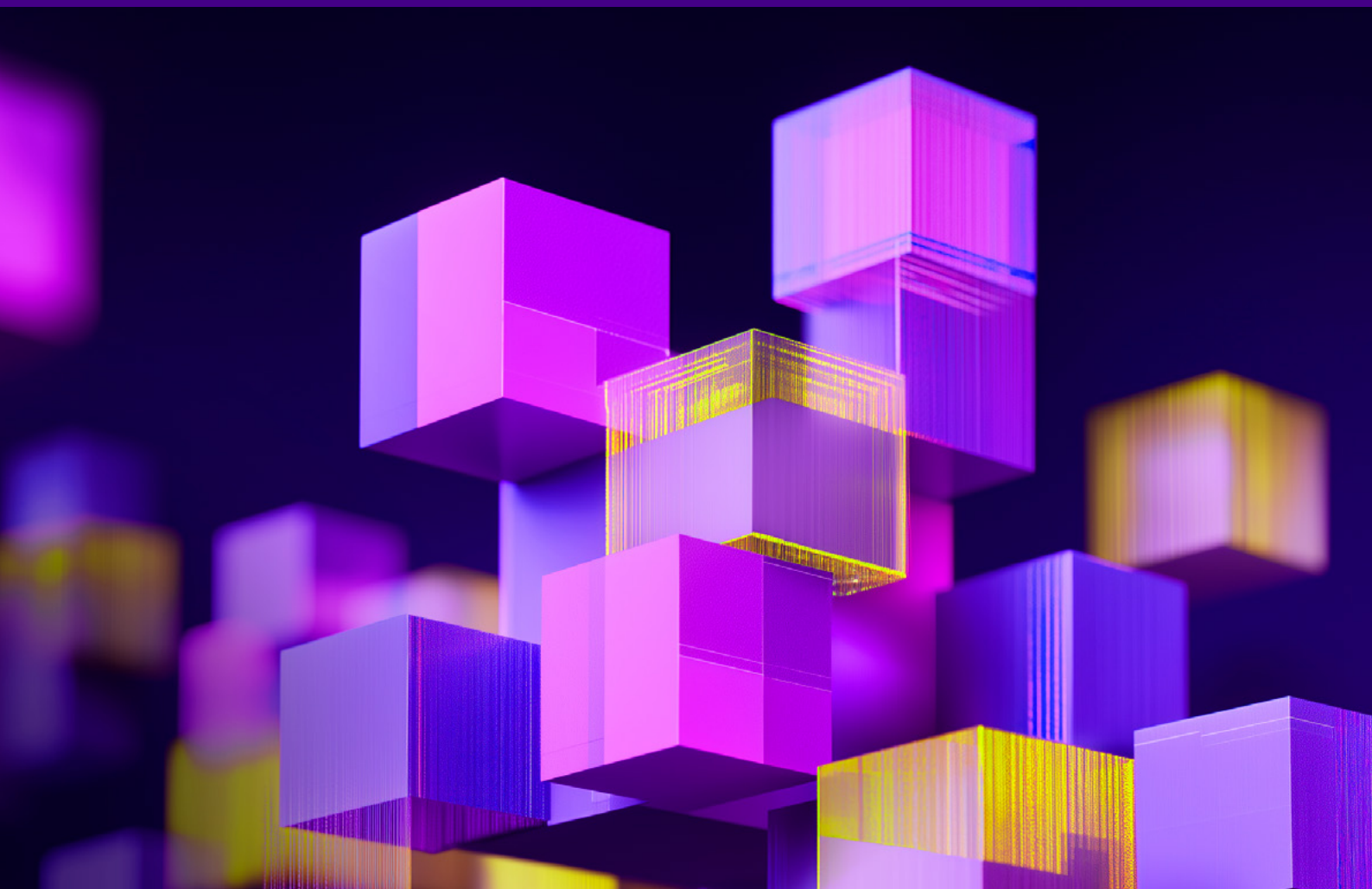
MISSION KI

November 2025

Quality Standard for Low-Risk AI



©acatech – German National Academy of Science and Engineering This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) licence. https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en

Note

Some of the indicators in the MISSION KI Quality Standard are based on the "Specification for the assessment of the trustworthiness of AI systems" developed jointly by the IEEE, VDE, Positive AI and IRT SystemX.

MISSION KI
c/o acatech
German National Academy
of Science and Engineering
Karolinenplatz 4, 80333 Munich, Germany

mission-ki.de

Gefördert durch:

Bundesministerium
für Digitales und
Staatsmodernisierung

aufgrund eines Beschlusses
des Deutschen Bundestages

acatech
DEUTSCHE AKADEMIE DER
TECHNIKWISSENSCHAFTEN

Foreword

Promoting trustworthy artificial intelligence (AI) is critical to the long-term development and acceptance of AI technologies. The **MISSION KI** National Initiative for Artificial Intelligence and Data Economy was launched within this context. Funded by the German Federal Ministry for Digital Transformation and Government Modernisation (BMDS), the initiative is being implemented by acatech – the German National Academy of Science and Engineering. **MISSION KI** iaims to strengthen Germany's digital competitiveness. As a key initiative in the German Federal Government's digital strategy, **MISSION KI** supports the development of uniform standards and improves the availability and accessibility of data.

A key element of the initiative is the development of a voluntary quality standard for low-risk AI[1], enabling transparent, standardised and comparable verification of AI system quality. The standard outlines the framework conditions and assessment procedure for quality verification, and clarifies the relationships between the main components of the assessment.

The **MISSION KI** Quality Standard for artificial intelligence was developed in close collaboration with a project consortium consisting of PwC Germany (lead), the AI Quality & Testing Hub, CertifAI, Fraunhofer IAIS, TÜV AI.Lab and VDE.

1 To improve readability, this document uses the terms 'MISSION KI Quality Standard' and 'Quality Standard' as abbreviations for the 'MISSION KI Quality Standard for Low-Risk AI' with the same intended meaning.

Contents

List of figures

Preamble

The **MISSION KI** Quality Standard is aimed at providers of AI systems, particularly start-ups and small to medium-sized enterprises (SMEs), whose applications fall below the high-risk threshold set out in the European AI Act. These systems shape a wide range of economically and socially significant processes, including production, customer service and administration. These systems require practical and verifiable criteria to demonstrate quality and trustworthiness in a clear, understandable manner.

The standard provides a structured approach to the internal assessment, documentation and external communication of quality measures. It helps organisations

- standardise their processes and responsibilities regarding AI,
- create transparency for customers, partners and supervisory authorities,
- and prepare for future regulatory requirements.

The standard benefits not only AI providers, but also customers, investors and public clients. Standardised assessment criteria foster comparability, build trust and support informed decisions in procurement and assessment processes.

The **MISSION KI** Quality Standard thus aids in establishing quality as a competitive advantage and in enhancing the widespread acceptance of trustworthy AI applications in business and administration.

Introduction

Purpose and use

This document describes the **MISSION KI** Quality Standard for Low-Risk AI and specifies the criteria, assessment procedure and evidence and validation requirements. The standard provides organisations with a foundation for self-assessment of their AI systems and can be supplemented by optional validation from external assessment bodies.

Structure of the document

Section 1 defines the scope and framework conditions of the assessment. It specifies which system components should be included based on the intended purpose, describes the structural elements (quality dimensions, criteria, indicators and observables) and presents the resulting assessment statement.

Section 2 explains the concept of assessment depth and details its definition within the procedure, including the requirements for authorised/validating persons.

Section 3 describes how to perform the assessment: Use case description, protection needs analysis, rating according to the Assessment Catalogue, evidence/tests, validation, overall assessment, assessment report and validity monitoring.

The details and templates required for implementation, including use case description, protection needs analysis, Assessment Catalogue, Test Method Catalogue and glossary) can be found in the appendices.

Guiding principles of the standard

The quality standard is based on six guiding principles, which are explained below and defined in Table 1. With the entry into force of the European AI Act, many providers are already subject to a conformity assessment. On the other hand, assessment in accordance with the **MISSION KI** Quality Standard is done voluntarily and primarily out of economic interest.

To supplement the EU AI Act, the standard is directed towards organisations in the unregulated sector and those already impacted by regulation or likely to be affected by it in the future due to the development of their systems, which require a practical approach to prepare for this. Organisations seeking to meet voluntary trustworthiness requirements will also benefit from its use.

To address regulatory and standardisation requirements, the Quality Standard is compatible and consistent with the EU AI Act and other relevant directives. For effective use within the voluntary sector, implementation must be efficient in terms of time, personnel and finances, and proportionate regarding cost and benefit. Therefore, the standard is essentially designed as a self-assessment tool. Reliable assessment results are required for a reliable quality promise. This is achieved through results that are verifiable, reproducible and objective. The simultaneous demands of efficiency, reliability and comparability require a high degree of systematisation.

Finally, the standard must be accessible so that results are communicated, understood and verified, giving assessed companies a competitive advantage.

Guiding principles of the MISSION KI Quality Standard

Guiding principles of the MISSION KI Quality Standard	
Compatibility	• There is an overlap between the Quality Standard and the EU AI Act or sectoral regulation, such as the Medical Device Regulation (MDR) or other regulations/standards. • The requirements of the Quality Standard are compatible with the high-risk requirements from Section 2 of the EU AI Act. • The Quality Standard may also "exceed" the requirements of the regulations and standards.
Soundness	• The Quality Standard should represent a minimum level that convincingly indicates an adequate standard of quality. Therefore, the assessment results should offer the most reliable assurance possible.
Efficiency	• Conducting the assessment in accordance with the Quality Standard must be feasible in terms of time, personnel and financial expenditure. • A greater assessment depth must proportionate to the added value delivered. • A high level of quality should be maintained.
Voluntary nature	• An assessment in accordance with the Quality Standard is conducted primarily for economic reasons, not because of any legal requirements (e.g. providing a clear, understandable and credible indication of quality) • The assessment requirements should also provide the requesting target group with clear added value (i.e. a needs-oriented interpretation of quality) • The assessment requirements should provide further clear added value to the supplying target group (e.g. partial compliance with the EU AI Act, process improvement, etc.)
Comparability	• Assessments based on the Quality Standard should be comparable with each other. • Tests should be uniform and reproducible. • The assessment procedure should be as objective as possible.
Accessibility	• The assessment statement should be understandable for all parties involved. • The results should be easy to communicate. • The assessment requirements should be specific, tangible, verifiable and, where possible, quantifiable.

Table 1: Guiding principles of the MISSION KI Quality Standard

Basics of the assessment procedure

The Quality Standard was developed taking into account the guiding principles above. At the same time, there is already established preliminary work such as VDE SPEC 90012[2], the Fraunhofer IAIS AI Assessment Catalogue[3] and the Joint Specification V1.0 for the Assessment of the Trustworthiness of AI Systems[4], which describe tried-and-tested AI assessment approaches. These have been taken into consideration when developing the **MISSION KI** Quality Standard.

2 VCIO-based description of systems for AI trustworthiness, VDE SPEC 90012 V1.0 (en) characterisation, 2022. URL: https://www.vde.com/re-source/blob/2176686/a24b13db01773747e6b7bba4ce20ea60/vde-spec-vcio-based-description-of-systems-for-ai-trustworthiness-characterisation-data.pdf (Accessed: 21/06/2025).

3 Poretschkin, M., et al., KI-Prüfkatalog: Ein Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz, [AI Assessment Catalogue: Guideline for Designing Trustworthy Artificial Intelligence], Fraunhofer IAIS, 2021. URL: https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefkat-alog/202107_KI-Pruefkatalog.pdf (Accessed: 21/06/2025).

4 Joint Specification V1.0 for the Assessment of the Trustworthiness of AI Systems, 2024. URL: https://standards.ieee.org/news/joint-specification-trustworthy-ai-systems/ (Accessed: 06/11/2025).

The VDE SPEC 90012 specification presented in April 2022 is used to assess AI-specific risks of a given AI system in terms of ethical and other relevant characteristics. The catalogue uses the VCIO model to assess adherence to predefined criteria and display the results with an AI trust label ranging from "A" (best) to "D" (worst). The Fraunhofer IAIS AI Assessment Catalogue (April 2021) also aims to check the risks of an AI application within the framework of six dimensions of trustworthiness. It takes a risk-based and application context-dependent approach and sometimes proposes specific technical tests for mitigating risks.

The assessment procedure defined in the **MISSION KI** Quality Standard consolidates the strengths of these three approaches into an integrated procedure model. In the interest of accessibility and comprehensibility, the assessment system and the assessment basis according to the VCIO model are derived from VDE SPEC 90012.

To enable an assessment depending on the protection needs of an AI system, the VCIO model is expanded to consider the intended purpose of the AI system. To this end, a protection needs analysis (see. 3.2) similar to that in the AI Assessment Catalogue is performed prior to the VCIO approach. In addition, the observables (the lowest level of the VCIO approach) are expanded by obtaining application-related evidence and technical tests (see 3.4) to further increase the resilience of the assessment result.

The contents of VDE SPEC 90012 and the AI Assessment Catalogue are combined to formalise the values (referred to as quality dimensions in this assessment standard, see. 1.2), criteria, indicators and observables.

1. Scope of the assessment procedure

The assessment procedure applies to AI systems that have been assigned a specific purpose (see 1.1). The scope of consideration corresponds to a product-oriented process assessment. This assesses the specific quality measures taken during the development of the AI system and the precautions taken for quality assurance during deployment (see 1.2).

Therefore, the scope typically focuses on the provider side of the AI system (the "provider" as defined by the EU AI Act). The assessment procedure can also be performed downstream at the deployer's premises, provided that the necessary information and documentation are available.[5]

The organisation conducting the assessment is always referred to as the auditee. As a result of the assessment, the level of quality and assurance measures taken is compared with the protection needs resulting from the intended purpose of the AI system (see 1.3). The validity of the assessment results is firmly linked to the intended purpose of the AI system and the technical implementation specified as the basis for the assessment (see 1.4). These aspects of the scope are elaborated on further in the subsequent sections.

1.1 AI systems as assessment subjects

AI systems are permitted as assessment subjects, provided they are defined in accordance with the European AI Act (see Glossary). For the assessment procedure to be applied, an AI system must possess both of the following characteristics: implementation of an AI method and having a clearly defined intended purpose.

1.1.1 AI system

An AI system is "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations or decisions that can influence physical or virtual environments" (Article 3 (1) EU AI Act).

For the purposes of this Quality Standard, an "AI system" is a technical term denoting a functional combination of one or more AI components and non-AI components, with a specific intended purpose and application context. A key feature of AI is the ability to derive or learn from data inputs. This can be achieved by one or more AI components, for example, based on machine learning (ML) processes.

An AI system can either be standalone and interact directly with users via an interface (e. g. a translation model as a web application) or be integrated into larger products or systems (e. g. control or monitoring elements in machines).

5 It should be noted that, in accordance with the European AI Act, companies using an externally supplied AI system already assume the role of provider, for example, by assigning or modifying the intended purpose of the AI system and by making significant technical changes. See Recital 84 of the AI Act for more information.

For the performance and validity of the assessment, it is essential that the AI system is clearly defined as the assessment subject and, if applicable, delineated from other systems.[6] Consistency with the intended purpose of the AI system (see 1.1.2) must be ensured. All components necessary for achieving the specific purpose should be considered part of the AI system and delineated from the system environment via specified interfaces. Section 2 of the Fraunhofer IAIS Assessment Catalogue, for example, provides assistance with technical delineation. Figure 1, taken from this catalogue, shows a simplified representation of an AI system with an ML component.

For instance, an AI system for machine monitoring could have interfaces to sensors and a control element. The sensors measure the machine data, and the control element receives the AI system's forecast as additional input. Based on this, it initiates actions such as shutting down the machine if necessary.

In this scenario, all components required to generate the forecast based on the sensor data (e.g. data preprocessing and the ML model) could be considered part of the AI system. Any other components (e.g. an expert system that reacts to the forecasts or the sensors) are delineated separately as part of the wider system environment that is not subject to the assessment.

Furthermore, in order to perform the assessment, the AI system must be implemented at a sufficiently high level of maturity in accordance with the described specification and delineation (see also 3.1) The assessment procedure does not apply to prototypes. The maturity level of the AI system must be assessed by the auditee on a case-by-case basis.

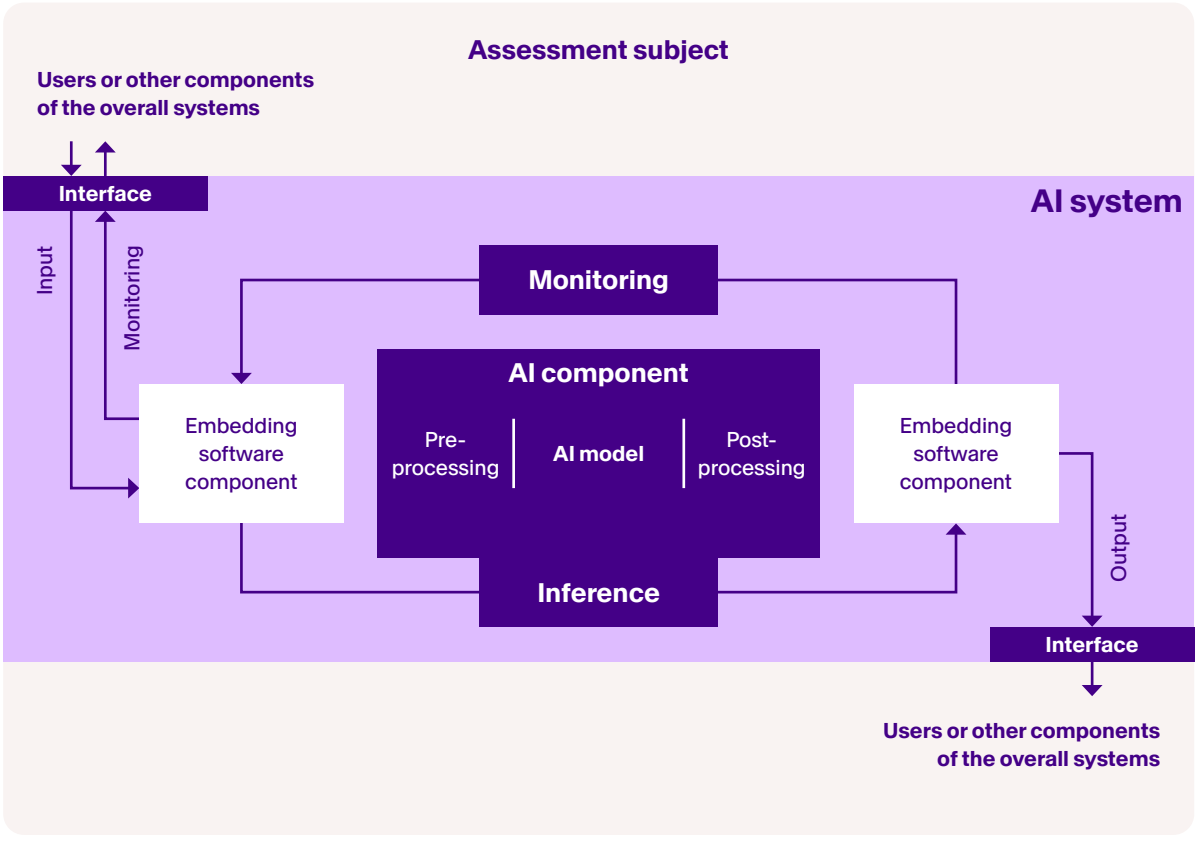


Figure 1: Abstract structure and delineation of an AI system. Source: Fraunhofer IAIS Assessment Catalogue (2021)

6 The structure of the AI system, including how it is delineated from other systems (e.g. via interfaces), must be specified at the beginning of the assessment. This can be done in the form of an architecture diagram (see Section 4.1), among other things. If an AI system comprises several AI components, individual assessment requirements may need to be applied repeatedly for each component

1.1.2 Intended purpose

The intended purpose of an AI system is "the use for which an AI system is intended by the provider, including the specific context and conditions of use, as specified in the information supplied by the provider in the instructions for use, promotional or sales materials and statements, as well as in the technical documentation" (see Article 3 (12), AI Act).

The **MISSION KI** Quality Standard focuses on AI providers and the assessment of an AI system in relation to product processes and measures up to the point of handover to the deployer. It is therefore typically not necessary (or possible, e.g. if the provider and deployer are not the same entity) for the intended purpose to provide detailed information about the (subsequent) deployment environment or production data.

However, the purpose should be defined so precisely in the use case description (see 3.1) that sufficient quality assurance by the provider in relation to this purpose can be substantiated.

