteria. There follows an analysis of the components of an AI system that require consideration when testing. In addition, the AI-specific challenges that need to be addressed to close the gap illustrated above are described in the following.

4.3.2.2.1 Components of an AI system

Components of an AI system include algorithms, databases and interfaces to the overall system. In principle, AI-based system components are based on symbolic and sub-symbolic methods of artificial intelligence. These include techniques for decision-making (e.g. decision-theoretical expert systems), knowledge representation (e.g. ontologies and knowledge graphs), methods for applying knowledge (e.g. logical reasoning and probabilistic methods) and machine learning methods (e.g. supervised learning and unsupervised learning). A detailed description of the classification of AI components can be found in 4.1.2.

Here the methods of artificial intelligence in an AI application can be realized by software. Depending on the capability spectrum of an AI application, hybrid methods (e.g. hybrid neural network models) can also be used in which symbolic and sub-symbolic techniques are combined. In AI applications there is an adaptability (dynamics) of the partial components of methods of artificial intelligence. For example, in machine learning processes, activation, transfer and summation functions determine the dynamics of a neural network [158]. On the other hand, a dynamic can manifest itself in the changeability of knowledge through AI methods, for example based on the AGM theory [159], [160].

Regardless of the actual realization of the AI application, a quality assessment should include the following aspects:

→ The quality of the data used: This includes, among other things, a possible bias in the data, which can negatively affect the fairness of the overall system, and the integrity of the data, since these significantly determine the behaviour of an AI component and thus make it necessary to secure the training data sets against indirect attacks through their manipulation. This applies in particular to continuously learning (self-learning) systems that are further trained in the field and whose input data for continuous learning is not under the direct control of the manufacturer. Therefore, quality assurance of the data supply chain itself is also necessary, as it plays an essential role with regard to the quality aspects of the data. Also, the data used for the training of a model and its distribution (e.g. image resolution, statistical distribution) must correspond to the operational environment.
Synthetically generated data are increasingly being used in the development of AI systems. Here, an artificial representation of an original data set is created, which has the most important statistical properties of the original

data set. Such synthetically generated data sets are especially helpful if either the amount of original data is too small (an example is the training of ML models for autonomous driving) or if the original data contains sensitive personal characteristics. The quality of such synthetically generated data is measurable and should meet the same quality requirements as real data sets.

→ The selection of the method/algorithms, their hyperparameters and the evaluation of a learned model. In general, empirical methods of testing as well as verification methods for quality assurance of a trained system are suitable here. Both are subject to the challenges described in the following chapter. To check the quality, it is necessary to consider alternative hyperparameters and their influence on the quality. For parameter selection and other areas of engineering, methods of automated machine learning under the term AutoML are in development and first use, among others as a service.

→ Also, the assessment of the overall IT system in which the AI component is embedded. This results in particular in interfaces to other technical IT environments such as cloud architectures, server farms, data repositories and data supply chains, statistical analysis packages, etc.

→ The man-versus-machine interface. Here, human and machine factors need to be considered. But machine-versus-machine and AI-system-versus-AI-system require validation. The interface can be facilitated by the AI system giving the human feedback explaining what it has "understood".

→ The behaviour of the AI application after delivery during its use in the operational environment (product observation), until the end of its life cycle (see 4.3.2.3.2.1 and 4.3.2.3.2.4)

4.3.2.2.2 AI-specific challenges

In contrast to conventional IT systems, AI applications have some special features for which quality criteria and test methods must be established and which pose substantial challenges for existing and future test methods. This includes:

→ A correctness term for KI systems: Rule-base algorithms have a clear source code that can be tested using classical methods. Examples of suitable verification methods are classical proving and proof assistants. Certain definite parameters can also be appropriately tested. With learning systems, not only the software architecture (e.g. NN model selection) and the source code quality are involved, but also what has been learned. Also, in contrast to classical systems, AI-based systems often work statistically and will therefore not achieve an accuracy of 100 % of the specified behaviour. Therefore, a test of AI-based systems must define a sufficient requirement for accuracy and aim to argue for this requirement on the one hand and for the remaining cases to secure the system by further measures, e.g. safeguards. It remains that certain residual risks can be tolerated as part of the application-specific risk management.