1.1.3 Third-party components

The assessment procedure does not fundamentally exclude AI systems that integrate third-party components or are based on them. For example, so-called "general-purpose AI models" (e.g. large language models) fall within the scope of the assessment procedure if they are assigned a specific purpose as the basis for the assessment (see 1.1.2).

The integration of external components is explicitly considered in the Assessment Catalogue from the perspectives of supply chain reliability and cybersecurity. However, when working with third-party providers, it is crucial to the success of the assessment that relevant data and documentation are obtained.

In general, if evidence or the possibility of validation is lacking, a high or good quality rating of the AI system cannot be achieved. (see Section 2).

1.2 Quality dimension as a structural element of the assessment

As part of the assessment procedure, an AI system is recorded and analysed against six criteria. These are referred to as quality dimensions in the context of the **MISSION KI** Quality Standard:

- Data quality, protection and governance,
- Non-discrimination,
- Transparency,
- Human oversight and control,
- Reliability,
- AI-specific cybersecurity.

These quality dimensions form the basis of the **MISSION KI Assessment Catalogue**, which covers all assessment requirements. The quality dimensions originate from the "Ethics guidelines for trustworthy AI", developed by the High-Level Expert Group convened by the European Commission (AI HLEG[7]). This document lists a number of key principles intended for use in assessing the trustworthiness of AI systems. These principles also form the basis of the concept of trustworthiness in the EU AI Act and are mentioned in Recital 27 accordingly as the basis for developing trustworthy AI systems.

7 Ethics Guidelines for Trustworthy AI, High-level Expert Group on AI, 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (Accessed: 27/10/2025).

To ensure compatibility with the European AI Act, it is only logical that the **MISSION KI** Quality Standard should also adopt and integrate these principles. Several (criteria) catalogues and standards for AI trustworthiness based on the AI HLEG have now been published in Europe and Germany. **MISSION KI**'s quality dimensions are specifically based on VDE SPEC 90012 and the Fraunhofer IAIS AI Assessment Catalogue.

These six quality dimensions provide the structure for the two most important components of the assessment: the protection needs analysis (see 3.2) and the Assessment Catalogue requirements, which follow the VCIO rating from VDE SPEC 90012 (see 3.3). This structure further subdivides the quality dimensions into criteria, indicators and observables (see Figure 3 and the descriptions in 3.3).[8]

The protection needs analysis identifies the potential protection needs of an AI system based on its intended purpose, while the rating assesses the extent to which appropriate quality measures and technical precautions have been implemented. The results of the protection needs analysis and the rating against the assessment requirements are structured separately for each quality dimension and its associated criteria.

1.3 Assessment statement

The overall assessment statement is mainly based on two central sections. The protection needs analysis (see 3.2) determines the level of protection required for each quality dimension, derived from potential damage levels. The rating against the assessment requirements (see 3.3) shows the extent to which state-of-the-art measures have been implemented and documented, through an evidence-based assessment.

These two results are then compared at the criteria level (see 3.6), to assess whether the quality measures implemented and documented correspond to the protection needs identified. Passing certifies that the documentation shows the measures implemented in the Assessment Catalogue (i. e. the assessment requirements) are appropriate for the protection needs.

As the quality assessment is a plausibility check rather than a comprehensive risk analysis, passing the assessment does not provide any information about the AI system's residual risks.

1.4 Validity

The assessment procedure describes an assessment at a specific point in time. The intended purpose and technical implementation of the AI system at the time of the assessment are particularly relevant for the assessment and evaluation. These are recorded in the use case description (see 3.1).

The assessment result, based on the assessment procedure and Assessment Catalogue, applies exclusively to the system specified at the beginning of the assessment. The result becomes invalid as soon as the original intended use or technical implementation changes significantly and these changes affect the Quality Standard requirements.

8 VCIO stands for subdivision into values, criteria, indicators and observables. This structure has been adopted, but "value" has been replaced by "quality dimension".

For AI systems that continue to learn during deployment (e.g. incrementally), the methodology and conditions of further training (e.g. training data requirements, loss value logging) are also recorded in the use case description and considered in the assessment requirements[9]. The assessment result depends on this too, and would become invalid in the event of significant deviations.

At the same time, even AI systems that do not continue to learn during deployment are often subject to external dynamics due to data or concept drifts in the specific application. Therefore, the validity of the assessment results must be limited.

9 This only applies to the assessment requirements to the extent that continuous assessment and monitoring aspects are included within the scope. Currently, the Assessment Catalogue focuses on an assessment for the provider of an AI system at a specific point in time. Aspects of application and continuous implementation are only considered in detail in the requirements to the extent that this is possible for a provider.

2. Assessment depth

The assessment depth significantly influences the implementation of the assessment procedure and the effort required to perform the assessment. It determines the degree of assurance and plausibility verification of assessment statements and defines the roles required to validate the results within the assessment procedure.

The assessment depth is determined by three central characteristics of the assessment:

1) The level of detail of measures that must be fulfilled according to the minimum requirements in the Assessment Catalogue. These minimum requirements are determined by the need for protection and result in a clearly defined, minimum level of requirements. A greater assessment depth goes hand in hand with more extensive requirements for data, models, processes, assessments, tests or organisational structures, among other things.

2) The evidence to be provided in the form of documentation and technical measures performed, such as tests, must also be reproducible at higher assessment depths.

3) The degree to which assessment statements are validated by one or more auditors is also a factor.

The exact composition of these characteristics for the individual assessment requirements depends on the selected protection need. These relationships are summarised in Table 2.

A greater assessment depth is usually associated with increased assessment effort, potentially involving independent auditors at higher validation levels. In the context of self-assessment, independent auditors refers to calling in experts from the organisation to be assessed who are not involved in the development.

Additionally, the voluntary involvement of external assessment bodies is expressly possible. Their involvement can further strengthen the level of assurance, particularly in cases involving high assessment depths and increased protection needs.

While an increase in the requirements of the Assessment Catalogue beyond the highest level is not explicitly provided for, it is also possible on a voluntary basis. Deployer-specific evidence relating to the deployment environment and monitoring can be added to the Assessment Catalogue if desired.

The **MISSION KI** Quality Standard's basic approach makes resilience and efficiency decisive factors in defining the assessment depth. The scope of measures and evidence, and the need for validation, is linked to the protection needs analysis, ensuring the highest level of quality assurance for the assessment statement in areas with the highest protection needs.

In areas with lower protection needs, the effort required to provide evidence is reduced accordingly. Even at the highest assessment depth, the required measures do not exceed proven best practices. Instead, validation according to the four-eyes principle and the reproducibility of the results ensure a high degree of resilience of the assessment statements.

Assessment depth

	Protection need			
	–	low	moderate	high
Minimum requirement from the Assessment Catalogue	No measures need to be implemented (Level D Observable)	**Simple** measures must be implemented (Level C Observable)	**Advanced** measures must be implemented (Level B Observable)	**Best practice** measures must be implemented (Level A Observable)
Provision of evidence	–	Create and maintain **simple evidence** (e.g. documentation) of the implemented measures; Create and maintain **technical evidence** of the **tests performed and the associated results** (e.g. system extracts, dashboards);	Create and maintain **detailed evidence** (e.g. documentation) of the implemented measures; Create and maintain **technical evidence** of the **tests performed and the associated results with metadata** (e.g. system extracts, dashboards, logs for model versions and data sets);	Create and maintain **detailed evidence** (e.g. documentation) of the implemented measures; Create and maintain **reproducible technical evidence** of the **tests performed and the associated results with metadata** (e.g. system extracts, dashboards, logs for model versions, and data sets); Reproduce the results; Evidence that the internal assessment has been performed (e.g. via an audit trail);
Validation	–	Validation by an **authorised person from the responsible team** (e.g.: the product or technical owner); People involved in the development can determine and confirm protection needs and ratings; **Required qualifications:** Experience in developing/ operating AI systems or a comparable qualification	Validation by an **authorised person**, who was **not involved in the development of the AI system** (e.g. the product or technical owner from another team); People involved in the development can determine protection needs and ratings; **Required qualifications:** Experience in developing/ operating AI systems or a comparable qualification	Additional validation by **authorised persons from a hierarchically independent body** (e.g.: internal audit); People involved in the development can determine protection needs and ratings; **Validations** are performed according to the **four-eyes principle** i.e. by two people not involved in developing the AI system; **Required qualifications:** They must have experience in the development/ deployment of AI systems or a comparable qualification and experience in auditing/ assessment.

Table 2: Protection needs and assessment depth

3. Performing the assessment

The following sections provide a detailed description of how the assessment is performed (see Figure 2).

First, the organisation being assessed prepares a use case description (see 3.1), which specifies the subject and scope of the assessment. A protection needs analysis is then performed to identify particularly critical quality dimensions for safeguarding the AI system and any criteria that may not be applicable to the AI system (see 3.2). In the third step, the AI system is rated against all the relevant quality dimensions in accordance with the assessment requirements (see 3.3). In the fourth step, controls and evidence for the rating are documented, particularly using technical assessment methods (see 3.4). The rating is then validated in the fifth step on the basis of the evidence at the required assessment depth (see 3.5). Finally, the protection needs analysis rating is systematically compared to provide an overall result (see 3.6). The assessment concludes with the results documentation (see 3.7), and the validity of the assessment results can be monitored or extended if necessary (see 3.8).

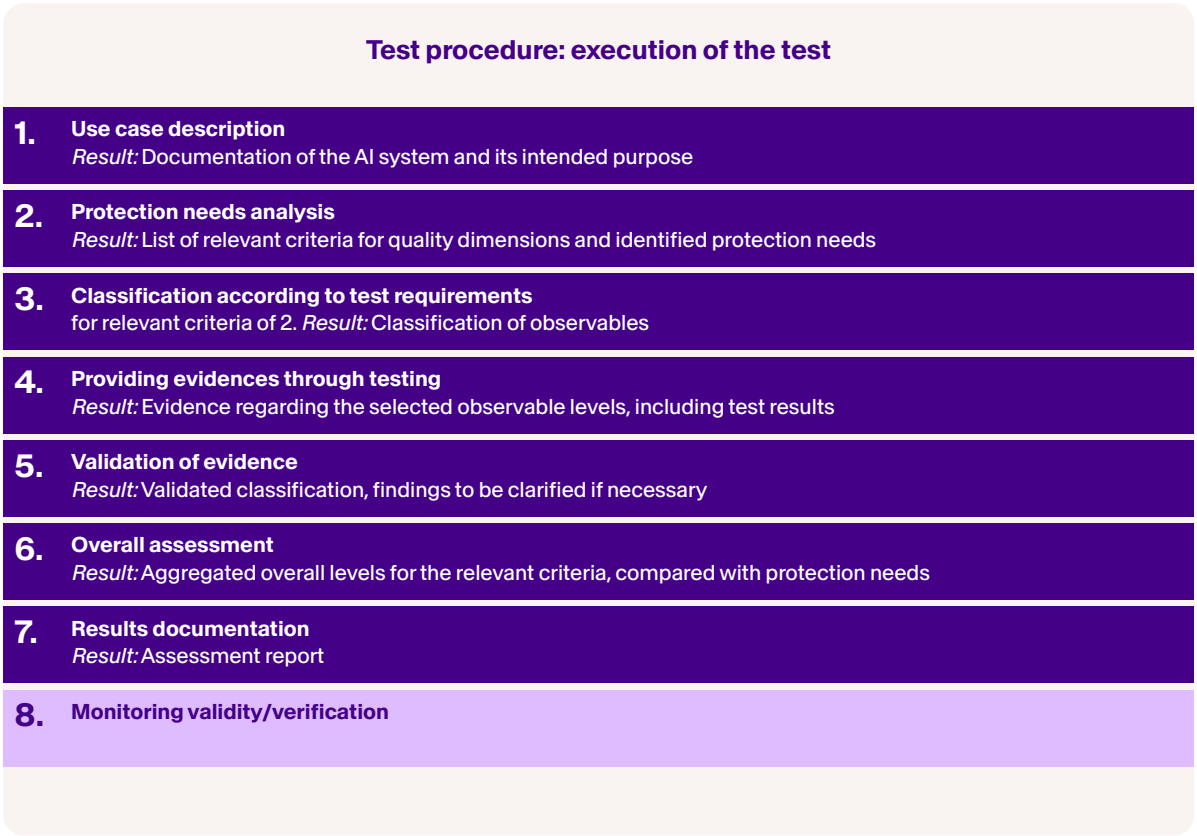


Figure 2: Overview of the various steps in the assessment procedure

3.1 Use case description

The use case description serves as the initial specification of the assessment subject, particularly with regard to its intended purpose and the scope of the AI system's area of application. It enables (internal) auditors called in for validation who have had no previous contact with the AI system to make an initial assessment of the assessment subject.

The **Use case description template** (see template) provided in the appendix can be used to prepare the relevant information for the self-assessment.

Any statement about the trustworthiness of an AI system is always linked to its intended purpose and area of use. Protection needs can vary significantly between different areas of use for the same AI system.

Therefore, the protection needs analysis and the rating of the criteria must be understood within the context of the scope of use as defined in the use case description. For this reason, the scope of use is also a fundamental element of the assessment report and the assessment statement.

In addition to the scope of use, the limits on the input range and the output format, additional relevant context information is requested. This includes potential changes to the AI system during deployment, including after the assessment period, and the type of deployment (local or cloud-based). It also specifies whether and how humans should be involved in the deployment and supervision of the AI system.

The information collected provides the necessary context in the **assessment report template** (see appendix), while avoiding the collection of sensitive data. This allows companies to present the terms of use of their AI system to which the assessment relates in a transparent and clearly understandable manner in the assessment report.

3.2 Protection needs analysis

The **protection needs analysis** (see appendix) is used to filter out irrelevant criteria for the given use case. This is based on the assumption that not every criterion of a quality dimension is equally relevant for the assessment of quality, depending on the task and scope of use of the AI system.

In this way, the protection needs analysis offers an efficient method of identifying relevant quality dimensions at the criterion level and categorising them according to their level of protection needs. At the same time, the protection needs analysis provides a practicable way to ensure compatibility with the European AI Act, which requires "high-risk" AI systems to have a specific intended purpose.

The level of protection needed is examined separately for each quality dimension at the criterion level. This is based on the potential damage that could result if the requirements of the individual quality dimensions at the criteria level cannot be met. Specifically, possible damage ("worst case") scenarios are derived from the protection objectives listed in Art. 1 (1) of the AI Act. Each protection need is rated in three categories – low, moderate or high. Alternatively, a quality dimension's criteria can also be assessed as "not applicable". More detailed explanations of the analysis and the assessment scheme can be found in Appendix 4.3 Based on the assessments of the relevance of the quality dimensions and their associated criteria (by excluding or determining the protection needs), further assessment can be performed in a targeted and efficient manner.

In principle, there is no in-depth analysis in the form of a rating (see 3.3) for "not applicable" criteria, or for "not applicable" quality dimensions if all the criteria of a quality dimension are "not applicable". However, at the provider's express request, it is possible to deviate from this principle and assign a rating for "not applicable" criteria or quality dimensions as well. The identified protection needs in turn implicitly serve as a minimum level in the overall assessment (see Section 2).

3.3 Assessment requirements

A core component of the assessment procedure is determining the quality of an AI system – the "rating". To allow for systematic and efficient rating, the Quality Standard requirements are organised hierarchically within a structure based on the VDE SPEC 90012 VCIO framework (see Figure 3).

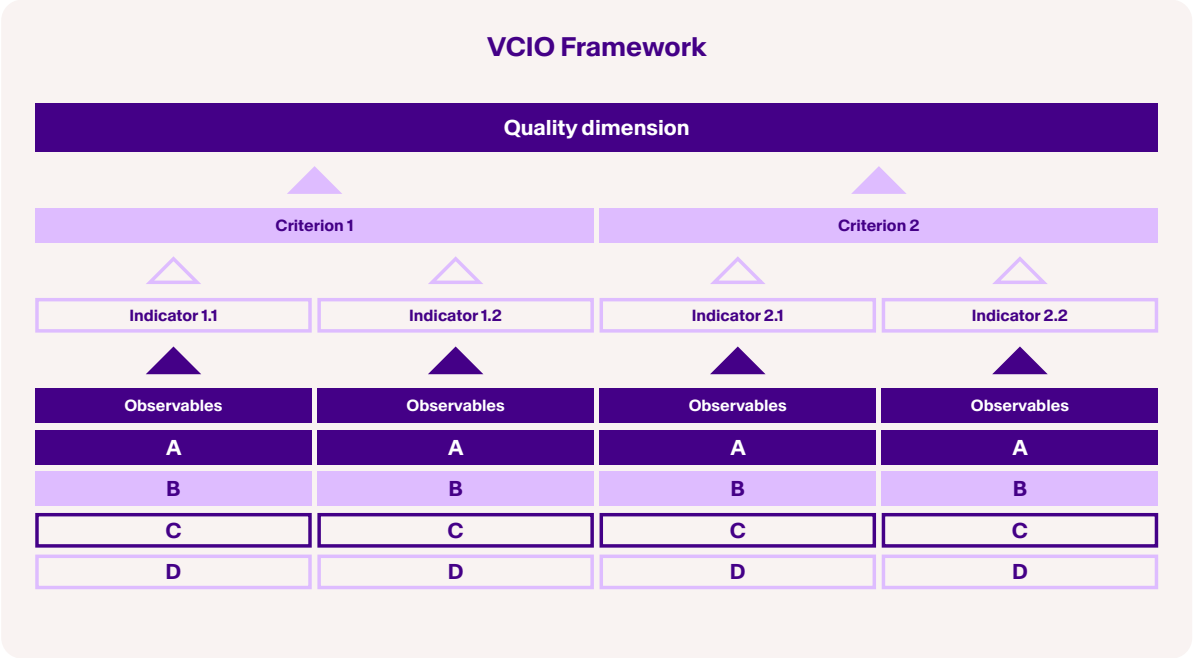


Figure 3: VCIO framework

The appended Assessment Catalogue (see appendix) contains all of the quality dimensions, criteria, indicators and observables. Within the framework of the **MISSION KI** Quality Standard, the terms from the VCIO framework have the following meanings:

Quality dimension: These form the basis of the concept of quality in the **MISSION KI** Quality Standard and define the overarching quality objectives that an AI system should fulfil.

Criteria: Quality dimensions are divided into criteria (see Figure 4) to further narrow down quality objectives, such as "Reliability" to "Performance & Robustness" and "Fallback Plans & General Safety". At this level, the protection needs analysis is compared with the assessment requirements, since quality dimensions are extremely broad, whereas the downstream indicators are too detailed.

Indicators: Each criterion is specified by indicators that are crucial for ensuring the quality of the AI system. These indicators are divided into three types of required action:

- **Type 1 (analysis):** Fokus auf Risikoanalysen und Zweckbestimmung des KI-Systems.
- **Type 2 (measures):** These are subdivided into organisational measures (e.g. governance or training) and technical measures (e.g. data preparation or testing), which must be implemented to mitigate the identified protection needs/risks.
- **Type 3 (assessment):** Requires justification for and acceptance of remaining residual risks after the measures have been implemented.

Some indicators may be relevant to several criteria, as indicated in the Assessment Catalogue (see Assessment Catalogue at the observable level) by a reference.

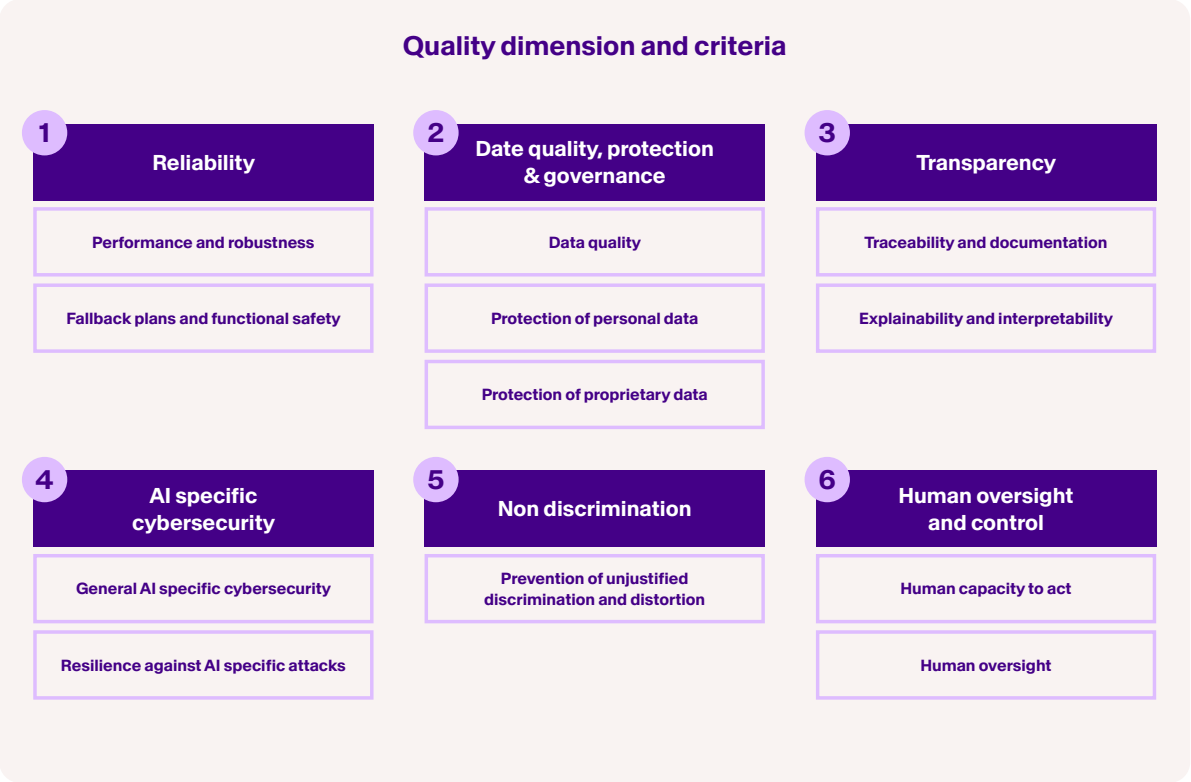


Figure 4: Quality dimensions and criteria

Observables: These are the detectable/measurable characteristics that determine the quality of a system, rated at levels ranging from A (best) to D (not fulfilled at all). Each observable indicates the degree to which an indicator is met. The observables are graded according to the complexity and scope of the requirements to be met.

The protection needs analysis determines the applicability of the criteria, and thus the associated indicators, for a specific AI system. Therefore, rating should only be performed for applicable indicators. Since protection needs and observables have the same number of levels, they can be compared for the final assessment of an AI system (see 3.6).

In the context of conducting the assessment, "rating" means selecting the observable level that applies to the specific AI system for each indicator. The rating, and thus the implementation of the required analyses, measures and assessments, must be verifiable by providing evidence (see 3.4). Finally, an overall assessment of the system can be made through aggregation (see 3.6).

3.4 Provision of evidence

To justify or prove the correctness of the AI system's rating, the auditee must systematically record and provide appropriate controls and evidence. The exact requirements for the scope of the evidence can be found in the Assessment Catalogue and are derived from the respective observables. All points mentioned in the observables must be fully met and substantiated by appropriate evidence.

Individual pieces of evidence may cover several indicators across different quality dimensions. Duplicate evidence is not necessary. Evidence and certifications already in place (e. g. GDPR elements for the protection of personal data, ISO 27001 or BSI IT-Baseline Protection (IT-Grundschutz)) are recognised as evidence.

Various types of evidence are used in the assessment. (Table 3) shows how the observables can be implemented and provides a list of evidence types.

List of observables and the associated evidence types

Analyses and assessments	
Subcategory	**Types of evidence**
(Purpose) definition	written documentation – description, justification
Risk	written documentation – analysis, test results – results documentation, justification
Metrics & thresholds	written documentation – analysis, test results – results documentation
Assessments	written documentation – description, justification

Organisational measures	
Subcategory	**Types of evidence**
Governance	written documentation – guideline/policy, process documentation, plan, description
System-related processes	written documentation – process documentation, plan, description, non-written documentation – screenshots, computer script/code
User instructions	written documentation – process documentation, plan, description
Training materials	written documentation – plan, description

Technical measures	
Subcategory	**Types of evidence**
Data	Data – sample, data – data documentation, data – full data access, non-written documentation – screenshots, non-written documentation – computer script/code
Testing	Test results – test execution, test results – results documentation, test results – screenshot
Models	Model features – model documentation, model features – model access
Deployment	written documentation – process documentation, plan, description, non-written documentation – logs, non-written documentation – screenshots, non-written documentation – computer script/code

Table 3: List of observables and the associated evidence types

Indicators in the "analyses and assessments" category require evidence and the associated observables primarily in the form of documentation, descriptions of the AI system and its components (e.g. system or architecture diagrams), and logs or other process artefacts. If the same documentation is relevant for different criteria or indicators, it does not need to be provided multiple times but can be referenced instead.[10]

The "organisational measures" category uses corresponding indicators and downstream observables to assess the extent to which an organisation has taken measures necessary for the quality-oriented development and implementation of an AI system. This category includes processes, training courses or guidelines, for example.

The third category describes "technical measures". For some indicators and their observables from the "measures" type, evidence of specific system properties is required. This can be generated by evaluating suitable metrics, tests and analyses of the AI system and its components, for example.

The evidence provided in this way is predominantly procedural, creating transparency. If it covers several criteria, it can be referenced instead of duplicated. If this evidence is insufficient for assessing technical system properties, an in-depth, measurement-based assessment follows, and the results are documented in a comprehensible and reproducible manner. Where assessment requirements address technical aspects, auditees should include representative and reproducible test results (e.g. dataset/split description, test configuration, versions and seed/run information). The assessment procedure does not stipulate any particular metrics for this, as different ones may be relevant depending on the AI system and application context. The Quality Standard includes a **Test Method Catalogue**, which serves as an aid (see appendix). The technical Test Method Catalogue focuses on procedures that substantiate AI system properties by mapping the specific properties of data or AI components to test outputs. Each technical assessment method in the collection is linked to the appended **Assessment Catalogue** via indicator IDs. This means auditees receive an overview of established procedures that can be used to generate evidence for an indicator. However, specific technical assessment methods must still be selected based on an analysis of the AI system to be assessed, particularly with regard to its task and area of application (see VE1.2 and VE1.3 im **Assessment Catalogue**, for example). Therefore, it is not possible to specify the use of certain technical assessment methods.

3.5 Evidence validation

In the next step, the rating made by the auditee is validated. This involves confirming the existence and plausibility of the measures that the evidence is intended to prove. The validation is performed in relation to the selected observable levels, taking into account the evidence provided (see 3.4).

As described in Section 2 the required degree of validation for each indicator is determined directly by the observable rating to be validated.

Where applicable and depending on the assessment depth, evidence for technical measures must include reproducible tests under identical configurations (e.g. data/splits, seeds and hyperparameters). The **Test Method Catalogue** provides methodological assurance. This can also be used by independent auditors to validate evidence collected systematically (e.g. expected value or acceptable range).

10 Example: For the requirement "Have target groups been defined with regard to fairness?", an artefact could be the "Fairness Analysis Documentation v2.pdf", with the link provided in "Table 2: Potentially Disadvantaged Groups". Additionally, risk analysis indicators can refer to a single risk analysis document.

3.6 Overall assessment

To complete a full assessment, an overall assessment of the AI system is performed after the observables have been rated. To achieve this, the rating is aggregated at the criterion level and systematically compared with the corresponding protection needs (see Figure 5).

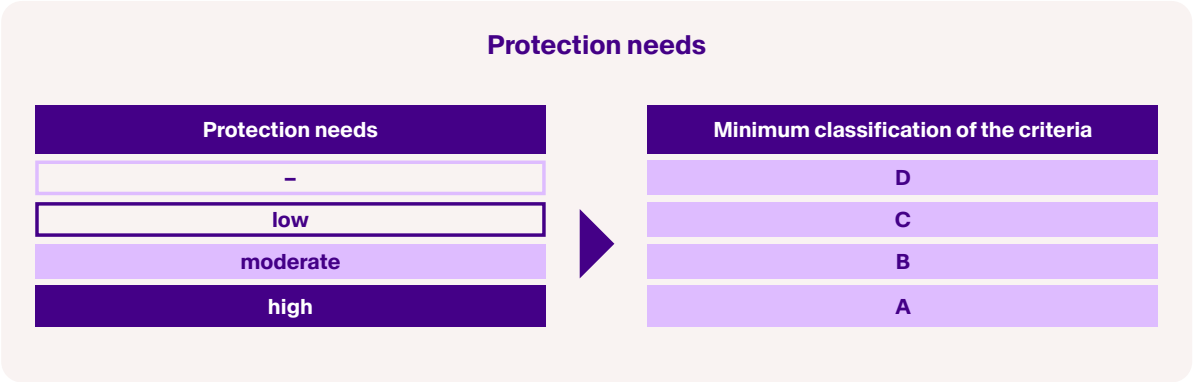


Figure 5: Protection needs as implicit minimum levels in the overall assessment, which determine the degree to which the minimum standard for an AI system in a specific use case is confirmed.

Thus, the overall assessment checks whether the rating determined for the AI system, aggregated on the basis of the criteria, meets the protection needs for each criterion. The minimum levels required, as determined by the protection needs, must be achieved or exceeded.

When comparing the aggregated rating with the protection needs analysis, the different levels of protection needs (high, moderate or low) serve as the minimum required level. This level must be achieved for each (applicable) criterion in the rating to pass the assessment. For example, a high protection need would require the aggregated criteria level to correspond to an "A". For a moderate protection need, at least the aggregated criteria level of "B" or higher must be achieved.

If the minimum level is achieved for all criteria, the quality assessment is passed overall. If the minimum level is not met for all criteria, the quality assessment fails. If the minimum level is exceeded in individual criteria, the assessment is considered to have been exceeded.

Instructions on the exact **Process flow of the overall evaluation** can be found appended to the Quality Standard.

3.7 Assessment report

Once the assessment has been completed, the results are summarised in an assessment report. The assessment report contains a formal declaration of the accuracy of the statements made and clearly reflects the most important aspects of the assessment:

- Information about the AI system, including a description of the application context, the data used and the (AI) components, and how they interact.
- Identified protection needs, broken down by criteria and accompanied by an explanation if the criterion was assessed as "not applicable".
- The level of assessment requirements achieved in each case is compared with the protection needs at the criterion level.
- Selected quality measures and precautions to safeguard the AI system that underpin the rating.
- Information about the assessment depth, i. e. the technical tests and the degree of validation

performed.
- Explanations and comments that facilitate the interpretation of the results or contextualise them.
- Explanation of which criteria have been met, not met or exceeded.
- Persons responsible for the assessment.

The presentation of the results must comply with the (see appendix). The assessment report is only valid if signed by the persons responsible for the assessment and any validations. An accompanying declaration confirming the accuracy of the statements made is mandatory as part of the **MISSION KI** Quality Standard.

3.8 Validity monitoring

Specific measures and provisions must be in place to ensure the validity of the assessment results (see Validity 1.4) is maintained.

The assessment statement is generally only valid for the clearly defined version of the AI system available at the time of the assessment. This should be evident from the documentation and clear versioning within the context of the assessment report.

The assessment statement ceases to be valid as soon as:

- The original intended use changes.
- External conditions (e. g. concept or data drift) change the assessment statement.
- The technical implementation of the AI system (software or hardware) changes significantly.

The following non-significant changes do not affect the validity:

- Equivalent redistribution of responsibilities.
- Changes to hardware or software components (e. g. a new generation of GPUs or a new version of a software library, etc.) that can be proven to have no significant impact on the assessed quality dimensions (see Monitoring).

4. Appendices

4.1 Glossary

4.1.1 Generic terms

Quality dimension
A quality dimension refers to a desirable, high-level, and general property of an →AI system, the presence of which can be indirectly tested – i.e., through concretizations and operationalizations (→criterion, →indicator, →observable) and which, together with other properties of equal rank, defines the overall quality of this →AI system when a corresponding protection needs exists. The model consisting of quality dimension, →criterion, →indicator, and →observable is collectively referred to in the literature as "VCIO" ("Value–Criterion–Indicator–Observable"). Within **MISSION KI** the term "value" used there is largely replaced by "quality dimension", with essentially the same meaning.

Criterion
To determine whether an →AI system meets individual →quality dimensions, criteria are specified. Criteria thus represent a concretization of a quality dimension toward operationalization and observable specific circumstances, as well as toward concrete risks or protection needs.

Indicator
It is usually not possible to measure directly whether individual →criteria are met. To assess this, the →criteria are further broken down into indicators. Indicators are derived from conditions that contribute to the fulfillment of higher-level →criteria at an abstract level. Thus, indicators provide information about specific properties of an →AI system that are decisive for the qualitative fulfillment, non-fulfillment, or degree of fulfillment of a criterion (VDE SPEC 90012: 2.28). Maximum-value indicators are indicators that play a special role in the evaluation process. Due to their criticality for the determined protection need, their classification represents a binding upper limit for the average rating.

Observable
To assess the fulfillment of the →indicators, observables are defined that indicate, in graded levels, the extent to which an →indicator is met. The levels of the observables are predefined and reflect the degree of goal achievement. An observable is thus a measurable value used to determine the state or properties of a system based on an indicator (VDE SPEC 90012: 2.36).

Examples: quality of dataset documentation or effectiveness of risk prevention.

Test
Attempt to determine specific properties, performance, or comparable characteristics.

4.1.2 Individual quality dimensions

(AI-specific) cybersecurity
Resilience of →AI models, →AI components, and →AI systems against AI-specific, malicious external interference and manipulation occurring via general telecommunication networks.

Data governance
Handling, processing, and safeguarding of data used throughout the lifecycle of an →AI system, with the aim of ensuring that the data meet high quality and integrity standards and are used in compliance with applicable regulations on privacy protection and →data protection (cf. Recital 27 AI Act).

Data quality
Property of the training, validation, and test data of an →AI system with respect to their factual correctness, completeness, and freedom from unjustified bias.

Data protection
Preservation and restricted access of certain documented personal data.

Human oversight and control
Property of an →AI system, including its embedding in the application context, concerning the ability of a human individual with appropriate professional competence to adequately observe and modify the behavior and/or functioning of this →AI system both in principle and during ongoing operation, and, if necessary, to terminate it.

Non-discrimination
Characteristic of an open process carried out by an →AI system, in which multiple human individuals are treated in comparison to one another, and which, from a legal perspective, is free from disadvantaging any human individual on the basis of a legally protected characteristic.

Transparency
Property of an →AI system that is →explainable and →interpretable. Within the scope of this quality standard, "transparency" also includes documentation of the properties of the →AI system.

Reliability
Property of an →AI system that exhibits sufficient →performance, sufficient →robustness, and allows for adequate →monitoring.

4.1.3 Individual criteria

Explainability
Property of an →AI system concerning the fundamental comprehensibility and traceability of its functionality, behavior, and output for human experts as well as affected individuals and →users. Explainability is often understood as a property of an →AI model or →AI system that is measured locally and independently of the system design ("post hoc")

Interpretability
Property of an →AI model to be, in principle, as directly traceable and understandable to experts as possible in its model parameters, weights, or other (mathematical) properties. Interpretability is often explicitly provided as part of the model architecture design, in contrast to the deliberate choice of opaque "black box" models (e.g., choosing an interpretable decision tree instead of a neural network for a classification task).

Performance
Property of an →AI system concerning its ability to achieve its intended goals and purposes as completely as possible.

Monitoring
Procedure in which deviations between observable actual states and intended target states are detected during the operation of an →AI system.

Robustness
Ability of an →AI system to maintain its regular and expected behavior and functionality as effectively as possible, even in the presence of non-malicious, adverse, disruptive, or faulty inputs or external influences.

Traceability
Property of an →AI system concerning the traceability of the consecutive sequence of all decisions that enter or have entered the →AI system throughout its entire lifecycle.

4.1.4 Horizontal concepts

Documentation
Systematic recording, collection, storage, and provision of various types of information in accordance with legal, internal, or external requirements.

Fairness
Characteristic of an open process carried out by an →AI system, in which multiple human individuals are treated in comparison to one another, and which, from a legal perspective, is free from disadvantaging any human individual on the basis of a legally protected characteristic, while also aligning with the notions of fairness held by designated individuals.

Safety
Property of an →AI system concerning the safety of the system for human individuals with respect to risks to life, limb, and health, as well as for objects with respect to damage, under intended functional normal operation.

Security
Resilience of an →AI system against malicious external interference and manipulation.

4.1.5 Further terms

Compatibility
Property of the →quality standard depending on its compatibility and consistency with other AI regulations, such as the European AI Act.

Application domain
The entirety of possible input data relevant to an →AI system. It encompasses the contexts in which an →AI system can be applied or used. In certain contexts, the term "application domain" may be used synonymously with "Operational Design Domain" (ODD) to define the specific conditions and parameters under which an →AI system operates effectively.

Soundness
Property of the quality standard depending on the →test depth and the objectivity of the assessment, whereby the degree of validation, external audits, and comparable factors correspondingly increase it.

Affected persons
Natural persons who are influenced or affected by an →AI system without necessarily interacting directly or actively with the system. The difference from →users lies in the type of interaction: while →users operate the →AI system directly and actively, for example, as end users or operators, affected persons stand in a more indirect and passive relationship to the →AI system but may nonetheless be influenced by its decisions or functions.

Efficiency
Property of the quality standard depending on the temporal, personnel, and financial effort required, while simultaneously ensuring a high level of quality.

AI provider
A "provider" is a natural or legal person, public authority, agency, or other body that develops or has developed an →AI system or a →general-purpose AI model and places it on the market under its own name or trademark, or puts the →AI system into service under its own name or trademark, whether for payment or free of charge (cf. Art. 3(3) AI Act)

AI deployer
A "deployer" is a natural or legal person, public authority, agency, or other body that uses an →AI system under its own responsibility, unless the →AI system is used in the course of a personal and non-professional activity (cf. Art. 3(4) AI Act)

AI component
An "AI component" comprises an implemented →AI model, possibly together with the methods directly related to the pre- or post-processing of this model's inputs and outputs, as well as their interfaces.

Example: an AI image recognition model, including the methods for preprocessing the images, which takes raw image or video data as input and outputs a statement on whether a human is visible in the image.

AI life cycle
Development of a system, product, service, project, or other human-made entity that uses AI, from conception to decommissioning (based on and extending ISO/IEC 22989:2022)

AI model
An "AI model" includes only the functional, AI-specific parameters, potentially including weights and biases (inferential input-output mappings), as well as the architecture. It does not include the further implementation and integration, which are only covered by the term →AI component.

Example: a neural network for image processing that receives numerical values assigned to pixels as input and outputs the probability that a human is visible in the image.

AI system
An "AI system" is a machine-based system designed to operate with varying levels of autonomy and capable of adaptation after deployment. Based on the inputs it receives, it infers how outputs such as predictions, content, recommendations, or decisions are generated to achieve explicit or implicit objectives, and these outputs may influence physical or virtual environments (cf. Art. 3(1) AI Act) Within the scope of the present quality standard, an AI system is understood in a technical sense as a functional integration of one or more →AI component(s) and non-AI components, aimed at a specific purpose and concrete application context. Multiple AI systems may be interconnected to form a larger AI system. Components with which the AI system can interact, but which are not essential for its functioning or are interchangeable, are not considered part of the AI system.

Network
A "network" is a connection of at least two computers or other electronic devices that enables the exchange of data and the use of shared resources.

Low-threshold accessibility
An offering or service characterized by low required effort (e. g., short evaluation duration) while simultaneously ensuring high benefit (high quality guaranteed).

User
Natural persons who interact directly and actively with an →AI system, either as end users who use the →AI system for personal or commercial purposes or as operators who use the →AI system in a professional context. The term includes both those who operate the →AI system and those who use it to achieve specific objectives. Depending on the context, this may involve the use of →AI systems in everyday applications or in specialized professional scenarios.

Proprietary data
Data that are subject to ownership rights, including licensed and copyright-protected data.

Assessor
Independent internal or external natural persons who must not, at any point, be involved in the development of the →AI system or related structures. As natural persons, they act on behalf of legal entities, for example, as internal →auditors of the →AI provider or →AI operator, or as external →auditors representing an independent party.

Auditee
The →AI provider or →AI operator of the →AI system under evaluation is referred to as the auditee. The auditee is represented by one or more natural persons.

Test method
Methodical procedure for collecting individual or multiple thematically related pieces of test evidence and evaluating them within the context of the audit process. A test method may include both manual and automated components. Technical test methods refer to those in which the collection of test evidence primarily relies on technical procedures and/or tools.

Test depth
The test depth defines the level of effort with which the assessment is conducted, determines the degree of assurance and plausibility of assessment statements, and specifies the roles required to validate the results within the audit process. The test depth consists of three key characteristics of the assessment: 1) The level of detail of the measures according to the test requirements from the audit catalog, 2) The evidence to be provided, and 3) The degree of validation.

Audit process
Process within the audit that includes the determination of the →test depth, the methodology for collecting evidence, the comparability, and the evaluation of the →criteria.

Quality standard
A standard is a document that defines a procedure, making it possible to objectively verify whether a test object conforms to the criteria established in the standard. A quality standard is a standard whose criteria allow conclusions to be drawn about the quality of the test object.