→ Dynamics of AI systems: AI systems, which are based on machine learning methods, are often subject to a dynamic during operation that has two causes: On the one hand, the operating environment can change so that the originally learned model only inadequately reflects reality (concept drift). On the other hand, the model can continue learning during operation, for example through user feedback. This is designated model drift. For a potential AI test, this means that the result about the assured properties of the AI system need not be valid at a later date. This represents another central difference to the verification of conventional software. The following measures are conceivable to counter this problem: 1) Model drift can be avoided by introducing structured model updates. Potential quality requirements can be defined for such updates, so that the asssured and tested properties are maintained after the update. 2) Similar to cloud certifications, a continuous test of the AI system by monitoring suitable KPIs is conceivable. However, a suitable selection of such KPIs is currently still the subject of research and development. Alternatively, suitable measures, e.g. safeguards, can be taken to ensure that the system cannot assume critical states. 3) Possible uncontrollable behaviour of AI systems can be prevented by involving a supervisor (human-in-the loop).

→ Uncertainty: Uncertainty regarding the correctness of an output is an intrinsic property of data-driven AI applications. Apart from the simple observation that the application of a model created by a machine learning process to a new, so far unknown input can lead to a correct or even incorrect result, research on the uncertainty of models in the narrower sense is concerned with the view that a learned model can be regarded as a probabilistic function, and thus each statement made by the model is provided in principle with a confidence, the knowledge of which in turn allows various conclusions regarding the use of the model in a given case. Unfortunately, for com-

plex learned models, the actual valid confidence values are not directly visible. The situation is complicated by the fact that the uncertainties arising in the application of the model can be caused not only by different aspects, but also by interacting aspects: insufficient or imprecise data, limitations in the expressiveness of the chosen model class, or an immanent, non-deterministic behaviour of the modelled objective function (e.g. long-term weather forecast). Accordingly, a precise knowledge of the model uncertainty would in turn allow conclusions to be drawn about the data situation, model complexity and prediction quality in the application. The latter in turn is a central element of layered security architectures, where alternative mechanisms are applied at upper levels (e.g. driver takes the wheel), if the AI application on the lower level signals too much uncertainty (monitoring approach). There is a broad spectrum of research approaches to capture the uncertainties associated with a learned model under restrictive conditions, ranging from simple subsequent "model calibration" and targeted interventions in the actual learning process to complex redundancy procedures and more or less holistic mathematical analyses. In view of the ever-increasing model complexity and breadth of applications of learned models in safety-critical areas as well, the development of efficient, precisely effective and generally applicable methods for determining and testing the uncertainty of models is urgently required.

→ Transparency/traceability: An AI system is transparent if its genesis and mode of action are presented openly, completely and understandably. This includes particularly the data basis and the algorithmic component. The decisions/proposals of an AI system are traceable if the factors that led to their creation can be understood by a person.
The following aspects in particular play an important role in transparency: Transparency of the data used for training, the annotation of the data (e.g. inter-annotator agreement using Cohen's kappa or Fleiss' kappa). Transparency in the selection of methods. Transparency and traceability of results (influence weighting of the entered variables). Transparency in the approach (e.g. through a history of the hypotheses tested during parameter optimization or model generation). Transparency in the secured application (i.e. when a model can make sound decisions or when it operates outside or in peripheral areas of the input data). In general, a distinction must be made between transparency for the end user and interpretability.
From a technical point of view the question of basic transparency is not easy to answer, and the tension between higher accuracy or robustness and the explainability of models is a well-known dilemma in the AI world. Although "black box" models are in many cases more accurate or more robust than, for example, rule-based models, they are only conditionally interpretable. In part, this explainability can also be achieved by downstream procedures, such as training of explanatory models or an analysis of the input/output behaviour of models, so-called Local Interpretable Model-agnostic Explanations (LIME) analysis. Currently, the interpretability of models is an active field of research and many efforts are being made to better understand the learning processes of "black box" models, to visualize their internal processes and to explain the resulting decisions.

→ IT security: AI components and AI-based systems are now exposed to IT security risks such as adversarial attacks. Since these often work statistically and their mode of operation is not yet fully understood, quality assurance poses major problems for the IT security of AI components. Modifications of data that are imperceptible to humans, e.g. in images, lead to misclassifications when using adversarial samples, e.g. by subtle manipulation of traffic signs on the road or by adding targeted noise in already existing images. AI systems themselves and the models they contain are also subject to IT security risks. The trained model represents a business value to be protected and must therefore be protected against reverse engineering and its training data. Corresponding attacks can also have an impact on data protection, since techniques already exist that allow the extraction of individual training data records. Detailed explanations can be found in Chapter 4.4.