Comparability
Property of the →quality standard depending on the uniformity of assessments and the replicability of tests, with the goal of ensuring objectivity.

Accessibility
Property either of an assessment statement, indicating its comprehensibility, or of an assessment approach, describing the extent of its applicability to specific target groups.

4.2 Use case description template

General information

	Question	Use case details
1	Name of the use case	
2	Describe the use case (maximum 300 words)	
3	Which task is solved by the AI system? (maximum 300 words)	
4	What is the input range of the AI system, i.e., in which situations/with which inputs should the AI system function[11]?	
5	What are the limits of the AI system's input range, i.e., in which situations/with which inputs does the AI system only work to a limited extent or not at all?	
6	What is the output range of the AI system?	
7a	For which user group is the use of the AI system conceivable? Are there any requirements in terms of training or knowledge for this user group?	
7b	Which groups of people may be directly affected (positively and/or negatively) by the AI system's output[12]?	
8a	Is it envisaged that humans will be involved in the deployment of the AI system?	☐ Yes ☐ No
8b	Are humans involved in overseeing the AI system?	☐ Yes ☐ No
8c	If so, do they need to actively intervene to exert influence, or is human scrutiny a necessary part of the process?	☐ Yes ☐ No
9	Are there special regulatory requirements relating to the usage context[13]?	

11 The input range may be restricted, for example, by general conditions or technical requirements. In the context of autonomous driving, this is also referred to as an Operational Design Domain (ODD). One example would be object recognition that only works under certain lighting conditions.

12 This refers to the groups of people who are affected by undetected incorrect output from the AI system, e.g., applicants in an application tool or readers in automatically generated newspaper articles.

13 For example, requirements resulting from the rating according to the AI Act or requirements from sectoral regulation. Proof of fulfilment of requirements can serve as evidence in the further course of the test.

General information

	Question	Use case details
10	Do you deploy your AI models on-premise/ locally or in the cloud?	☐ Cloud ☐ on-premise ☐ local ☐ hybrid
11	Is at least one model being trained further during deployment?	☐ Yes, the model continues to train continuously ☐ Yes but further training must be triggered manually ☐ No
12	Architecture	☐ Large Language Model ☐ Decision tree ☐ Convolutional NN ☐ Genetic algorithm ☐ Random forest ☐ Regression ☐ SVM ☐ Clustering ☐ Others

4.3 Protection needs analysis

In general, the protection needs analysis is an examination of the protection needs with regard to the individual quality dimensions and their associated criteria. It is used to pre-filter the quality dimensions relevant for the respective application at criteria level and to determine a target value to be achieved for the subsequent assessment.

The background to this is that, depending on the task and area of application of an AI system, not all quality features are equally important. Even within a feature, individual points can have different weightings when assessing the quality of the AI system.

In this way, the protection needs analysis makes it possible to efficiently identify relevant quality dimensions at criteria level and determine the level of the corresponding protection needs.

At the same time, the protection needs analysis is a practicable way to achieve compatibility with the European Artificial Intelligence Act (AI Act). For high-risk AI systems in particular, but also for AI systems in general, this is based on their intended purpose and considers the combination of the AI system with the respective task and application context. The object of the protection needs analysis is therefore the AI system in its intended application context ('use case').

In addition to the information in the main body of this document, the following points are explained in more detail in this annex: firstly, some relevant definitions for the internal system of protection needs analysis are given (4.3.1) and the trade-off between risk assessment and protection needs analysis is discussed (4.3.2). After a systematic detailed description of the protection needs analysis (4.3.3) and the protection objectives (4.3.4) the application and implementation (4.3.5) of the framework is finally explained.

4.3.1 Definitions

To describe the factual constellation of the protection needs analysis, the following terms are defined in advance for the internal aspects of this document:

- **Danger** refers to a situation in which there is a possibility of damage occurring.
- **Risk** describes the product of the probability of damage occurring and the extent of the expected damage.
- **Protection** means the defence against, mitigation or prevention of damage or a hazard.
- **Protection need** is the appropriate or necessary level of protection for a protection objective and is based on the extent of the damage threatened in the event of violations.
- **Protection objective** means the respective property on the part of the entities concerned with regard to regulation or assessment. The protection objectives for the purpose of this protection needs analysis are the values set out in Art. 1(1) AI Act (health, safety and the fundamental rights enshrined in the EU Charter, including democracy, the rule of law and environmental protection).

4.3.2 Risk assessment vs. protection needs analysis

After careful consideration and in-depth discussions as part of the project, the protection needs analysis is used instead of a risk assessment. This modification is based on several arguments.

The first thing to mention here is the difficulty of the concept of risk in the context of AI. Based on the common definition described above, the concept of risk encompasses the consideration of probabilities. This includes both the probability of occurrence of a loss and the probability of failure

of the entity under review. So far, however, there is no experience in the AI sector to the extent and breadth required. It is currently very difficult to determine the respective probabilities, particularly with regard to scalability.

Furthermore, a risk assessment is accompanied by a significantly more complex assessment due to the combined consideration of the probability of damage and the extent of damage. In contrast, a protection needs analysis, as described below, can be carried out much more efficiently.

A protection needs analysis allows the quality dimensions relevant to the AI system to be pre-filtered at criteria level, as the questions in the protection needs analysis make it possible to determine which of the quality dimensions and the criteria grouped under them are specifically applicable to the AI system being analysed. This significantly increases the efficiency of the assessment process and consequently also the attractiveness of the test product.

For these reasons, a protection needs analysis is carried out instead of a risk assessment. This reduces the assessment effort and at the same time addresses the respective protection needs for each quality dimension at criteria level. The basic guiding principle for determining the protection need is to include questions on hazards and risk factors if these actually have an influence on the amount and type of damage. Only so-called "first order effects" are considered as part of the protection needs analysis. This is due to the fact that it is not possible to comprehensively and efficiently cover a more extensive recording of subsequent effects in view of the resulting possibilities.

4.3.3 Systematic description of the protection needs analysis

The following overall constellation must be considered for a systematic understanding of the protection needs analysis: Through its specific constitution and intended purpose, an AI system has the ability to influence other entities (people, animals, things, society, the environment as a whole). These entities have properties that are ascribed intrinsic value. These include fundamental values such as life, health, integrity, survival and freedom from harm, but also the values formulated in terms of fundamental rights, such as gender equality.

Depending on its constitution, an AI system in its basic constitution, but above all due to its intended use, potentially poses a threat to these properties of the aforementioned entities in specific respects and thus a potential for damage. In the intended prevention of this damage, the basic values mentioned become protection objectives. Depending on the extent of the threat and the potential damage, the AI system requires a certain level of a protection need. As explained in Section 3, a protection needs analysis is based only on the amount and extent of the damage, but not on precisely quantified probabilities of occurrence.

The existing hazard and damage potential and the selected protection objectives determine the properties that the AI system under investigation must have to minimise these hazards. These characteristics of the AI system are ensured by the corresponding assessment of the various quality dimensions (such as VE: reliability, CY: AI-specific cybersecurity etc.). Insofar as the protection need can be graduated per quality dimension and criterion, the degree to which these quality dimensions are required on the part of the AI system also varies.

In accordance with this overall constellation, the protection needs analysis consequently determines the protection needs for each individual quality dimension at criteria level, depending on the selected protection objectives. It therefore uses targeted questions to determine how high the protection need is for each individual quality dimension at criteria level. It takes into account the specific AI system in its context of use – for example, the criterion "fallback plans and functional safety" of the quality dimension VE (reliability). As a target state, this statement forms a minimum value that must be achieved in the subsequent tests for this quality dimension at criteria level.

Two further decisions were made to ensure that the protection needs analysis (and subsequent review) was carried out as efficiently as possible: firstly, with regard to the applicability of the quality dimensions at criteria level and, secondly, with regard to the graduation of the protection needs.

As part of the initial decision, before the protection need is determined in detail, it is checked whether the criteria of a quality dimension can be meaningfully applied to the AI system under review. For example, the criterion "avoidance of unjustified discrimination and distortion" of the quality dimension ND (non-discrimination) may only be applicable to a technical system for predictive maintenance to a limited extent. Only if this applicability is given in principle is the protection need for the criterion in question of a quality dimension determined in detail. This is to ensure that, both in the protection needs analysis itself and when conducting the assessments, only those quality dimensions are tested at criteria level that can be meaningfully applied to the specific AI system. It should also be ensured that these quality dimensions correspond to the given protection need. Exceptions to this approach are the criteria "performance and robustness" and "fallback plans and functional safety" of the quality dimension VE (reliability), the criterion "data quality" of quality dimension DA (data quality, protection and governance) and the criterion "traceability and documentation" of the quality dimension TR (transparency) – these are considered so fundamental to the quality of AI systems that they are always checked.

With regard to the graduation of the protection need, it was decided to take a middle course between a purely binary distinction (need for protection: yes/no) and excessive granularity in the interests of efficiency. For the purposes of this decision, the protection need is determined in three levels ("Low", "Moderate", "High"). If questions on the protection need can only be answered with "Yes" or "No", a high protection need is assumed if the answer is "Yes". One exception is the "health" protection objective. A moderate protection need is assumed here with "Yes" to differentiate from the protection objective of "life and limb".

After answering the individual questions, the protection need is aggregated at the level of the individual criteria of a quality dimension according to a specific procedure. As stated, this formulates the minimum value that must be achieved for each individual criterion of a quality dimension to successfully pass the test. The exact procedure for aggregation is described in Process flow of the overall evaluation (see 4.5).

Overall, the protection needs analysis uses an efficient question format to determine the protection needs for each individual quality dimension to be assessed at criteria level and with a view to the protection objectives listed below. The protection needs determined form the target state for the subsequent assessment of the AI system at the level of the individual criteria for each quality dimension.

4.3.4 Protection objectives

The protection objectives are derived, in the sense of compatibility with the AI Regulation, from Art. 1(1) (AI Act). This article lists health, safety and the fundamental rights enshrined in the EU Charter, including democracy, the rule of law and environmental protection. The AI Act requires a high level of protection to be guaranteed for these protection objectives. To keep the assessment effort accessible and low-threshold, the protection objectives are categorised and recorded in manageable categories to avoid unnecessary depth of detail. These categories include life and limb, health, fundamental rights, property and objects, environment, protection of personal data, human dignity and non-discrimination.

4.3.5 Carrying out the protection needs analysis

The template in 4.3.6 is used to carry out the protection needs analysis. The list of questions contained in the template is worked through for each criterion and each quality dimension to determine whether the AI system has potential hazards for the respective protection objectives. The basis for the answer is the use case description. If necessary, further information can be obtained from the auditee for individual questions of the protection needs analysis.

Each set of questions per quality dimension is introduced by application questions (German: Applikationsfrage; AF). In principle, these check which criteria per quality dimension are actually applicable to the given AI system in terms of protection needs. Criteria can have either none (provided one of the exceptions applies and the criterion is always applicable, see 04), one or more AF. If a criterion has more than one AF, it is considered "applicable" if at least one of the AFs was answered with "Yes". The general applicability has no influence on the level of protection need. If a criterion based on this AF is not applicable to the given AI system, the level of protection need in accordance with the explanations in Section 04 cannot be meaningfully determined and the answers to the other questions on this criterion are not applicable. This increases the efficiency of the test in two ways. On one hand, the assessment effort is reduced as no further questions relating to the criterion need to be answered. On the other hand, the corresponding criterion can be disregarded in the further assessment of the AI system.

Following the AF, so-called basic questions (German: Grundfrage; GF) are asked. Answers to these questions specify a mandatory minimum level with regards to the level of protection need. If there are several GFs, the highest protection need determines the protection need of the entire criterion. As a result, the GFs carry particular weight in this way: they relate to protection needs that, for example, in case of personal injury, cannot be 'offset' or reduced by other, lower protection needs.

The third type of question is the extension question (German: Erweiterungsfrage; EF). The answers to these questions form the average for each criterion per quality dimension. Unlike the GFs, the EFs determine an "average overall impression" of the protection needs for protection objectives whose relevance is assessed as relatively low, but nevertheless not irrelevant in comparison to the protection needs surveyed via the GFs. If < .5 in the first decimal point, the average is rounded down; if > = .5, it is rounded up.

For each question, a column is listed with the respective protection objective category. The selection of the protection objective categories is based on the potential damage that can occur if the requirements for the criterion of a quality dimension are violated. If a question cannot be assigned to a clear protection objective category, it is listed under the protection objective category "general". Individual questions are relevant for rating several criteria or quality dimensions and therefore appear several times.

To adequately answer the questions, the AI system as such must be considered, i. e., both with regard to its intended purpose and its application context, but without taking into account implemented measures or similar. Accordingly, the protection need is determined with regard to the intended use as such. The existence of any necessary mitigation, safety or protective measures will be queried as part of the subsequent VCIO assessment.

The answers to the individual questions are aggregated according to a specific procedure for the overall protection need of a criterion per quality dimension. This aggregation process is closely interlinked with the protection needs analysis in the interests of efficiency. The aggregation of the protection needs, in conjunction with the implementation of the protection needs analysis, is organised as follows for each criterion per quality dimension:

Step 1: Answering the AF. If these consistently show that there is no applicability, the protection needs analysis for this criterion can be ended with step 7 of this description. The same applies to the exceptions (see. 04), which concern criteria that are always applicable and for which there is therefore no AF. If applicable, continue with step 2.

Step 2: Answering the GF.

Step 3: Determination of the maximum value for the GF. If one of the basic questions is answered with "3" or "High", the protection needs for this criterion should be set to "High" overall and the protection needs analysis, for the purpose of efficiency, should be ended for this criterion after the corresponding question with step 7 of this description. (If an overall impression of the protection need is desired, the analysis can also be carried out further).

Step 4: Answering the EF.

Step 5: Determination of the average value for the EF in accordance with the above provisions on the EF.

Step 6: If: **level from GF > = average from EF,** the protection need for this criterion is equal to the highest protection requirement from the GF.

Schritt 7: If: **level from GF < average from EF,** the protection need for this criterion is equal to the average of all GF and EF, again using the above provisions for averaging. If there is no GF, the average of EF corresponds to the protection need for the criterion.

Step 8: This determines the protection need for the criterion being analysed. It must be noted at the top of the tab for the analysed quality dimension of the criterion in the field provided.

Step 9: Hereinafter, continue with next criterion of the quality dimension or (if all quality dimension criteria have already been processed) **continue** to the next tab with criteria for the next quality dimension.

4.3.6 Protection needs analysis template

Protection needs analysis

MISSION KI
Nationale Initiative für Künstliche Intelligenz und Datenökonomie

Overview of tabs

Rider name	Contents
DA	Questionnaire on the quality dimension of data quality, data protection and data governance
ND	Questionnaire on the quality dimension of non-discrimination
TR	Questionnaire on the quality dimension of transparency
MA	Questionnaire on the quality dimension of human oversight and control
VE	Questionnaire on the quality dimension of reliability
CY	Questionnaire on the quality dimension of AI-specific cybersecurity

This document is part of the MISSION KI Quality Standard. ©acatech – National Academy of Science and Engineering.
This work is licensed under a Creative Commons Attribution-NoDerivs 4.0 International Licence (CC BY-ND 4.0).
https://creativecommons.org/licenses/by-nd/4.0/deed.en

Protection needs analysis

Quality dimension: Data quality, data protection and data governance (DA)

Overall assessment per criterion

Protection requirement criterion “Data quality”:			1 / 2 / 3	always applicable
Protection requirement criterion “Protection of personal data”:			1 / 2 / 3 / not applicable	
Protection requirement criterion “Protection of proprietary data”:			1 / 2 / 3 / not applicable	

Question index	Question	Protection goal category	Criterion	Assessment	Optional: Comment	Notes and examples
	Application questions	**Protection goal category**	**Criterion**	**Assessment**	**Optional: Comment**	**Notes and examples**
DA-Z12	Does the AI system process or generate personal data, or was this data used in the training of the AI model(s)? yes = 3 no = 1	General	Protection of personal data			"Processing" explicitly does not mean steps to remove, exclude, or anonymise this data for further processing, provided that it is not stored.
DA-Z13	Does the AI system process or generate proprietary data, or was this data used in the training of the AI model(s)? yes = 3 no = 1	General	Protection of proprietary data			"Processing" explicitly does not mean steps to remove, exclude, or anonymise this data for further processing, provided that it is not stored.
	Application questions	**Protection goal category**	**Criterion**	**Assessment**	**Optional: Comment**	**Notes and examples**
DA-Z16	What is the level of protection required with regard to the criterion "performance and robustness" (VE)? low = 1 moderate = 2 high = 3	General	Data quality			Automatic function can be inserted to take over the assessment of the need for protection.
DA-Z17	How high is the protection requirement with regard to the criterion "protection of personal data" (DA)? low = 1 moderate = 2 high = 3	General	Data quality			Automatic function can be inserted to take over the assessment of the need for protection.
DA-Z18	How high is the need for protection with regard to the criterion "avoidance of unjustified discrimination and distortion" (ND)? low = 1 moderate = 2 high = 3	General	Data quality			Automatic function can be inserted to take over the assessment of the need for protection.
DA-Z19	Does the AI system process personal data in accordance with Article 9(1) GDPR? yes = 3 no = 1	General	Protection of personal data			Personal data based on Article 9(1) GDPR are data revealing a natural person's racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership.
DA-Z20	Does the AI system process genetic data, biometric data for the unique identification of a natural person, health data, or data relating to a natural person's sex life or sexual orientation? yes = 3 no = 1	General	Protection of personal data			
DA-Z21	Could inadequate protection or inadequate management of the data used or generated by the AI system – in a realistic worst-case scenario – cause fatal personal injury? yes = 3 no = 1	Body and life	Protection of personal data			Consider, for example, political persecution: in such a case, inadequate protection of sensitive data through access by relevant authorities can mean a risk to the life of the person concerned.
DA-Z22	Could inadequate protection or inadequate management of the data used or generated by the AI system – in a realistic worst-case scenario – cause health damage (physical health) to natural persons? yes = 2 no = 1	Health	Protection of personal data			One could, for example, think of attacks on politicians, also in this case against the backdrop of inadequate data protection.
DA-Z23	Can the AI system facilitate or strengthen the surveillance of natural persons, or does it directly serve the purpose of monitoring natural persons? yes = 3 no = 1	General	Protection of personal data			

Question index	Question	Protection goal category	Criterion	Assessment	Optional: Comment	Notes and examples
	Application questions	**Protection goal category**	**Criterion**	**Assessment**	**Optional: Comment**	**Notes and examples**
DA-Z26	Given the task and data basis, is the AI system capable of generating personal data? yes = 3 no = 1	General	Protection of personal data			
DA-Z27	Can the results of the AI system, in principle, be used to establish a link to a specific person in the data? yes = 3 no = 1	General	Protection of personal data			
DA-Z28	Has a Data Protection Impact Assessment (DPIA) for the AI system resulted in a high or critical rating? yes = 3 no = 1	General	Protection of personal data			
DA-Z29	How frequently is the AI system retrained or updated with new versions during operation? never = 1 occasionally = 2 often = 3	General	Data quality			Example boundaries: **never** = not even once **now and again** = monthly or less frequently **often** = more frequently than monthly
DA-Z30	Is personal data anonymized before being used for the AI system? all and always = 1 with slight to moderate limitations = 2 with major or severe limitations = 3	Protection of personal data	Protection of personal data			Example boundaries: **all and always** = Anonymisation (all data is completely anonymized) **with mild to moderate limitations** = Pseudonymisation (data is pseudonymised, but there is a risk of re-identification) **with major or severe limitations** = Data is neither anonymised nor pseudonymised, or the percentage is below 50%.
DA-Z31	What is the maximum potential financial damage to a company if intellectual property is not adequately protected? The potential financial damage serves here as a proxy measure of the damage to the affected party or parties. No to low financial damage = 1 moderate financial damage = 2 severe or catastrophic financial damage = 3	Property and things	Protection of proprietary data			Example boundaries: **no or low financial damage** = Damage amounting to up to 1% of the previous year's turnover or less than €1,000 **moderate financial damage =** Damage amounting to 1% to 5% of the previous year's turnover or between €1,000 and €50,000 **serious or catastrophic financial damage** = Damage amounting to more than 5% of the previous year's turnover or more than €50,000
DA-Z32	What is the maximum potential financial damage to a company if personal data is not adequately protected (cumulative damage)? The potential financial damage serves here as a representative measure of the damage to the affected individual(s). No to low financial damage = 1 moderate financial damage = 2 severe or catastrophic financial damage = 3	Property and things	Protection of personal data			Example boundaries: **no or low financial damage** = Damage amounting to up to 1% of the previous year's turnover or less than €1,000 **moderate financial damage** Damage amounting to 1% to 5% of the previous year's turnover or between €1,000 and €50,000 **serious or catastrophic financial damage** = Damage amounting to more than 5% of the previous year's turnover or more than €50,000