→ Hyperparameters: In addition to the selected AI method or algorithm and the data used for training and testing, the associated hyperparameters significantly determine its quality and can lead to effects such as overfitting, where the system achieves a particularly high level of accuracy for the training data, but only a low level of accuracy in operation. Hyperparameters include properties of the model regarding its size (e.g. number of layers of a deep neural network) as well as learning parameters like the number of epochs and the learning rate.

4.3.2.3 Test methods

4.3.2.3.1 Verification of AI systems

Suitable verification is the basis of any conformity assessment in the development of systems. The specific challenges of learning systems mentioned in the Introduction also make demands on the verification during development. Beside a corresponding documentation with configuration management, here the consideration lies on testing (check against criteria), verification (formal check of the AI module against the specification) and validation (formal check of the application in the use environment).

The "Product Quality Model" of ISO/IEC 25010 [146], addresses two fields of topics:

1. Functional testing: "what the system does"
2. Non-functional testing: "how the system does it".

For a test environment, an application-specific and boundary condition-dependent action framework should be created, within which test methods can be defined. The test procedures and test depths depend on the identification of relevant user groups such as developers and users, as well as application scenarios, data protection and potential for harm (see Figure 8). The boundary conditions, structures and interfaces of a possible location of the AI-based system should be simulated within the test environment. The test environment should be separated from the external environment so that distortions in results can be avoided. In the test environment, sequences for tests of different depths should be guaranteed. The test depth can be determined based on deployment risk, complexity of the AI application, effort and cost. In order to demonstrate the conformity of a system in a traceable way (conformity assessment), it is necessary to define the underlying requirements unambiguously. For systems based on AI, a catalogue of requirements should be developed in which aspects such as system requirements, system architecture, software requirements, software architecture, source code structure, module structure, software integration, software quality, training and test data quality, system integration and system quality are documented [161]. On the basis of a framework for action, AI methods and capabilities can be subjected to a conformity assessment with a view to appropriate suitability depending on quality and validation requirements, taking into account ethical, legal and social assessment schemes. AI applications can be described based on the criticality pyramid in 4.1.2. Similar to the determination of measurement capabilities in the calibration/testing of measuring instruments using traceable, validated metrological standards and references, reference data, benchmarks and reference methods can be an important part of the test in certain areas. For example, benchmarks validated in ECG analysis can be performed with test data not previously known with the AI method and compared with the results of reference methods.

When testing AI systems, two approaches can be followed: Process tests can be used to verify quality standards for the operation and development of the AI system, while product tests verify assured properties of AI systems. Both test approaches must fit into an overarching testing framework that ensures the comparability of tests of different AI systems. This testing framework should be open with regard to the selection of subsequent test methods, but should be connectable and compatible with established test methods. Examples of established test methods are the conformity assessment of the New Legislative Framework (NLF) or CC [47].

4.3.2.3.2 Process tests

The product AI brings a host of new challenges. Among other things, depending on the method used, transparency or traceability is limited with regard to a decision made by an AI. Therefore, transparency with regard to the AI development process in the form of a process test is even more important [162]. It should also be noted that AI products are often provided in the form of Internet-based services or access Internet and cloud services. Such services are often continuously updated: A test of AI products, especially service-based AI products, should therefore be supplemented by an audit of the processes of the organization providing these products. In addition, organizations using AI products should also be able to obtain proof of their responsible use of such technologies, for example in the form of a test report or an appropriate certificate.

4.3.2.3.2.1 Assessment of the consequences of using AI

At the beginning of these processes, in addition to the requirements adapted to the AI challenges, there is also a consideration of the expectations, demands and fears of other affected parties, e.g. customers and partners of an organization, end users of AI products, etc. Organizations should be able to understand the impact and consequences

of using such products and, if necessary, to reconcile them with their own objectives: Such an extended management of risks of the use of AI, which considers not only risks for an organization but also the impact on third parties, should be implemented and verifiably documented by appropriate management functions, roles and responsibilities.

4.3.2.3.2.2 Development process of AI systems

Transparency with regard to the AI development process in the form of a process test should include the documentation of important decisions regarding the selection of certain criteria and indicators (e.g. metrics, accuracy, precision, recall, specificity and sensitivity). Furthermore, requirements for continuously learning AI systems should be appropriately designed (goal alignment) and documented. Since the training process has a significant influence on the quality of an AI, the training progress must be ensured. This requires a versioning of the software, including the data used for training. In addition to the versioning of the software, the documentation of central and system-relevant decisions is important, e.g. decisions and decision changes regarding model selection, data preparation (feature engineering) and the classification into training and test data. Before validating an AI, the documentation for testing and verification must be completed.