Protection needs analysis

Quality dimension: Non-discrimination (ND)

		Overall assessment per criterion	
Protection requirement criterion "avoidance of unjustified discrimination and distortion":			1 / 2 / 3 / not applicable

Question index	Question	Protection goal category	Criterion	Assessment	Optional: Comment	Notes and examples
	Application questions	**Protection goal category**	**Criterion**	**Assessment**	**Optional: Comment**	**Notes and examples**
ND-Z10	Do the results or behavior of the AI system affect natural persons differently, and is this unequal treatment linked to the presence of protected characteristics (race, ethnic origin, gender, religion or belief, disability, age, sexual identity) in these persons? In the case of a self-assessment, an explanation in the comment section (how exactly are natural persons affected?) is mandatory. yes = 3 no = 1	General	Avoidance of unjustified discrimination and distortion			Here, **affected persons** are natural persons who are influenced or affected by an AI system without necessarily interacting directly or actively with the system. The difference between affected persons and users lies in the nature of the interaction: While users operate the AI system directly and actively, for example as end users, affected persons have a more indirect and passive relationship to the AI system, but can nevertheless be influenced by its decisions or functions (see Glossary). **Protected properties** are (based on the German General Equal Treatment Act (AGG)): race, ethnic origin, gender, religion or belief, disability, age, sexual identity.
	Basic questions	**Protection goal category**	**Criterion**	**Assessment**	**Optional: Comment**	**Notes and examples**
ND-Z14	Does the AI system use data that reveals whether a natural person belongs to a group of people with protected characteristics? yes = 3 no = 1	General	Avoidance of unjustified discrimination and distortion			**Protected properties** are (based on the German General Equal Treatment Act (AGG)): race, ethnic origin, gender, religion or belief, disability, age, sexual identity.
ND-Z15	Does the AI system regulate access to services or activities that are essential to one's personality? yes = 3 no = 1	General	Avoidance of unjustified discrimination and distortion			
ND-Z16	Is the AI system involved in decision-making processes that significantly affect personal rights? yes = 3 no = 1	General	Avoidance of unjustified discrimination and distortion			
	Extension questions	**Protection goal category**	**Criterion**	**Assessment**	**Optional: Comment**	**Notes and examples**
ND-Z19	Could misuse (e.g., surveillance of individuals) of the AI system target groups that define themselves based on protected characteristics, or cause particularly severe damage? yes = 3 no = 1	General	Avoidance of unjustified discrimination and distortion			**Protected properties** are (based on the German General Equal Treatment Act (AGG)): race, ethnic origin, gender, religion or belief, disability, age, sexual identity.
ND-Z20	How many users and (through its use) affected individuals does the AI system have within a specific time period? Low number of users/affected persons = 1 moderate number of users/affected persons = 2 high number of users/affected persons = 3	General	Avoidance of unjustified discrimination and distortion			Examples of thresholds: **low number of users/affected persons per month** = 0-1,000 **moderate number of users/affected persons per month** = 1,001-10,000 **high number of users/affected persons per month** = >10,000 Here, **affected persons** are natural persons who are influenced or affected by an AI system without necessarily interacting directly or actively with the system. The difference to users lies in the type of interaction: While users operate the AI system directly and actively, for example as end users, affected persons have a more indirect and passive relationship to the AI system, but can be influenced by its decisions or functions (see Glossary).

Question index	Question	Protection goal category	Criterion	Assessment	Optional: Comment	Notes and examples
ND-Z22	Does the AI system, with regard to its scope of application, have the potential to reproduce legally prohibited discrimination, or is there reasonable cause to believe that it exacerbates such inequalities?? yes = 3 no = 1	Non-discrimination	Avoidance of unjustified discrimination and distortion			
ND-Z23	Does the AI system, due to its scope of application or its functionality, have the potential to produce outputs that may be prohibited by law (e.g., insults), or can it be reasonably assumed that the underlying (training/test) database contains such instances or content? yes = 3 no = 1	Human dignity	Avoidance of unjustified discrimination and distortion			

Protection needs analysis

Quality dimension: Transparency (TR)

Overall assessment per criterion

Protection requirement criterion “Traceability & Documentation”:			1 / 2 / 3	always applicable
Protection requirement criterion “Explainability & Interpretability”:			1 / 2 / 3 / not applicable	

Question index	Question	Protection goal category	Criterion	Assessment	Optional: Comment	Notes and examples
	Application questions	**Schutzzielkategorie**	**Criterion**	**Assessment**	**Optional: Comment**	**Notes and examples**
TR-Z11	Does the safety of the system decrease if there is a limitation in the explainability and interpretability of the system? Note: This is about understanding the "how" of the system and not about traceability and documentation. yes = 3 no = 1	General	Explainability & Interpretability			**Low level** can be understood as having effects that users and/or those affected would probably accept without complaint.
TR-Z12	Does the behavior or result of the AI system substantially affect the freedom of action of natural persons or their personal rights? yes = 3 no = 1	General	Explainability & Interpretability			**Substantial** can be understood as effects that restrict freedom of action or autonomy.
	Basic questions	**Protection goal category**	**Criterion**	**Assessment**		**Notes and examples**
TR-Z16	Could a restriction of the safe and appropriate use of the AI system caused by a lack of explainability and interpretability – in a realistic worst-case scenario – lead to fatal personal injury? yes = 3 no = 1	Body and life	Explainability & Interpretability			
TR-Z17	Could a restriction of the safe and appropriate use of the AI system caused by a lack of traceability and documentation – in a realistic worst-case scenario – lead to fatal personal injury? yes = 3 no = 1	Body and life	Traceability & Documentation			
TR-Z18	Could a restriction of the safe and appropriate use of the AI system, caused by a lack of explainability and interpretability, – in a realistic worst-case scenario – cause health damage (physical health) to natural persons? yes = 2 no = 1	Health	Explainability & Interpretability			
TR-Z19	Could a restriction of the safe and appropriate use of the AI system caused by a lack of traceability and documentation – in a realistic worst-case scenario – cause health damage (physical health) to natural persons? yes = 2 no = 1	Health	Traceability & Documentation			
TR-Z20	Does the AI system regulate access to services and activities that are essential to personality or personal scope of action? yes = 3 no = 1	General	Traceability & Documentation			Identical to question TR-Z21 (different criterion)
TR-Z21	Does the AI system regulate access to services and activities that are essential to personality or personal scope of action? yes = 3 no = 1	General	Explainability & Interpretability			Identical to question TR-Z20 (different criterion)
TR-Z22	Is the AI system involved in decision-making processes that affect fundamental rights or democratic procedures? yes = 3 no = 1	General	Traceability & Documentation			Identical to question TR-Z23 (different criterion) Example: Surveillance of natural persons
TR-Z23	Is the AI system involved in decision-making processes that affect fundamental rights or democratic procedures? yes = 3 no = 1	General	Explainability & Interpretability			Identical to question TR-Z22 (different criterion) Example: Surveillance of natural persons

Question index	Question	Protection goal category	Criterion	Assessment	Optional: Comment	Notes and examples
TR-Z24	Can the AI system – intentionally or unintentionally – be misused for another purpose if traceability and documentation are lacking, thereby causing societal harm or harm to the fundamental rights of natural persons? yes = 3 no = 1	General	Traceability & Documentation			For example, an AI system that regulates the flow of information on platforms like social networks could inadvertently amplify the spread of misinformation by prioritizing polarizing or sensationalist content. This could lead to social division, the rise of extremism, and the erosion of trust in public institutions.
TR-Z25	To what extent does the AI system interact with users in a way that – in a realistic worst-case scenario – could significantly and negatively restrict or influence their perception, decisions, or actions? Not at all or to a small extent = 1 To a moderate extent = 2 To a high extent = 3	General	Traceability & Documentation			Identical to question TR-Z26 (different criterion). Here, **users** means... Natural persons who interact directly and actively with an AI system, either as end users who use the AI system for personal or business purposes, or as users who apply the AI system in a professional context. The term encompasses both those who operate the AI system and those who use it to achieve specific goals. Depending on the context, this can include the use of AI systems in everyday applications or in specialized, professional scenarios (see Glossary).
TR-Z26	To what extent does the AI system interact with users in a way that – in a realistic worst-case scenario – could significantly and negatively restrict or influence their perception, decisions, or actions? Not at all or to a small extent = 1 To a moderate extent = 2 To a high extent = 3	General	Explainability & Interpretability			Identical to question TR-Z25 (different criterion). Here, **users** mean natural persons who interact directly and actively with an AI system, either as end users who use the AI system for personal or business purposes, or as users who apply the AI system in a professional context. The term encompasses both those who operate the AI system and those who use it to achieve specific goals. Depending on the context, this can include the use of AI systems in everyday applications or in specialized, professional scenarios (see Glossary).
TR-Z27	How frequently can situations arise in which the safety or usefulness of the AI system is limited due to a lack of validation or explainability of individual results? never = 1 occasionally = 2 often = 3	General	Explainability & Interpretability			
	Extension questions	**Protection goal category**	**Criterion**	**Assessment**	**Optional: Comment**	**Notes and examples**
TR-Z30	Is the use of the AI system mandatory or involuntary for users? yes = 3 no = 1	General	Explainability & Interpretability			Here **users** mean natural persons who interact directly and actively with an AI system, either as end users who use the AI system for personal or business purposes, or as users who apply the AI system in a professional context. The term encompasses both those who operate the AI system and those who use it to achieve specific goals. Depending on the context, this can include the use of AI systems in everyday applications or in specialized, professional scenarios (see Glossary).
TR-Z31	How frequently is the AI system retrained or updated with new versions during operation? never = 1 occasionally = 2 often = 3	General	Traceability & Documentation			Identical to question VE-Z24 (different dimension) Example limits: **never** = not even once **now and again** = monthly or less frequently **often** = more frequently than monthly
TR-Z32	How well can malicious errors caused by the AI system be detected as such? always easy and clear = 1 with slight to moderate effort = 2 with great or greatest effort = 3	General	Traceability & Documentation			
TR-Z33	Could a restriction of the safe and appropriate use of the AI system caused by a lack of traceability and documentation – in a realistic worst-case scenario – cause fatal harm to animals? yes = 3 no = 1	Body and life	Traceability & Documentation			
TR-Z34	Could a limitation of the safe and appropriate use of the AI system caused by a lack of explainability and interpretability – in a realistic worst-case scenario – cause fatal harm to animals? yes = 3 no = 1	Body and life	Explainability & Interpretability			

Question index	Question	Protection goal category	Criterion	Assessment	Optional: Comment	Notes and examples
TR-Z35	If users require expertise in artificial intelligence: To what extent can users or those affected by the AI system be expected to possess expertise in artificial intelligence? High level of expertise expected or not required = 1 Moderate level of expertise expected = 2 No or low level of expertise expected = 3	General	Explainability & Interpretability			Here **affected persons** are natural persons who are influenced or affected by an AI system without necessarily interacting directly or actively with the system. The difference to users lies in the nature of the interaction: While users operate the AI system directly and actively, for example as end users, affected persons have a more indirect and passive relationship to the AI system, but can be influenced by its decisions or functions (see Glossary). Here, **users** mean natural persons who interact directly and actively with an AI system, either as end users who use the AI system for personal or business purposes, or as users who apply the AI system in a professional context. The term encompasses both those who operate the AI system and those who use it to achieve specific goals. Depending on the context, this can include the use of AI systems in everyday applications or in specialized, professional scenarios (see Glossary).
TR-Z36	Due to the way the AI system functions and its tasks, is it fundamentally possible that the AI system could influence or direct the decisions or actions of users or affected persons in a way that they cannot fully understand? yes = 3 no = 1	General	Traceability & Documentation			Identical to question MA-Z23 (other dimension) **"Sufficiently permeated"** means that users/affected individuals cannot adequately understand how their decisions and actions are being influenced. Here **affected persons** are natural persons who are influenced or affected by an AI system without necessarily interacting directly or actively with the system. The difference to users lies in the nature of the interaction: While users operate the AI system directly and actively, for example as end users, affected persons have a more indirect and passive relationship to the AI system, but can be influenced by its decisions or functions (see Glossary). Here, **users** are natural persons who interact directly and actively with an AI system, either as end users who use the AI system for personal or business purposes, or as users who apply the AI system in a professional context. The term encompasses both those who operate the AI system and those who use it to achieve specific goals. Depending on the context, this can include the use of AI systems in everyday applications or in specialized, professional scenarios (see Glossary).
TR-Z37	In a realistic worst-case scenario, what is the potential financial damage that could occur if explanations or insights into the origins of AI system expenditures are lacking? The potential financial damage serves here as a proxy measure of the harm to the affected party(ies). No to low financial damage = 1 moderate financial damage = 2 severe or catastrophic financial damage = 3	Property and things	Explainability & Interpretability			Example boundaries: **no or low financial damage =** Damage amounting to up to 1% of the previous year's turnover or less than €1,000 **moderate financial damage** Damage amounting to 1% to 5% of the previous year's turnover or between €1,000 and €50,000 **serious or catastrophic financial damage** = Damage amounting to more than 5% of the previous year's turnover or more than €50,000
TR-Z38	In a realistic worst-case scenario, what is the potential financial damage that could occur if explanations or accountability for errors or malfunctions of the AI system cannot be provided retrospectively? The potential financial damage serves here as a proxy measure of the harm suffered by the affected party(ies). No to low financial damage = 1 moderate financial damage = 2 severe or catastrophic financial damage = 3	Property and things	Traceability & Documentation			Example boundaries: **no or low financial damage** = Damage amounting to up to 1% of the previous year's turnover or less than €1,000 **moderate financial damage =** Damage amounting to 1% to 5% of the previous year's turnover or between €1,000 and €50,000 **serious or catastrophic financial damage** = Damage amounting to more than 5% of the previous year's turnover or more than €50,000

Protection needs analysis

Quality dimension: Human oversight and control (MA)

Overall assessment per criterion

Protection requirement criterion “Human capacity to act”:			1 / 2 / 3 / not applicable
Protection requirement criterion “Human oversight”:			1 / 2 / 3 / not applicable

Question index	Question	Protection goal category	Criterion	Assessment	Optional: Comment	Notes and examples
	Application questions	**Protection goal category**	**Criterion**	**Assessment**	**Optional: Comment**	**Notes and examples**
MA-Z11	Can the AI system, in principle, perform its task completely without human intervention? yes = 3 no = 1	AF	Human oversight			**"In principle"** means the manner of use within the intended purpose. **"Task"** refers to the task of the entire AI system.
MA-Z12	Can the AI system have a direct impact on people? yes = 3 no = 1	AF	Human capacity for action			Examples: personal data during the training of the AI model(s), processing or generation of personal data, human recipients of the output
	Basic questions	**Protection goal category**	**Criterion**	**Assessment**	**Optional: Comment**	**Notes and examples**
MA-Z15	How high is the protection requirement for the criterion "performance and robustness" (VE)? low = 1 moderate = 2 high = 3	General	Human oversight			
MA-Z16	What is the level of protection required for the criterion "fallback plans and functional safety" (FS)? low = 1 moderate = 2 high = 3	General	Human oversight			
MA-Z17	Is the use of the AI system mandatory or involuntary for users? yes = 3 no = 1	General	Human capacity for action			Here, **users** mean natural persons who interact directly and actively with an AI system, either as end users who use the AI system for personal or business purposes, or as users who apply the AI system in a professional context. The term encompasses both those who operate the AI system and those who use it to achieve specific goals. Depending on the context, this can include the use of AI systems in everyday applications or in specialized, professional scenarios (see Glossary).
MA-Z18	How urgent is a quick fix or recovery in the event of a failure or malfunction of the AI system? No or low urgency = 1 moderate urgency = 2 high urgency = 3	General	Human oversight			A quick fix or restoration is urgent, for example, when the damage is continuously increasing due to ongoing failure or malfunction.
	Extension questions	**Protection goal category**	**Criterion**	**Assessment**	**Optional: Comment**	**Notes and examples**
MA-Z21	How can the alternative options for users or the task of the AI system be described if the desired functionality of the AI system is not available? Adequate and quickly available = 1 available with slight to moderate limitations = 2 available only with major and severe limitations = 3	General	Human capacity for action			Here, **users** are natural persons who interact directly and actively with an AI system, either as end users who use the AI system for personal or business purposes, or as users who apply the AI system in a professional context. The term encompasses both those who operate the AI system and those who use it to achieve specific goals. Depending on the context, this can include the use of AI systems in everyday applications or in specialized, professional scenarios (see Glossary).
MA-Z22	To what extent does the AI system interact with users in a way that – in a realistic worst-case scenario – could significantly and negatively restrict or influence users' perceptions, decisions, or actions? Not at all or to a small extent = 1 To a moderate extent = 2 To a high extent = 3	General	Human capacity for action			Here, **users** are natural persons who interact directly and actively with an AI system, either as end users who use the AI system for personal or business purposes, or as users who apply the AI system in a professional context. The term encompasses both those who operate the AI system and those who use it to achieve specific goals. Depending on the context, this can include the use of AI systems in everyday applications or in specialized, professional scenarios (see Glossary).

Question index	Question	Protection goal category	Criterion	Assessment	Optional: Comment	Notes and examples
MA-Z23	Due to the way the AI system functions and its tasks, is it fundamentally possible that the AI system could influence or direct the decisions or actions of users or affected persons in a way that they cannot fully understand? yes = 3 no = 1	General	Human capacity for action			Identical to question TR-Z36 (different dimension), **users** are natural persons who interact directly and actively with an AI system, either as end users who use the AI system for personal or business purposes, or as users who apply the AI system in a professional context. The term encompasses both those who operate the AI system and those who use it to achieve specific goals. Depending on the context, this can include the use of AI systems in everyday applications or in specialized, professional scenarios (see Glossary).