4.3.2.3.2.3 Use of AI systems and their provision as services

Processes involved in the use of AI systems, particularly in their provision as services, include the continuous review and evaluation of performance and security metrics, the determination of appropriate responses to incidents, and the establishment of appropriate countermeasures. In addition to these generic processes, AI must also be considered and supported by appropriate management processes, e.g:

→ The impact of automated decisions made by AI systems and the resulting loss of control.
→ The loss of organizational knowledge that can be caused by the use of automated decision-making systems and the resulting strong attachment to such systems (“blind trust”).
→ The possibility that services of third parties are used for purposes that are questionable within the ethical self-image of an organization.
→ Dealing with limited transparency and explainability of AI systems.

Process tests should be based on established MSS (e.g. ISO 9001 [120], ISO/IEC 27001 [122], ISO/IEC 27701 [163], etc.); however, such standards, as far as they are currently published, only cover parts of the development and use of AI (quality, security, data protection, etc.) The development of a stand-alone MSS for AI, which has already been discussed in Chapter 4.1, is therefore recommended. This can be used in addition to product testing (see 4.3.2.3.3) for conformity assessment and certification as a result of an audit.

4.3.2.3.2.4 Quality assurance after delivery through product monitoring

As a quality assurance measure during operational use in the AI life cycle (see 4.1.2.3.1), an active product observation with evaluation of acquired field data should be normatively defined for AI systems, as is already practised for safety-critical systems in the automotive [164], aerospace and defence [165] industries.

To ensure that recognized problems and risks in the application of an AI system, e.g. unethical or unnecessarily endangering behaviour in the operational environment, lead to appropriate corrective and sustainable improvement measures, it is necessary to feed back the quality-relevant information gained through product monitoring to the corresponding point of action in the AI life cycle. This feedback must include the possibility of warnings to customers and authorities, as well as a product recall.

A systematic feedback for product, system and process improvement through internal [166] and external [167] quality-relevant information is described, for example, in the “Aachen QM Model (AQMM)” [168] as a “Quality Backward Chain” (see Figure 20).

For AI systems, this product observation can be done by storing evaluation results together with the corresponding source data (e.g. sensor data), transferring them to the AI development company and evaluating them there. In commercial aircraft today, this is done partly during flight or when the aircraft is electrically coupled to the gate at the airport, with around 1 GB of data per flight hour. Current motor vehicles also have integrated mobile radio interfaces for transmitting rudimentary field data to the vehicle manufacturer and, in some cases, digital data loggers in order to record at least a certain period of time before a damage event. For AI systems, the amount of data required can be very large, so a workable

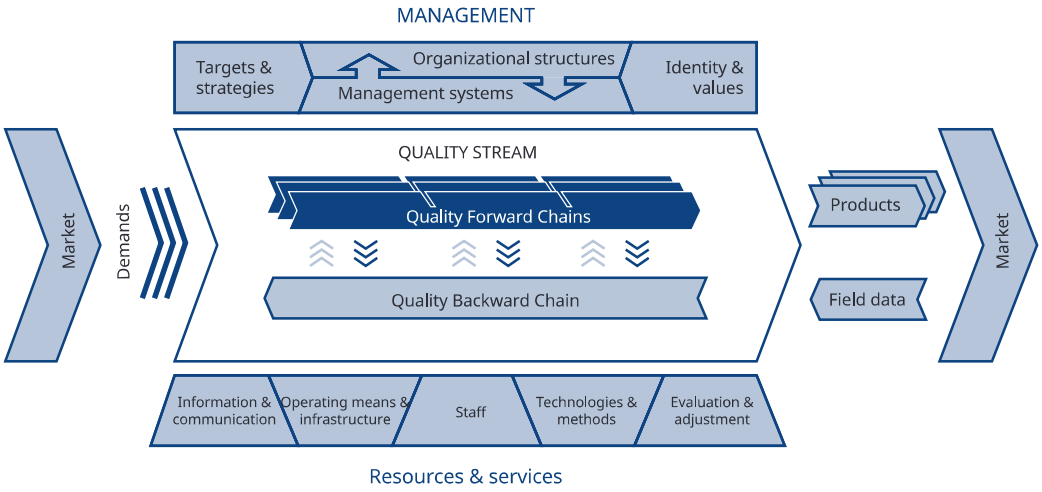

Figure 20: The Aachen QM Model with "field data" and "Quality Backward Chain" [168]

solution for each AI application must be determined during development.