Protection needs analysis

Quality dimension: AI-specific cybersecurity (CY)

Overall assessment per criterion

Protection requirement criterion "General AI-specific cybersecurity":			1 / 2 / 3	always applicable
Protection requirement criterion "Resistance against AI-specific attacks":			1 / 2 / 3 / not applicable	

Question index	Question	Protection goal category	Criterion	Assessment	Optional: Comment	Notes and examples
	Applikationsfragen	**Protection goal category**	**Criterion**	**Assessment**	**Optional: Comment**	**Notes and examples**
CY-Z10	Is the AI system designed to be connected to a network, or is the AI system a network? yes = 3 no = 1	General	Resilience against AI-specific attacks			**Connection** refers to wireless connections. It includes both internal and external networks.
	Grundfragen	**Protection goal category**	**Criterion**	**Assessment**	**Optional: Comment**	**Notes and examples**
CY-Z14	How critical is the data processed by the AI system? low = 1 moderate = 2 high = 3	General	"General AI-specific cybersecurity" and "Resilience against AI-specific attacks"			Example in the context of company data (Information Classification Model): Internal Only = 1 Confidential = 2 Restricted = 3 Example in the context of personal data: General personal data (e.g., name, address, …) = 1 Personal data with behavioral relevance or similar (e.g., purchasing, browsing, movement, …) = 2 Special categories of personal data according to Art. 9 GDPR = 3
CY-Z15	To what extent are the damages that can be caused by the deliberate misuse or manipulation of the AI system – in a realistic worst-case scenario – repairable or recoverable? Completely fixable = 1 Partially fixable = 2 Slightly to not fixable = 3	General	"General AI-specific cybersecurity" and "Resilience against AI-specific attacks"			**Abuse** is to be understood in a broad sense.
CY-Z16	To what extent are the damages which – in a realistic worst-case scenario – can be caused by data access by unauthorized persons or by data use outside the contractual conditions, can be remedied, or to what extent are states prior to access or use of the data restorable? Completely fixable = 1 Partially fixable = 2 Slightly to not fixable = 3	General	"General AI-specific cybersecurity" and "Resilience against AI-specific attacks"			This section focuses solely on the direct consequences of unauthorized data access and use. The consequences of misuse and manipulation are addressed elsewhere. While compromised passwords can be changed, the immediate damage is remediable. However, the loss of sensitive information is irreversible, and the resulting damage is irreparable.
CY-Z17	Could deliberate misuse or manipulation of the AI system – in a realistic worst-case scenario – cause fatal personal injury? yes = 3 no = 1	Body and life	"General AI-specific cybersecurity" and "Resilience against AI-specific attacks"			
CY-Z18	Could deliberate misuse or manipulation of the AI system – in a realistic worst-case scenario – cause health problems (physical health) in natural persons? yes = 2 no = 1	Health	"General AI-specific cybersecurity" and "Resilience against AI-specific attacks"			
CY-Z19	Could deliberate misuse or manipulation of the AI system cause societal harm or violations of fundamental rights? yes = 3 no = 1	General	"General AI-specific cybersecurity" and "Resilience against AI-specific attacks"			
CY-Z20	Could deliberate misuse or manipulation of AI systems have a particularly strong impact on groups that define themselves by protected characteristics? yes = 3 no = 1	General	"General AI-specific cybersecurity" and "Resilience against AI-specific attacks"			**Protected properties** are (based on the German General Equal Treatment Act (AGG)): race, ethnic origin, gender, religion or belief, disability, age, sexual identity.
	Erweiterungsfragen	**Protection goal category**	**Criterion**	**Assessment**	**Optional: Comment**	**Notes and examples**
CY-Z23	In a realistic worst-case scenario, how many harmful results or outputs could potentially be generated by the AI system through intentional misuse or manipulation, until safe, regular operation is restored? low number = 1 moderate number = 2 high number = 3	General	"General AI-specific cybersecurity" and "Resilience against AI-specific attacks"			Example boundaries: **low number** = 0-1,000 **moderate number** = 1,001-10,000 **high number** =>10,000

Question index	Question	Protection goal category	Criterion	Assessment	Optional: Comment	Notes and examples
CY-Z24	How well can malicious errors in the AI system, caused by intentional misuse or manipulation, be detected as such? always easy and clear = 1 with slight to moderate effort = 2 with great or greatest effort = 3	General	“General AI-specific cybersecurity” and “Resilience against AI-specific attacks”			
CY-Z25	Could deliberate misuse or manipulation of the AI system – in a realistic worst-case scenario – cause fatal harm to animals? yes = 3 no = 1	Body and life	“General AI-specific cybersecurity” and “Resilience against AI-specific attacks”			
CY-Z26	In a realistic worst-case scenario, what is the potential financial damage resulting from a breach of confidentiality or integrity of sensitive business or licensed data, or intellectual property, processed by the AI system? The potential financial damage serves here as a proxy measure of the harm suffered by the affected party or parties. No to low financial damage = 1 moderate financial damage = 2 severe or catastrophic financial damage = 3	Property and things	“General AI-specific cybersecurity” and “Resilience against AI-specific attacks”			Example boundaries: **no or low financial damage** = Damage amounting to up to 1% of the previous year's turnover or less than €1,000 **moderate financial damage** Damage amounting to 1% to 5% of the previous year's turnover or between €1,000 and €50,000 **serious or catastrophic financial damage** = Damage amounting to more than 5% of the previous year's turnover or more than €50,000
CY-Z27	How can the environmental impact (sustainability) of intentional misuse or manipulation of the AI system – in a realistic worst-case scenario – be described? Minimal negative impact = 1 Moderate negative impact = 2 High or catastrophic impact = 3	Environment	“General AI-specific cybersecurity” and “Resilience against AI-specific attacks”			Environmental impacts can relate to wasted material resources or electricity. Examples of limits: **at most minor negative effects** = Individual unusable products are being produced **moderate negative effects** = Some unusable products are being **produced high or catastrophic impacts** = Large quantities of unusable products are being produced.

Protection needs analysis

Quality dimension: Reliability (VE)

Overall assessment per criterion

Protection requirement criterion “performance and robustness”:			1 / 2 / 3	always applicable
Protection requirement criterion “fallback plans and functional safety”:			1 / 2 / 3	always applicable

Question index	Question	Protection goal category	Criterion	Assessment	Optional: Comment	Notes and examples
	Basic questions	**Protection goal category**	**Criterion**	**Assessment**	**Optional: Comment**	**Notes and examples**
VE-Z11	To what extent are the damages that can be caused by the AI system – in a realistic worst-case scenario – repairable or recoverable? Completely fixable = 1 Partially fixable = 2 Not fixable = 3	General	Fallback plans and functional safety			
VE-Z12	Could the AI system – in a realistic worst-case scenario – cause immediate fatal personal injury? yes = 3 no = 1	Body and life	Performance and robustness			Identical to question VE-Z13 (different criterion)
VE-Z13	Could the AI system – in a realistic worst-case scenario – cause immediate fatal personal injury? yes = 3 no = 1	Body and life	Fallback plans and functional safety			Identical to question VE-Z12 (different criterion)
VE-Z14	Could the AI system – in a realistic worst-case scenario – cause health problems (physical health) in people? yes = 2 no = 1	Health	Performance and robustness			Identical to question VE-Z15 (different criterion)
VE-Z15	Could the AI system – in a realistic worst-case scenario – cause health problems (physical health) in natural persons? yes = 2 no = 1	Health	Fallback plans and functional safety			Identical to question VE-Z14 (different criterion)
VE-Z16	Does the AI system regulate access to services and activities that are essential to personality or personal scope of action? yes = 3 no = 1	General	Performance and robustness			
VE-Z17	Is the AI system involved in decision-making processes that affect fundamental rights or democratic procedures? yes = 3 no = 1	General	Performance and robustness			
	Extension questions	**Protection goal category**	**Criterion**	**Assessment**	**Optional: Comment**	**Notes and examples**
VE-Z20	How well can harmful results or outputs caused by the AI system be identified as such? always easy and clear = 1 with slight to moderate effort = 2 with great or greatest effort or not at all = 3	General	Performance and robustness			
VE-Z21	In a realistic worst-case scenario, how many harmful results or outputs could potentially be generated by the AI system due to a faulty state or behavior before safe, normal operation is restored? low number = 1 moderate number = 2 high number = 3	General	Fallback plans and functional safety			Example boundaries: **low number** = 0–1,000 **moderate number** = 1,001–10,000 **high number** => 10,000
VE-Z22	Is it unclear how the validity of the AI system's results can be measured? yes = 3 no = 1	General	Performance and robustness			Here, **users** means natural persons who interact directly and actively with an AI system, either as end users who use the AI system for personal or business purposes, or as users who apply the AI system in a professional context. The term encompasses both those who operate the AI system and those who use it to achieve specific goals. Depending on the context, this can include the use of AI systems in everyday applications or in specialized, professional scenarios (see Glossary).

Question index	Question	Protection goal category	Criterion	Assessment	Optional: Comment	Notes and examples
VE-Z23	How can the alternative options for users or the task of the AI system be described in the event that the desired functionality of the AI system is not available? as adequate and quickly available = 1 as available with slight to moderate limitations = 2 as available only with major and severe limitations = 3	General	Fallback plans and functional safety			
VE-Z24	How frequently is the AI system retrained or updated with new versions during operation? never = 1 occasionally = 2 often = 3	General	Performance and robustness			Identical to question TR-Z31 (other dimension) example limits: **never** = not even once **now and again** = monthly or less frequently **often** = more frequently than monthly
VE-Z25	Could the AI system – in a realistic worst-case scenario – cause fatal harm to animals? yes = 3 no = 1	Body and life	Performance and robustness			Identical to question VE-Z26 (different criterion)
VE-Z26	Could the AI system – in a realistic worst-case scenario – cause fatal harm to animals? yes = 3 no = 1	Body and life	Fallback plans and functional safety			Identical to question VE-Z25 (different criterion)
VE-Z27	What is the potential financial damage – for all involved stakeholders, in a realistic worst-case scenario – caused by the AI system? No to low financial damage = 1 moderate financial damage = 2 severe or catastrophic financial damage = 3	Property and things	Performance and robustness			Note: This is about the **directly measurable financial damage** Examples of limits: **no or low financial damage** = Damage amounting to less than €1,000 **moderate financial damage** = Damage amounting to between €1,000 and €50,000 **serious or catastrophic financial damage** = Damage amounting to over €50,000
VE-Z28	What is the potential financial damage – for all involved stakeholders, in a realistic worst-case scenario – caused by the AI system? No to low financial damage = 1 moderate financial damage = 2 severe or catastrophic financial damage = 3	Property and things	Fallback plans and functional safety			Note: This is about the directly measurable financial damage Examples of limits: **no or low financial damage** = Damage amounting to less than €1,000 **moderate financial damage** = Damage amounting to between €1,000 and €50,000 **serious or catastrophic financial damage** = Damage amounting to over €50,000
VE-Z29	How can the environmental impact (sustainability) of the AI system be described – in a realistic worst-case scenario? Minimal negative impact = 1 Moderate negative impact = 2 High or catastrophic impact = 3	Environment	Performance and robustness			Identical to question VE-Z30 (different criterion): Environmental impacts can relate to wasted material resources or electricity. Examples of limits: **at most minor negative effects** = Individual unusable products are being produced. **moderate negative effects** = Some unusable products are being produced. **high or catastrophic impacts** = Large quantities of unusable products are being produced.
VE-Z30	How can the environmental impact (sustainability) of the AI system be described – in a realistic worst-case scenario? Minimal negative impact = 1 Moderate negative impact = 2 High or catastrophic impact = 3	Environment	Fallback plans and functional safety			Identical to question VE-Z29 (different criterion)

4.4 Assessment Catalogue

Navigation

Quality dimensions	Data quality, protection and governance			Non-discrimination	Transparency		Human oversight and control		Reliability		AI-specific cybersecurity	
Criteria	**DA1**	**DA2**	**DA3**	**ND1**	**TR1**	**TR2**	**MA1**	**MA2**	**VE1**	**VE2**	**CY1**	**CY2**
Indicators	DA1.1	DA2.1	DA3.1	ND1.1	TR1.1	TR2.1	MA1.1	MA2.1	VE1.1	VE2.1	CY1.1	CY2.1
	DA1.2	DA2.2	DA3.2	ND1.2	TR1.2	TR2.2	MA1.2	MA2.2	VE1.2	VE2.2	CY1.2	CY2.2
	DA1.3	DA2.3	DA3.3	ND1.3	TR1.3	TR2.3	MA1.3	MA2.3	VE1.3	VE2.3	CY1.3	CY2.3
	DA1.4	DA2.4	DA3.4	ND1.4	TR1.4	TR2.4	MA1.4	MA2.4	VE1.4	VE2.4	CY1.4	CY2.4
	DA1.5	DA2.5	DA3.5	ND1.5	TR1.5		MA1.5	MA2.5	VE1.5	VE2.5	CY1.5	
		DA2.6	DA3.6	ND1.6	TR1.6		MA1.6		VE1.6	VE2.6	CY1.6	
		DA2.7		ND1.7	TR1.7		MA1.7		VE1.7	VE2.7		
				ND1.8	TR1.8							
				ND1.9								

This document is part of the MISSION KI quality standard. ©acatech – National Academy of Science and Engineering. This work is licensed under a Creative Commons Attribution-NoDerivs 4.0 International Licence (CC BY-ND 4.0).
https://creativecommons.org/licenses/by-nd/4.0/deed.en

Data quality, protection and governance

DA1 Data quality

DA1.1	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
The characteristics of the data sets must be documented.	**Analysis – Definition** Data composition – number of data instances / data size – standards for data structures/formats Data collection process Method of data collection – Source of data – Person responsible for data collection Data processing steps – Explanation of the features and their possible quality dimensions and areas of the quality dimensions – Preprocessing – Labeling – Cleaning Data maintenance – Storage periods – Data updates	**Analysis – Definition** Data composition – number of data instances / data size – standards for data structures/formats Data collection process – data collection method – data source Data processing steps – explanation of the features and their possible quality dimensions and areas of the quality dimensions – Labeling – Cleaning Data maintenance – Storage periods – Data updates	**Analysis – Definition** Data composition – number of data instances / data size – standards for data structures/formats Data collection process – Source of data Data processing steps – Explanation of the features and their possible quality dimensions and areas of the quality dimension Data maintenance – Storage periods – Updating the data	The characteristics of the data were not documented.		Component	Measure	Maximum value	TR1.3

DA1.2	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
The risks arising from a lack of data quality in the context of the purpose of the AI system must be analysed, and data quality requirements must be derived from this.	**Analysis – Purpose Definition** Based on the intended use and application area of the AI system and the potential risks, relevant data quality characteristics must be defined and specific requirements must be set for them. Analysis risk Detailed risk analysis including quantification of probabilities of occurrence and risks: – Identification of possible risks and their causes – Assignment of risk responsibility – Estimation of the probability of occurrence – Estimation of the probability of detection – Estimation of the impact – Structured grading and prioritisation of risks At least the following risk sources must be considered: – Data quality characteristics do not fit the intended use and scope – Data quality characteristics are poorly met.	**Analysis – Purpose Definition** Based on the intended use and application area of the AI system and the potential risks, relevant data quality characteristics must be defined and specific requirements must be set for them. Analysis risk Limited risk analysis focusing on protection needs without quantifying probabilities: – Identification of potential risks – Assignment of risk responsibility – Estimation of impact – Structured grading and prioritisation of risks At least the following risk sources must be considered: – Data quality characteristics do not fit the intended use and application area – Data quality characteristics are poorly met.	**Analysis – Purpose Definition** Based on the intended use and application area of the AI system and the potential risks, relevant data quality characteristics must be defined and specific requirements must be set for them. Analysis risk Primarily qualitative assessment without probabilities: Identification of potential hazards – Assignment of risk responsibility – Qualitative assessment of the impact – Qualitative grading and prioritisation of hazards. At least the following risk sources must be considered: – Data quality characteristics do not fit the intended use and application area – Data quality characteristics are poorly met.	The risks and hazards were not analysed, and no data quality requirements were derived.	ISO/IEC 5259	System/Component	Analysis	Maximum value	TR1.1

DA1.3	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
The datasets must correspond to the intended purpose of the AI system and meet the derived data quality requirements.	**Technical measures – data** The data quality characteristics defined in DA 1.2 must be tested and evaluated, taking into account the scope and intended use. This includes: – Tests and metrics to demonstrate the data quality characteristics – At a minimum, completeness, timeliness, and correctness must be considered. The selected metrics and tests should ideally have the following characteristics: – If justified by the selection, the metrics and tests should include both important basic methods and, where possible, advanced methods (in terms of complexity, information content, implementation effort, e.g., comprehensive testing tools) – The selected metrics and tests should, where possible, allow for continuous evaluation of the AI models through a high degree of automation – Definition of thresholds (as minimum requirements for functionality/performance) and justification of the thresholds, taking into account the scope and intended use – If applicable: Differentiation of the metrics and thresholds according to different use cases, with reference to the definition of the scope.	**Technical measures – data** The data quality characteristics defined in DA 1.2 must be tested and evaluated, taking into account the scope and intended use. This includes: – Tests and metrics to demonstrate the data quality characteristics – At a minimum, completeness and correctness must be considered. The selected metrics and tests should ideally have the following characteristics: – If justified by the selection, the metrics and tests should include both important basic methods and, where possible, advanced methods (in terms of complexity, information content, implementation effort, e.g., comprehensive testing tools) – Definition of thresholds (as minimum requirements for functionality/performance) and justification of the thresholds, taking into account the scope and intended use.	**Technical measures – data** The data quality characteristics defined in DA 1.2 must be tested and evaluated, taking into account the scope and intended use. This includes: – Tests and metrics to demonstrate the data quality characteristics. The selected metrics and tests should ideally have the following characteristics: – Basic methods are sufficient (regarding complexity, information content, implementation effort, e.g., a simple metric) – Definition of thresholds (as minimum requirements for functionality/performance) and justification of the thresholds, taking into account the scope and intended use.	No evidence was provided that the dataset meets the purpose of the AI system and the data quality requirements.	During training ISO/IEC 5259	Component	Measure	Maximum value	DA1.4

DA1.4	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
The data from the AI system must (be able to) be monitored during development and operation.	MA2.3					Component	Measure	Normal	DA1.3

DA1.5	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
It must be assessed whether the measures taken have reduced the identified risks to an acceptable level and whether the quality of the AI system meets the set targets.	**Evaluation** – Summary review of the effects of implementing the technical and organisational measures – Assessment of the interactions of the measures taken, and consideration of the mitigation of risks in conjunction with the risks of other quality dimensions – Identification and description of the residual risk after implementation of the technical and organisational measures – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk.	**Evaluation** – Weighing the interactions of the measures taken, and weighing the mitigation of risks in conjunction with the risks of other quality dimensions – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk.	**Evaluation** – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk.	No evaluation was conducted.		System	Evaluation	Maximum value	

Data quality, protection and governance

DA2	Protection of personal data								
DA2.1	**A**	**B**	**C**	**D**	**Additional information**	**Reference level**	**Type**	**Weighting**	**Link**
A risk analysis for the protection of personal data used in the AI system must be carried out taking into account the intended use.	**Data protection impact assessment implemented in accordance with GDPR Art. 35 or Risk analysis.** Full risk analysis including quantification of probabilities of occurrence and risks: – Identification of potential risks and their causes – Assignment of risk responsibility – Estimation of the probability of occurrence – Estimation of the probability of detection – Estimation of the impact – Structured grading and prioritisation of risks. At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Data processing operations: Collection, storage, processing, and transfer of personal data, in particular risks in implementing the right to data portability or erasure. – Purpose limitation: Use of data beyond the originally defined purpose. – Data minimisation: Collection of unnecessary or excessive personal data. – Data access rights and storage: Unauthorised or uncontrolled access to sensitive personal data and retention of data beyond necessary periods. – Data transfer to third parties: Uncontrolled transfer of personal data to external partners or service providers, risks in transferring personal data to countries with inadequate data protection. – Misuse of data: Potential misuse, such as identity theft or discrimination through improper data processing. – Anonymisation and pseudonymisation: Inadequate anonymisation or pseudonymisation techniques. – Automated decision-making: Lack of transparency or control over automated decisions based on personal data.	**Analysis – Risk** Limited risk analysis focusing on protection needs without quantifying probabilities – Identification of potential risks – Allocation of risk responsibility – Impact assessment – Structured risk grading and prioritisation. At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Data processing activities: Collection, storage, processing, and transfer of personal data, in particular risks related to the implementation of the right to data portability or erasure. – Purpose limitation: Use of data beyond the originally defined purpose. – Data minimisation: Collection of unnecessary or excessive personal data. – Data access rights and storage: Unauthorised or uncontrolled access to sensitive personal data and retention of data beyond necessary periods. – Data transfer to third parties: Uncontrolled transfer of personal data to external partners or service providers, risks associated with transferring personal data to countries with inadequate data protection. – Misuse of data: Potential misuse, such as identity theft or discrimination through improper data processing. – Anonymisation and pseudonymisation: Inadequate anonymisation or pseudonymisation techniques. – Automated decision-making: Lack of transparency or control over automated decisions based on personal data.	**Analysis – Risk** Primarily a qualitative assessment of risks without probabilities – Identification of potential risks: – Assignment of risk responsibility – Qualitative assessment of impact – Qualitative grading and prioritisation of risks. At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Data processing operations: Collection, storage, processing, and transfer of personal data, in particular risks in implementing the right to data portability or erasure. – Data access rights and storage: Unauthorised or uncontrolled access to sensitive personal data and retention of data beyond necessary periods. – Data transfer to third parties: Uncontrolled transfer of personal data to external partners or service providers, risks in transferring personal data to countries with inadequate data protection. – Data misuse: Potential misuse, such as identity theft or discrimination through improper data processing. – Anonymisation and pseudonymisation: Inadequate anonymisation or pseudonymisation techniques.	No risk analysis was conducted.	Relevant here are the GDPR and the requirements contained therein for a data protection impact assessment (DPIA), Art. 35, GDPR.	System	Analysis	Minimum value	DA1.1, DA1.2, TR1.1 (purpose)
DA2.2	**A**	**B**	**C**	**D**	**Additional information**	**Reference level**	**Type**	**Weighting**	**Link**
Non-AI-specific measures must be taken to ensure the protection of personal data, taking into account the identified risk.	**Measures implemented in accordance with the data protection impact assessment pursuant to GDPR Art. 35 or** **Organisational measures – Governance / System-related processes** – Implementation of data minimization mechanisms to collect and process only the necessary personal data. – Introduction of strict access controls and authorisation management to ensure that only authorised persons have access to personal data. – Implementation of retention period management mechanisms to securely delete data after the specified periods have expired. – Ensuring that clear contractual agreements exist when data is transferred to third parties. – Implementation of clear consent processes and transparent information methods for data subjects. – Implementation of an emergency plan for the detection, reporting, and remediation of data breaches within the legally required timeframe. – Ensuring that appropriate safeguards (e.g., standard contractual clauses) are implemented when transferring personal data to third countries. **Technical measures – data** – Ensuring that data protection measures are already integrated into the development and implementation of the systems ("data protection by design") – Use of encryption techniques to protect stored data – Use of anonymisation, pseudonymisation and other technologies such as differential privacy to prevent information from being classified as personal data, with justification for the necessity and effectiveness of the measures – Implementation of processes to inform data subjects about automated decisions and their logic (where applicable, reference to AI-specific measures on explainability, see TR2) **Technical measures – Operation** – Regular reviews and audits of data processing processes and practices to ensure that data protection requirements are met in accordance with the intended use.	**Organisational measures – Governance / System-related processes** – Implementation of strict access controls and authorisation management to ensure that only authorised persons have access to personal data. – Implementation of retention period management mechanisms to securely delete data after the specified periods have expired. – Ensuring that clear contractual agreements exist when data is transferred to third parties. – Implementation of clear consent processes and transparent information methods for data subjects. – Implementation of an emergency plan for the detection, reporting, and remediation of data breaches within the legally required timeframe. – Ensuring that appropriate safeguards (e.g., standard contractual clauses) are implemented when transferring personal data to third countries. **Technical measures – data** – Ensuring that data protection measures are already integrated into the development and implementation of the systems ("data protection by design") – Use of encryption techniques to protect stored data – Use of anonymisation, pseudonymisation and other technologies such as differential privacy to prevent information from being classified as personal data, with justification for the necessity and effectiveness of the measures **Technical measures – Operation** – Regular reviews and audits of data processing processes and practices to ensure that data protection requirements are met in accordance with the intended use.	**Organisational measures – Governance / System-related processes** – Implementation of strict access controls and authorisation management to ensure that only authorised persons have access to personal data. – Implementation of retention period management mechanisms to securely delete data after the specified periods have expired. – Ensuring that clear contractual agreements exist when data is transferred to third parties. – Implementation of an emergency plan for the detection, reporting, and remediation of data breaches within the legally required timeframe. – Ensuring that appropriate safeguards (e.g., standard contractual clauses) are implemented when transferring personal data to third countries. **Technical measures – data** – Use of encryption techniques to protect stored data – Use of anonymisation, pseudonymisation and other technologies such as differential privacy to prevent information from being classified as personal data, with justification of the necessity and effectiveness of the measures.	No non-AI-specific measures were taken to protect personal data.	Relevant here are the GDPR and the requirements contained therein for a data protection impact assessment (DPIA), Art. 35, GDPR.	System/Component	Measure	Maximum value	DA1.2, DA1.3, TR2, MA1, CY1.5
DA2.3	**A**	**B**	**C**	**D**	**Additional information**	**Reference level**	**Type**	**Weighting**	**Link**
Measures and security controls must be in place to mitigate AI-specific attacks that occur during the preparation and model training phase of the AI models.	CY2.3				For reasons of plausibility, not all measures in CY2.3 are necessarily relevant to the protection of personal data, e.g. when it comes to the protection goal of availability.	System/Component	Measure	Maximum value	CY2.2, CY2.3
DA2.4	**A**	**B**	**C**	**D**	**Additional information**	**Reference level**	**Type**	**Weighting**	**Link**
Measures and security controls must be in place to mitigate AI-specific attacks that may occur during operation.	CY2.2				For reasons of plausibility, not all measures in CY1.6 are necessarily relevant to the protection of personal data, e.g. when it comes to the protection goal of availability.	System/Component	Measure	Maximum value	CY2.2, CY2.3
DA2.5	**A**	**B**	**C**	**D**	**Additional information**	**Reference level**	**Type**	**Weighting**	**Link**
Measures and security controls must be in place to prevent misuse and intentional abuse of the AI system.	CY1.4				This should primarily cover the incorrect use or misuse of the system, which falsely uses or discloses personal data.	System/Component	Measure	Normal	