4.3.2.3.3 Product tests

In addition to process tests, which ensure compliance with good standards for the development and operation of AI applications, there is a need for product tests which test the properties of an AI system itself. On the one hand, the test can include the product properties assured by the developer, on the other hand it can confirm compliance with certain industry- or product- specific standards.

For such a product test, a framework is required according to which the functionalities of the AI application can be specified uniformly. In addition, evaluation principles are required which indicate when the promised functionalities can be considered fulfilled. These evaluation principles should particularly include an overview of common metrics that make performance measurable with respect to different technical properties (e.g. robustness against adversarial attacks, reliability etc.). The challenge here is that the adequacy of the metrics used can be highly dependent on the usage context or use case. Some AI models (e.g. word embeddings) do not have their own quality criterion, but can only be compared in the application by further procedures. A solution should be found for this as well.

Example of a use case from the mobility sector autonomous driving

These issues are currently being investigated in initial pilot projects, such as the Federal Ministry for Economic Affairs and Energy (BMWi) project AI security, and are expected to yield valuable findings for the standardization of AI systems:

The goal of the AI security project is the development and investigation of methods and measures for the security of AI-based driving functions for the use case "pedestrian detection". The knowledge gained should make it possible to better determine and assess the technology. In addition, this is intended to create a stringent chain of argumentation which, from the expert's point of view, justifies the security of AI functions. Ultimately, communication with normative committees and certification bodies should support an industry consensus on an AI testing strategy.

Furthermore, it requires the specification of different levels of trustworthiness (see the criticality pyramid in 4.1.2.1.5 or the risk criticality model in 4.4.1.2), which are confirmed by an audit according to scope and depth. For this purpose it is necessary to define a suitable framework at different test depths. The range of methods includes document checks, audits, black box and white box tests, as well as validation and verification.

To carry out such product tests, suitable tools are also required with which the fulfilment of functionalities and performance can be measured in terms of some appropriate

metrics. These test tools need to be developed, and criteria for their evaluation and approval are needed. In addition, a designation requirement for implemented methods and capabilities can be established for AI applications, for example by using the classification matrix for methods and capabilities in 4.1.2.

4.3.2.3.4 Testability-by-design

Analogous to existing concepts such as "privacy-by-design" or "safety-by-design", quality requirements for AI systems should also be taken into account in the design of the application.

The full life cycle of an AI system from the specification of the input data, the processing of raw data to training data and the representative modelling of a purpose-specific, domain-specific knowledge, right up to the application scenarios, must be considered. This also applies in particular to transparency requirements.

The concepts and standards of a testability-by-design for AI applications is a medium-term research topic. With regard to the quality characteristics mentioned at the beginning, at least the following fundamental research questions arise when using AI systems:

- → How can an AI-specific FMECA (Failure Mode and Effects and Criticality Analysis) be performed?
- → How will this have to be updated beyond the period of its development?
- → For which purposes are an AI system and its AI components used? What are the resulting requirements for the testable design?
- → Which AI models are used for the AI components employed? Are there standardized designs that are testable?
- → Are people involved in the decision-making or prognosis by an AI component and if so, in what form? What are the responsibilities with regard to the input and output variables of the AI application?
- → How are the AI models, implementation and training methods selected? What are the requirements for a testable design of the application?
- → Which test methods are relevant, how can the AI component be tested if it has the appropriate properties?
- → And how is the ongoing operation of this component monitored with regard to compliance with the purpose? What conclusions should be drawn for a testable design to simplify testing?

It is also to be expected that certain quality characteristics of AI systems will be easier to verify or their verification will only be possible if the corresponding requirements are already considered during the design and further development of the AI systems. Possible starting points are, for example, documentation of the development process, logging of (intermediate) results or interfaces for corresponding test tools (see 4.3.2.3.6).

4.3.2.3.5 Test infrastructure for conformity assessment and certification

In order to be able to test the quality requirements formulated here, a testing infrastructure consisting of testing laboratories, technical inspectors and the necessary accreditation mechanisms and bodies is required. In particular, accreditation mechanisms should ensure that test bodies and testers have a sound technological understanding to perform these tests. When setting up the test infrastructure, the existing technical IT test infrastructure should be used as far as possible in order to develop marketable tests and establish connectivity to existing test methods. For the certification of persons, the competence of already established, accredited certification bodies can be extended with regard to methods and capabilities of artificial intelligence.