Data quality, protection and governance

DA2.6	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
It should be possible to exercise their rights relating to personal data, including data management, deletion and use, as well as information obligations, even during the operation of the AI system.	**Organisational measure – User instruction** – Providing a clear and understandable explanation of the data processing, the data subject rights and the controller – Documentation and logging of all data processing activities and the exercise of data subject rights (see TR1.7) **Organisational measure – Governance** – Ensuring that users are informed promptly about changes to data processing or use. – Establishing clear channels to support data subjects in exercising their rights (see MA1.4). – Implementing a system for obtaining and managing explicit consent for data processing. **Technical measure – Operation** – Providing easily accessible functions for data subjects to exercise their rights to information, rectification, erasure, and data portability, including the ability to object to data processing or withdraw their consent at any time. – Integrating mechanisms that enable the automatic deletion of personal data after the retention period has expired or upon request.	**Organisational measure – User instruction** – Providing a clear and understandable explanation of the data processing, the data subject rights and the controller – Documentation and logging of all data processing activities and the exercise of data subject rights (see TR1.7) **Organisational measure – Governance** – Establishment of clear channels to assist data subjects in exercising their rights (see MA1.4) – Implementation of a system for obtaining and managing explicit consent for data processing. **Technical measure – Operation** – Integration of mechanisms that enable automatic deletion of personal data after the storage period has expired or upon request.	**Organisational measure – User instruction** – Providing a clear and understandable explanation of data processing, data subject rights and the data controller **Organisational measure – Governance** – Establishment of clear channels to assist data subjects in exercising their rights (see MA1.4) – Implementation of a system for obtaining and managing explicit consent for data processing. **Technical measure – Operation** – Operation Integration of mechanisms that enable automatic deletion of personal data after the storage period has expired or upon request.	There are no options for natural persons to exercise their rights with regard to personal data.		Component	Measure	Maximum value	TR1.7, MA1.4

DA2.7	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
It must be assessed whether the measures taken have reduced the identified risks to an acceptable level and whether the quality of the AI system meets the set targets.	**Evaluation** – Summary review of the effects of implementing the technical and organisational measures. – Assessment of the interactions of the measures taken, and consideration of the mitigation of risks in conjunction with the risks of other quality dimensions. – Identification and description of the residual risk after implementation of the technical and organisational measures. – Assessment by a qualified and authorised person as to whether the residual risk is tolerable. – Justification of the tolerability of the residual risk.	**Evaluation** – Weighing the interactions of the measures taken, and weighing the mitigation of risks in conjunction with the risks of other quality dimensions – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	**Evaluation** – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk.	No evaluation was conducted.		System	Evaluation	Maximum value	

DA3 Protection of proprietary data

DA3.1	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
A risk analysis for the protection of proprietary data used in the AI system must be carried out taking into account the intended use.	**Analysis – Risk** Full risk analysis including quantification of probabilities of occurrence and risks: – Identification of potential risks and their causes – Assignment of risk responsibility – Estimation of the probability of occurrence – Estimation of the probability of detection – Estimation of the impact – Structured grading and prioritisation of risks. At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Unauthorised access to and lack of control over access to sensitive company data by internal or external actors. – Risks of accidental loss, deletion, or damage to proprietary data due to system errors or human error. – Threat from external attacks, hacking, or industrial espionage aimed at the theft, disclosure, or manipulation of confidential company data. – Loss of control over company data during its storage or processing by external service providers (e.g., in cloud environments). – Contractual or agreement breaches by partners or service providers regarding unauthorised data access or disclosure. – Non-compliance with industry-specific regulations regarding the use and disclosure of data.	**Analysis – Risk** Limited risk analysis focusing on protection requirements without quantifying probabilities – Identification of potential risks – Allocation of risk responsibility – Estimation of impact – Structured grading and prioritisation of risks. At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Unauthorised access to and lack of control over access to sensitive company data by internal or external actors. – Risks of accidental loss, deletion, or damage to proprietary data due to system errors or human error. – Threat from external attacks, hacking, or industrial espionage aimed at the theft, disclosure, or manipulation of confidential company data. – Loss of control over company data during its storage or processing by external service providers (e.g., in cloud environments). – Contractual or agreement breaches by partners or service providers regarding unauthorised data access or disclosure. – Non-compliance with industry-specific regulations regarding the use and disclosure of data.	**Analysis – Risk** Primarily a qualitative assessment of risks without probabilities – Identification of potential risks: – Assignment of risk responsibility – Qualitative assessment of impact – Qualitative grading and prioritisation of risks. At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Unauthorised access to and lack of control over access to sensitive company data by internal or external actors. – Risks of accidental loss, deletion, or damage to proprietary data due to system errors or human error. – Danger from external attacks, hacking, or industrial espionage aimed at the theft, disclosure, or manipulation of confidential company data. – Loss of control over company data during its storage or processing by external service providers (e.g., in cloud environments). – Contractual or agreement breaches by partners or service providers regarding unauthorised data access or disclosure. – Non-compliance with industry-specific regulations regarding the use and disclosure of data.	No risk or hazard analysis was conducted.	Relevant here are the GDPR and the requirements contained therein for a data protection impact assessment (DPIA), Art. 35, GDPR.	System	Analysis	Maximum value	DA1.1, DA1.2, TR1.1 (purpose)

DA3.2	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
Non-AI-specific measures must be taken to ensure the protection of proprietary data, taking into account the identified risk.	**Organisational measures – Governance / System-related processes** – Implementation of retention period management mechanisms to securely delete data after the specified periods have expired. – Clear contractual agreements with external service providers regarding access to, use of, and protection of proprietary data. – Ensuring that companies retain control over their data at all times, especially when using external cloud or storage services. – Implementation of a contingency plan for the detection, reporting, and remediation of data breaches. **Technical measures – Data** – Implementation of strict access controls and authorisation management to ensure that only authorised personnel have access to proprietary data. – Ensuring that data protection measures are integrated into the development and implementation of systems ("data protection by design"). – Implementation of encryption techniques and security protocols to protect the confidentiality and integrity of company data, especially during transmission and storage. – Implementation of mechanisms to ensure the integrity and authenticity of company data to prevent manipulation and guarantee its genuineness. – Implementation of regular data backups and contingency plans for data recovery in case of loss or damage. – Ensuring the interoperability of data formats and the ability to easily transfer company data between systems or providers. **Technical measures – Operation** – Regular reviews and audits of data processing processes and practices to ensure that data protection requirements are met in accordance with the intended use.	**Organisational measures – Governance / System-related processes** – Implementation of retention period management mechanisms to securely delete data after the specified periods have expired. – Clear contractual agreements with external service providers regarding access to, use of, and protection of proprietary data. – Ensuring that companies retain control over their data at all times, especially when using external cloud or storage services. – Implementation of a contingency plan for the detection, reporting, and remediation of data breaches. **Technical measures – Data** Implementation of strict access controls and authorisation management to ensure that only authorised individuals have access to proprietary data. – Implementation of encryption techniques and security protocols to protect the confidentiality and integrity of company data, especially during transmission and storage. – Implementation of mechanisms to ensure the integrity and authenticity of company data to prevent manipulation and guarantee its genuineness. – Implementation of regular data backups and contingency plans for data recovery in case of loss or damage. **Technical measures – Operation** – Regular reviews and audits of data processing processes and practices to ensure that data protection requirements are met in accordance with the intended use.	**Organisational measures – Governance / System-related processes** – Implementation of retention period management mechanisms to securely delete data after the specified periods have expired. – Clear contractual agreements with external service providers regarding access, use, and protection of proprietary data. – Ensuring that companies retain control over their data at all times, especially when using external cloud or storage services. **Technical measures – Data** – Implementation of access controls and authorisation management to ensure that only authorised personnel have access to proprietary data. – Implementation of encryption techniques to protect the confidentiality and integrity of company data, especially during transmission and storage. – Implementation of regular data backups and contingency plans for data recovery in case of loss or damage.	No non-AI-specific measures were taken to protect proprietary data.		System/Component	Measure	Maximum value	DA1.2, DA1.3, MA1, CY1.5

DA3.3	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
Measures and security controls must be in place to mitigate AI-specific attacks that occur during the preparation and model training phase of the AI models.	CY2.3				For reasons of plausibility, not all measures in CY2.3 are necessarily relevant for the protection of proprietary data, e.g., when it comes to the protection goal of availability.	System/Component	Measure	Maximum value	

Data quality, protection and governance

DA3.4	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
Measures and security controls must be in place to mitigate AI-specific attacks that may occur during operation.	CY2.2				For reasons of plausibility, not all measures in CY1.6 are necessarily relevant for the protection of proprietary data, e.g., when it comes to the protection goal of availability.	System/Component	Measure	Maximum value	

DA3.5	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
Measures and security controls must be in place to prevent misuse and intentional abuse of the AI system.	CY1.4				This should primarily cover the improper use or misuse of the system, which falsely uses or discloses proprietary data.	System/Component	Measure	Maximum value	

DA3.6	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
It must be assessed whether the measures taken have reduced the identified risks to an acceptable level and whether the quality of the AI system meets the set targets.	**Evaluation** – Summary review of the effects of implementing the technical and organisational measures – Assessment of the interactions of the measures taken, and consideration of the mitigation of risks in conjunction with the risks of other quality dimensions – Identification and description of the residual risk after implementation of the technical and organisational measures – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk.	**Evaluation** – Weighing the interactions of the measures taken, and weighing the mitigation of risks in conjunction with the risks of other quality dimensions – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk.	**Evaluation** – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk.	No evaluation was conducted.		System	Evaluation	Maximum value	

Non-discrimination

ND1 Avoidance of unjustified discrimination and distortion

ND1.1	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
The risk of discrimination related to the intended purpose of the AI system must be analysed.	**Analysis – Risk** Detailed risk analysis including quantification of probabilities of occurrence and risks: – Identification of potential risks and their causes – Assignment of risk responsibility – Estimation of the probability of occurrence – Estimation of the probability of detection – Estimation of the impact – Structured grading and prioritisation of risks. At least the following sources of risk must be considered, taking into account the intended use (see TR1.1): – Historical biases – Sampling bias – Model biases – Attribution errors – Implicit bias – Confirmation bias – Developer bias – Bias due to user interactions – Misinterpretation of results – Insufficient consideration of potential (user) groups – Disadvantage/discrimination against groups with protected attributes	**Analysis – Risk** Limited risk analysis focusing on protection needs without quantifying probabilities: – Identification of potential risks – Assignment of risk responsibility – Estimation of impact – Structured grading and prioritisation of risks. At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Historical biases – Sampling bias – Model biases – Attribution errors – Implicit bias – Confirmation bias – Developer bias – Bias due to user interactions – Misinterpretation of results – Insufficient consideration of potential (user) groups – Disadvantage/discrimination against groups with protected attributes	**Analysis – Risk** Primarily qualitative assessment without probabilities: – Identification of potential hazards – Assignment of risk responsibility – Qualitative assessment of impact – Qualitative grading and prioritisation of risks. At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Historical biases – Sampling bias – Model biases – Attribution errors – Implicit bias – Confirmation bias – Developer bias – Bias due to user interactions – Misinterpretation of results – Insufficient consideration of potential (user) groups – Disadvantage/discrimination against groups with protected attributes	No risk analysis was conducted.		System	Analysis	Maximum value	TR1.1

ND1.2	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
The necessary degree of avoidance of unjustified distortion or discrimination must be defined in the context of the intended purpose.	**Analysis – Definition / Metrics & Thresholds** Based on the defined purpose of the AI system (TR1.1), it must be defined what level of avoidance of bias or protection against discrimination is necessary. This includes: – Identifying sensitive or protected characteristics in the data with justification – Identifying groups with protected characteristics based on potential users and affected individuals – Identifying additional groups that may be unexpectedly affected (e.g., because their characteristics are not explicitly stated as features but only implicitly present in the data) – Defining the intended fairness – Collaborating with representatives of identified groups with protected characteristics – Defining tests, metrics, and thresholds to capture the objective within the framework of the fairness definition. The selected metrics and tests should have at least the following characteristics: – Methods should be applied to each AI model within and for the AI system as a whole, if it generates an aggregated output. – If justified by the selection, the metrics and tests should include both essential basic methods and advanced methods (regarding complexity, information content, implementation effort, e.g., comprehensive testing tools). – The selected metrics and tests should, where possible, allow for continuous evaluation of the AI models through a high degree of automation.	**Analysis – Definition / Metrics & Thresholds** Based on the defined purpose of the AI system (TR1.1), the necessary level of bias prevention and discrimination protection must be defined. This includes: – Identifying sensitive or protected characteristics in the data, with justification – Identifying groups with protected characteristics based on potential users and affected individuals – Defining the intended fairness – Establishing tests, metrics, and thresholds to measure the objective within the framework of the fairness definition. The selected metrics and tests should have at least the following characteristics: – Methods should be applied to each AI model within and for the AI system as a whole, if it generates an aggregated output. – If justified by the selection, the metrics and tests should include both essential basic methods and advanced methods (in terms of complexity, information content, implementation effort, e.g., comprehensive testing tools).	**Analysis – Definition / Metrics & Thresholds** Based on the defined purpose of the AI system (TR1.1), the necessary level of bias prevention and discrimination protection must be defined. This includes: – Identifying sensitive or protected characteristics in the data – Identifying groups with protected characteristics based on potential users and affected individuals – Defining the intended fairness – Establishing tests, metrics, and thresholds to measure the objective within the fairness definition. The selected metrics and tests should have at least the following characteristics: – Methods should be applied to each AI model within and for the AI system as a whole if it generates an aggregated output. – Basic methods are sufficient (in terms of complexity, information content, and implementation effort, e.g., a simple metric).	No definition or objective has been set to avoid unjustified distortion and discrimination.		System/Component	Analysis	Maximum value	TR1.1

ND1.3	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
An analysis of the AI system regarding existing biases and possible discrimination must be carried out.	**Technical measures – model, data & tests** The AI system and its models and data were examined with regard to the objective defined in ND1.2. This includes at least: – Justification of the model selection (including the task definition and optimisation strategy, as well as any pre- or post-processing measures in the AI system to mitigate biases) – Description of the scope and execution of the tests (see ND1.2) – Comparison with alternative tests/metrics – Evaluation of compliance with predefined thresholds or target values – Documentation of the test results and conclusions drawn from them – Description of the limitations of the validity of the tests performed and their results – Collaboration with relevant groups with protected attributes	**Technical measures – model, data & tests** The AI system and its models and data were examined with regard to the objective defined in ND1.2. This includes at least: – Description of the scope and execution of the tests (see ND1.2) – Comparison with alternative tests/metrics – Evaluation of compliance with predefined thresholds or target values – Documentation of the test results and conclusions drawn from them.	**Technical measures – model, data & tests** The AI system and its models and data were examined with regard to the objective defined in ND1.2. This includes at least: – Description of the scope and execution of the tests (see ND1.2) – Evaluation of compliance with predefined thresholds or target values – Documentation of the test results and conclusions drawn from them.	No analysis was conducted regarding existing biases and potential discrimination.		System/Component	Analysis	Maximum value	

ND1.4	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
When developing an AI system, the principles of universal design must be applied to ensure accessibility for people with disabilities.	**Technical measure – Operation** Where the AI system interacts with individuals, accessibility for people with disabilities is assessed and supported through appropriate measures. These measures include: – Designing the AI system's user interface to be accessible to people with disabilities (e.g., red-green color blindness or motor impairments) – Providing information about results and mechanisms, including information presented through the user interface, in a way that is tailored to the target audience and takes into account any potential limitations of the recipients – Designing the user interface and providing information is based on collaboration with user groups that include people with disabilities	**Technical measure – Operation** Where the AI system interacts with individuals, accessibility for people with disabilities is assessed and supported through appropriate measures. These measures include: – Designing the AI system's user interface to be inclusive of people with disabilities (e.g., red-green color blindness or motor impairments) – Providing information about results and mechanisms, including information presented through the user interface, in a way that is tailored to the target audience and takes into account any potential limitations of the recipients.	**Technical measure – Operation** Where the AI system interacts with individuals, accessibility for people with disabilities is assessed and supported through appropriate measures. These measures include: – Information about results and mechanisms, including, for example, a user interface, is presented in a way that is tailored to the target audience and takes into account any potential limitations of the recipients.	Principles of universal design were not considered during development.		System	Measure	Normal	

Non-discrimination

ND1.5	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
The staff involved in developing the AI system and implementing measures to reduce the risk of unjustified bias and discrimination must be informed and trained.	**Organisational measures – training** Target-group-oriented training program is offered to all internal and external employees involved in the development of the AI system. This program covers at least the following aspects: – Communicating the purpose of the AI system – Raising awareness of various potential biases and the recognition of potential discrimination in the context of AI systems – Handling relevant data types (e.g., training and validation data, operational data, customer data) in accordance with applicable guidelines and legal and regulatory requirements – Requirements for conducting training, validation, and testing of the AI model in the context of biases, fairness, and discrimination. The program is regularly updated based on insights gained and changes to guidelines. The content used and participation in the program are documented.	**Organisational measures – training** Training program is offered to all internal and external employees involved in the development of the AI system. This program covers at least the following aspects: – Raising awareness of various potential biases and the recognition of potential discrimination in the context of AI systems – Handling relevant data types (e.g., training and validation data, operational data, customer data) in accordance with applicable guidelines and legal and regulatory requirements – Requirements for conducting training, validation, and testing of the AI model with regard to biases, fairness, and discrimination. The content used and participation in the program are documented.	**Organisational measures – training** Training program is run for all internal and external employees involved in the development of the AI system. This program includes at least the following aspects: – Raising awareness of various possible biases and the recognition of potential discrimination in the context of AI systems – Requirements for conducting training, validation, and testing of the AI model in the context of biases, fairness, and discrimination. The content used and participation in the program are documented.	No guidance was provided to the staff responsible for AI development regarding biases and discrimination.		System	Measure	Normal	