Certification may be based on a potentially updated variant of ISO/IEC 17024 [41]. In order to prove specific competences, further documents such as recommendations, regulations and further standards with regard to AI should be drawn up, extended and consulted.

The use of innovative, AI-supported testing services requires proof of existing professional competence of testers, technical experts, assessors and auditors in order to guarantee quality assurance. Apart from the validation of technical aspects, the potential for harm of an AI application should be assessable by qualified persons on the basis of ethical and legal principles.

4.3.2.3.6 New test methods and new testing tools

Methodological approaches

According to a specification of the system to be tested, the test can be used, for example, with regard to machine learning procedures during training or on a fully trained system. This can be done by analyzing the input and output behav-

iour of models to evaluate invariance, regularity and equivalence. Sensitivity analyses, for example, are suitable for this purpose. In the case of training-accompanied learning, a learning curve can also be tracked and intentionally evaluated in terms of declaration, error probability and adaptivity. For already trained systems, key performance indices can be included which evaluate criteria for suitability and exclusion of the AI for research purposes or a market. This should make it possible to determine, via interpretable quality characteristics, in which environment individual methods and capabilities of the AI can be used.

Using the LIME approach as an example, the goal is to explain systems based on machine learning [44]. Furthermore, models for the interpretation of learning mechanisms can be included for individual AI methods. For multi-layer neural networks, the methods "activation maximization" and "deep Taylor decomposition" are suitable [169].

Methods like LIME, Shapley [170], DeepLIFT [171] and QII [172] can often only be applied to structured data. Methods for unstructured data sets of data types such as text, images and audio are currently in an early stage of development.

A verification of the source code of AI-based systems is only possible to a limited extent using conventional software test methods. This includes statistical code analysis (Grammatech's CODESURFER), runtime verification (Java Pathfinder) or model checking (SPIN model checker) [173].

For different classes of neural networks different verifications can be specified which can be derived from different theories of logic and mathematics. These include verification procedures based on the satisfiability of formulas of Boolean propositional logic (satisfiability theories, SAT), the satisfiability of formulas of first-order predicate logic (satisfiability modulo theories, SMT), reduction to linear problems (mixed integer linear programming, MIP) and robustness of multi-layer perceptron networks (multi-layer perceptron, MLP). For SAT and SMT verifications the classical AI (symbolic AI) of automated reasoning is combined with ML. MIP is based on the logic and algebra of linear programming. Robustness studies of MLP apply findings from the theory of complex dynamic systems in ML [174]. Verification procedures for sub-symbolic AI systems and ML require new techniques which are extremely computationally intensive due to their parameter explosion (e.g. neural networks during autonomous driving).

IT security tests for AI-based systems

An essential aspect of the tests for conformity assessment and certification are security tests, which are divided into static and dynamic tests. Dynamic security tests play a central role here, offering a wide range of methods and techniques. A brief overview is provided by overviews such as document ETSI TR 101 583 [175]. In this document is an enumeration and explanation of relevant methods and approaches for security testing, such as risk analysis and risk-based[24] security testing, functional testing of security functions, performance testing, robustness testing and penetration testing.

A number of techniques have been developed for the security testing of traditional software systems. These can only be applied to AI-based systems to a limited extent, if at all. Security tests for classical systems can partly be adapted for AI-based systems, e.g. the widely used fuzzing can also be used for AI-based systems in a modified form (see for example [176], [177]). In order to cover the security risks and attacks specific to AI-based systems, new techniques and approaches are needed that take AI-specific aspects into account, such as the relevance of training data. The techniques that address AI-specific security aspects are referred to as adversarial ML (AML) [178]. Coverage criteria are still a particular hurdle in security tests. There are a number of published metrics for this. However, a meta-study found only a low correlation between the existing metrics developed for AI systems and the robustness against attacks when these metrics were considered in tests [179]. An overview of existing techniques and metrics, including information on application, is currently the subject of the ongoing project ETSI DGS SAI-003 „Security Testing of AI".

4.3.2.4 National implementation programme for the Standardization Roadmap AI

The rapid spread and high complexity of AI systems is creating new technological challenges across industries. The dynamics prevailing in AI developments require a stable framework for action for all actors in research, industry and society in order to jointly use the available innovative power to shape the future and to promote the economic and social benefits of the use of AI systems in a converging manner. To operationalize those recommendations for action of the Standardization Roadmap AI that concern the technical

24 "Risk-based" in English.