ND1.6	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
Technical measures to avoid unjustified biases and discrimination must be implemented during the development of the AI system.	**Technical measures – data** The data used for training, validation, and testing reflect the fairness objective, and the test data is suitable for uncovering potential biases in the AI system. This includes at least: – Documentation of the data, selection or collection procedures, and processing steps (see DA 1.1) – Implementation of processing steps to reduce unjustified biases in the data – Justification of the appropriateness and benefits of measures applied to training and test data with regard to the risk of unjustified bias and discrimination.	**Technical measures – data** The data used for training, validation, and testing reflect the fairness objective, and the test data is suitable for uncovering potential biases in the AI system. This includes at least: – Documentation of the data, selection or collection procedures, and processing steps (see DA 1.1) – Implementation of processing steps to reduce unjustified biases in the data	**Technical measures – data** The data used for training, validation, and testing reflect the fairness objective, and the test data is suitable for uncovering potential biases in the AI system. This includes at least: – Documentation of the data, selection or collection procedures, and processing steps (see DA 1.1) – Implementation of processing steps to reduce unjustified biases in the data	No measures were taken to avoid unjustified distortions in the development.		System/Component	Measure	Normal	DA1.1

ND1.7	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
Mechanism for user feedback must be available to report problems related to potential discrimination.	MA1.4					System	Measure	Normal	

ND1.8	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
The AI system must (be able to) be monitored during development and operation to prevent unwanted bias and discrimination.	MA2.3					System	Measure	Normal	

ND1.9	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
It must be assessed whether the measures taken have reduced the identified risks to an acceptable level and whether the quality of the AI system meets the set targets.	**Evaluation** – Summary review of the effects of implementing the technical and organisational measures – Assessment of the interactions of the measures taken, and consideration of the mitigation of risks in conjunction with the risks of other quality dimensions – Identification and description of the residual risk after implementation of the technical and organisational measures – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	**Evaluation** – Assessment of interactions between measures, and weighing the mitigation of risks in conjunction with the risks of other quality dimensions – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	**Evaluation** – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	No evaluation was conducted.		System	Evaluation	Maximum value	

Transparency

TR1	Traceability & Documentation								
TR1.1	**A**	**B**	**C**	**D**	**Additional information**	**Reference level**	**Type**	**Weighting**	**Link**
The purpose and application area of the AI system are clearly defined and described.	**Analysis – Definition** The intended use is defined in terms of: i) improvements and benefits that can be achieved with the system; ii) functionalities of the AI system by means of which these improvements and benefits are to be realized; iii) intended/potential/permitted user groups and data subjects. The scope is defined with reference to the intended use and includes: – the expected input data and intended system outputs (including format and content); – a comprehensive description of the contexts/environments for which the AI system is suitable and also those in which the AI system may not be used; – a description of foreseeable abusive or misguided use of the AI system; – if applicable: definition of an ODD. The definitions are accessible to relevant stakeholders.	**Analysis – Definition** The intended use is defined in terms of the AI system's functionalities, as well as the intended/potential/permitted user groups and data subjects. The scope of application is defined in relation to the intended use and includes: – the expected input data and desired system outputs (including format and content) – a general description of the application contexts/environments for which the AI system is suitable and also those in which the AI system may not be used.	**Analysis – Definition** The intended use is defined in terms of the AI system's functionalities. The scope of application is defined in relation to the intended use and includes: – the expected input data and desired system outputs (including format and content) – a general description of the deployment contexts/environments for which the AI system is suitable.	The intended use and scope of application are not clearly defined.	See also Glossary	System	Analysis	Maximum value	relevant worldwide; especially TR1.2, TR2.1, VE1.1, VE2.1, CY1.1, DA2.1, DA3.1, ND1.1, MA1.2, MA2.2
TR1.2	**A**	**B**	**C**	**D**	**Additional information**	**Reference level**	**Type**	**Weighting**	**Link**
The risks associated with a lack of traceability and documentation of the AI system must be analysed.	**Analysis – Risk** Detailed risk analysis including quantification of probabilities of occurrence and risks: – Identification of potential risks and their causes – Assignment of risk responsibility – Estimation of the probability of occurrence – Estimation of the probability of detection – Estimation of the impact – Structured grading and prioritisation of risks. At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Lack of user trust – Unclear accountability – Concealment of biases or security vulnerabilities – Non-compliance with regulatory requirements – Interoperability issues	**Analysis – Risk** Limited risk analysis focusing on protection needs without quantifying probabilities: – Identification of potential risks – Assignment of risk responsibility – Estimation of impact – Structured grading and prioritisation of risks. At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Lack of user trust – Unclear accountability – Concealment of biases or security vulnerabilities – Non-compliance with regulatory requirements – Interoperability issues	**Analysis – Risk** Primarily qualitative assessment without probabilities: – Identification of potential hazards – Assignment of risk responsibility – Qualitative assessment of impact – Qualitative grading and prioritisation of risks. At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Lack of user trust – Unclear accountability – Concealment of biases or security vulnerabilities – Non-compliance with regulatory requirements – Interoperability issues	No risk analysis was conducted.		System	Analysis	Maximum value	TR1.1
TR1.3		**B**	**C**	**D**	**Additional information**	**Reference level**	**Type**	**Weighting**	**Link**
The architecture of the AI system must be documented.	**Technical measures – Other** The system architecture is documented, including: – AI components (see TR1.4) – Hardware and software integration and requirements for these – Interfaces, e.g., for users or to other systems – Information flow between individual system components – Justification for the choice of architecture with a description of the role of each component in the context of the AI system's purpose – If applicable, the license under which the system may be used – Planned modalities of the AI system for "independent" adaptation/further development during operation, including online learning	**Technical measures – Other** The system architecture is documented, including: – AI components (see TR1.4) – Hardware and software integration and requirements for these – Interfaces, e.g., for users or to other systems – Information flow between individual system components – If applicable, the license under which the system may be used	**Technical measures – Other** The system architecture is documented, including: – AI components (see TR1.4) – Hardware and software integration – Interfaces, e.g., for users or to other systems – Information flow between individual system components – If applicable, the license under which the system may be used	The system's architecture is not documented.		System	Analysis	Maximum value	VE1.1, VE2.1, VE1.2, VE1.3, CY1.5, CY1.6
TR1.4	**A**	**B**	**C**	**D**	**Additional information**	**Reference level**	**Type**	**Weighting**	**Link**
The characteristics of the AI model(s) must be documented.	**Technical measures – model** Properties are documented, including: – Model name – Model version (history) including date – Architecture description and diagram of the AI model – Expected input data – Expected output data – Training and test data used – Expected performance – Tests performed, test results obtained, and conclusions drawn. The choice of architecture and design of the AI system and/or model must be justified, including a description of the model's advantages and considerations regarding potential conflicts of objectives.	**Technical measures – model** Properties are documented, including: – Model name – Model version (history) including date – Architecture description and diagram of the AI model – Expected input data – Expected output data – Training and test data used – Expected performance	**Technical measures – model** Properties are documented, including: – Model name – Latest model version with date – General architectural description of the AI model – Expected input data – Expected output data – Training and test data used – Expected performance	There is no documentation of the AI model(s).	The architectural description of the AI model can include, for example, the type of model, the type and number of levels/layers in neural networks, as well as functions for activation or reward.	Component	Analysis	Maximum value	VE1.2, VE1.3, CY1.5, CY1.6
TR1.5	**A**	**B**	**C**	**D**	**Additional information**	**Reference level**	**Type**	**Weighting**	**Link**
The characteristics of the datasets used must be documented.	DA1.1					Component	Analysis	Maximum value	
TR1.6	**A**	**B**	**C**	**D**	**Additional information**	**Reference level**	**Type**	**Weighting**	**Link**
The design and development process of the AI system must be described.	**Organisational measure – System-related processes** The documentation for the design and development process includes: – Overview of the responsible person(s) and their tasks – Traceability and systematic approach of individual work elements (e.g., through a project management tool) – If applicable, description and justification of the adjustments to the AI system's architecture during development **Technical measures – Data** The documentation for the design and development process includes: – Data used and its origin – Relevant data preparation processes (e.g., cleaning, annotation, tagging, enrichment, aggregation, feature engineering) – Tracing the data lineage and possible data recovery **Technical measures – Model** The documentation for the design and development process includes: – A system for version control and tracking of changes to the AI models, and recording of the training and the data used in each case – Description of the changes to AI models during the development process	**Organisational measure – System-related processes** The documentation for the design and development process includes: – Overview of the responsible person(s) and their tasks – Traceability and systematic approach of individual work elements (e.g. through project management tool) **Technical measures – Data** The documentation for the design and development process includes: – Data used and its origin – Relevant data preparation processes (e.g., cleaning, annotation, tagging, enrichment, aggregation, feature engineering) – Data lineage tracing **Technical measures – Model** The documentation for the design and development process includes: – A system for version control and tracking of changes to the AI models and recording of the training and the data used in each case.	**Organisational measure – System-related processes** The documentation for the design and development process includes: – Overview of the responsible person(s) and their tasks – Traceability and systematic approach of individual work elements (e.g. through project management tool) **Technical measures – Data** The documentation for the design and development process includes: – Data used and its origin – Tracing the data lineage **Technical measures – Model** The documentation for the design and development process includes: – A system for version control and tracking of changes to the AI models and recording of the training and the data used in each case.	The design and development process was not documented.	For the documentation of data sets and models, common formats such as data sheets or model cards can be used.	System/Component	Analysis	Normal	VE1.4, VE1.5

Transparency

TR1.7	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
The AI system must include functions for monitoring, capturing, and recording its behavior.	**Organisational measures – System-related processes** A purpose-oriented logging concept for the AI system and the data used within it during development and operation must be created. The concept must include at least the following information: – Definition of the data and metrics to be monitored – Definition of a purpose-appropriate log retention period – Definition of a purpose-appropriate recording frequency – Definition of the recording structure – Description of the interface through which the logging can be accessed **Technical measures – Operation** An interface exists that allows for the collection and monitoring of at least the following information during operation: – If technically possible and meaningful in the application context, user interactions and model requests, including the model version used for the request, inputs and outputs over a defined period – Malfunctions (e.g., during the processing of automatic or manual actions) – User access to data, services, or functions – Changes to security-relevant configuration parameters (e.g., error handling and logging mechanisms, user authentication, action authorization, cryptography, and communication security) – Violations of the functionality, security, or quality objectives of the AI system. If logging certain information is not technically feasible, this must be justified. The metadata to be collected in each case captures, within the scope of technical possibilities, at least information about: – Type, time, duration, storage location, and actor/system of recorded events or actions.	**Organisational measures – System-related processes** A purpose-oriented logging concept for the AI system and the data used within it during development and operation must be created. The concept must include at least the following information: – Definition of the data and metrics to be monitored – Definition of a purpose-appropriate log retention period – Definition of a purpose-appropriate recording frequency – Definition of the recording structure – Description of the interface through which the logging can be accessed **Technical measures – Operation** An interface exists that allows for the recording and monitoring of at least the following information during operation: – If technically possible and meaningful within the application context, user interactions and model requests, including the model version used for the request, inputs, and outputs – Malfunctions (e.g., during the processing of automatic or manual actions). If logging certain information is not technically feasible, this must be justified. The metadata to be recorded in each case captures, within the scope of technical possibilities, at least the following information: – Type, time, duration, storage location, and actor/system of recorded events or actions.	**Organisational measures – System-related processes** Purpose-oriented logging concept for the AI system and the data used within it must be developed during development and operation. This concept must include at least the following information: – Definition of the data and metrics to be monitored – Description of the interface through which the logging can be accessed **Technical measures – Operation** An interface exists that allows for the recording and monitoring of at least the following information during operation: – If technically possible and meaningful within the application context, user interactions and model requests, including the model version used for the request, inputs, and outputs – Malfunctions (e.g., during the processing of automatic or manual actions). If logging certain information is not technically feasible, this must be justified. The metadata to be recorded in each case captures, within the scope of technical possibilities, at least information about: – Type, time, duration, storage location, and actor/system of recorded events or actions.	A record of important system data has not been prepared.		System/Component	Measure	Normal	VE1.6, DA2.6

TR1.8	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
It must be assessed whether the measures taken have reduced the identified risks to an acceptable level and whether the quality of the AI system meets the set targets.	**Evaluation** – Summary review of the effects of implementing the technical and organisational measures – Assessment of the interactions of the measures taken, and consideration of the mitigation of risks in conjunction with the risks of other quality dimensions – Identification and description of the residual risk after implementation of the technical and organisational measures – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	**Evaluation** – Weighing the interactions of the measures taken, and weighing the mitigation of risks in conjunction with the risks of other quality dimensions – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	**Evaluation** – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	No evaluation was conducted.		System	Evaluation	Maximum value	

TR2 Explainability & Interpretability

TR2.1	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
The risks associated with a lack of interpretability or explainability of the AI system must be analysed.	**Analysis – Risk** Detailed risk analysis for explainability, including quantification of probabilities of occurrence and risks: – Identification of potential risks and their causes – Assignment of risk responsibility – Estimation of the probability of occurrence – Estimation of the probability of detection – Estimation of the impact – Structured grading and prioritisation of risks. At least the following sources of risk must be considered, taking into account the intended use (see TR1.1): – Lack of trust among users – Lack of understanding of the system's outputs – Incorrect decisions due to an insufficient understanding of the system's outputs	**Analysis – Risk** Limited risk analysis for explainability, focusing on protection needs without quantifying probabilities: – Identification of potential risks – Assignment of risk responsibility – Impact assessment – Structured grading and prioritisation of risks. At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Lack of user trust – Lack of understanding of the system's outputs – Incorrect decisions due to an insufficient understanding of system's outputs	**Analysis – Risk** Primarily qualitative assessment without probabilities for explainability: – Identification of potential hazards – Assignment of risk responsibility – Qualitative assessment of impact – Qualitative grading and prioritisation of risks. At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Lack of user trust – Lack of understanding of the system's outputs – Incorrect decisions due to an insufficient understanding of system's outputs	No risk analysis was conducted.		System	Analysis	Maximum value	TR1.1

TR2.2	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
The interpretability and explainability of the AI system must be analysed with regard to the user groups, the purpose, and the risks.	**Technical measure – tests** The degree to which the AI system is interpretable or explainable for the permitted user groups must be analysed taking into account the risk analysis (TR2.1) and intended purpose (TR1.1). For this purpose, the relationship between inputs and outputs of the AI components of the system must be examined based on the inherent characteristics of the component or suitable technical tests. The selected tests should have at least the following characteristics: – If justified by the selection, the metrics and tests should include both important basic methods and advanced methods (regarding complexity, information content, implementation effort, e.g., comprehensive testing tools) – if necessary, defining thresholds (as minimum requirements for explainability) and justifying the thresholds, taking into account the scope and purpose – if applicable: differentiation of metrics and thresholds according to different application scenarios with reference to the definition of the scope.	**Technical measure – tests** The degree to which the AI system is interpretable or explainable for the permitted user groups must be analysed taking into account the risk analysis (TR2.1) and intended purpose (TR1.1). For this purpose, the relationship between inputs and outputs of the AI components of the system must be examined based on the inherent characteristics of the component or suitable technical tests. The selected tests should have at least the following characteristics: If justified by the selection criteria, the metrics and tests should include both important basic methods and advanced methods (in terms of complexity, information content, implementation effort, e.g., comprehensive testing tools).	**Technical measure – tests** The degree to which the AI system is interpretable or explainable for the permitted user groups must be analysed taking into account the risk analysis (TR2.1) and intended purpose (TR1.1). For this purpose, the relationship between inputs and outputs of the AI components of the system must be examined based on the inherent characteristics of the component or suitable technical tests. – Basic methods are sufficient (in terms of complexity, information content, implementation effort, e.g. a simple metric)	The interpretability and explainability of the AI system were not analysed.		System/Component	Analysis	Normal	

Transparency

TR2.3	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
Measures must be taken to make the AI system interpretable or adequately explainable with regard to its intended use.	**Technical measures – model** The AI system must be made adequately explainable within the context of its intended use and with the aim of minimizing risk. This can include: i) the selection of inherently interpretable algorithms, or ii) the selection of algorithms for which corresponding approaches exist for establishing the (local) explainability of individual outputs, and their subsequent implementation. **Technical measures – User instructions** The AI system must provide non-expert users with the necessary information to understand model outputs. A user interface is available for this purpose, containing the following aspects: – Description of the methods used to make system outputs understandable – Explanation of the generated explanation – Requirements for users (e.g., computer science knowledge) to be able to understand them in principle. **Organisational measures – training** Targeted training courses have been prepared for the qualification and training of personnel responsible for the operation and oversight of the AI system, as well as for human end users. These courses focus on the interpretation of results from the AI system and possible functions that support the explainability of the results. At a minimum, the following aspects are covered: – Relevance and benefits of the explainability of the AI system in the operational context – Methods and tools used for explainability – Practice in applying the methods and tools for explanation	**Technical measures – model** The AI system must be made adequately explainable within the context of its intended use and with the aim of minimizing risk. This can include: i) the selection of inherently interpretable algorithms, or ii) the selection of algorithms for which corresponding approaches exist for establishing the (local) explainability of individual outputs, and their subsequent implementation. **Organisational measures – User instructions** The AI system must provide non-expert users with the necessary information to understand model outputs. This includes the following information: – Description of the methods used to make system outputs understandable – Explanation of the generated explanation – Requirements for users (e.g., computer science knowledge) to understand the output in principle. User instructions can also be covered entirely through training.	**Technical measures – model** The AI system must be made adequately explainable within the context of its intended use and with the aim of minimizing risk. This can include: i) the selection of inherently interpretable algorithms, or ii) the selection of algorithms for which corresponding approaches exist for establishing the (local) explainability of individual outputs, and their subsequent implementation. **Organisational measures – User instructions** The AI system must provide non-expert users with the necessary information to understand model outputs. This includes the following information: – A description of the methods used to make system outputs understandable – Requirements for users (e.g., computer science knowledge) to understand them in principle. User instructions can also be covered entirely through training.	No appropriate measures were taken to establish interpretability or explainability.		System/Component	Measure	Normal	

TR2.4	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
It must be assessed whether the measures taken have reduced the identified risks to an acceptable level and whether the quality of the AI system meets the set targets.	**Evaluation** – Summary review of the effects of implementing the technical and organisational measures – Assessment of the interactions of the measures taken, and consideration of the mitigation of risks in conjunction with the risks of other quality dimensions – Identification and description of the residual risk after implementation of the technical and organisational measures – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	**Evaluation** Weighing the interactions of the measures taken, and weighing the mitigation of risks in conjunction with the risks of other quality dimensions – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	**Evaluation** – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	No evaluation was conducted.		System	Evaluation	Maximum value	

Human oversight and control

MA1	Human capacity for action								
MA1.1	**A**	**B**	**C**	**D**	**Additional information**	**Reference level**	**Type**	**Weighting**	**Link**
The degree of autonomy of the AI system must be analysed and documented.	**Analysis – Definition** – Determination of the level of autonomy or degree of human control and involvement in decision-making processes – Justified differentiation from other levels of autonomy – Derivation of specific legal requirements regarding human oversight and control	**Analysis – Definition** – Determination of the level of autonomy or degree of human control and involvement in decision-making processes – Derivation of specific legal requirements regarding human oversight and control	**Analysis – Definition** Determination of the level of autonomy or degree of human control and involvement in decision-making processes – Derivation of specific legal requirements regarding human oversight and control	No classification of the level of autonomy was made.	Autonomy levels are presented in ISO/IEC 22989.	System	Analysis	Maximum value	MA2.1
MA1.2	**A**	**B**	**C**	**D**	**Additional information**	**Reference level**	**Type**	**Weighting**	**Link**
The risks associated with the restriction of human agency by the AI system must be analysed.	**Analysis risk** Detailed risk analysis including quantification of probabilities of occurrence and risks: – Identification of potential risks and their causes – Assignment of risk responsibility – Estimation of the probability of occurrence – Estimation of the probability of detection – Estimation of the impact – Structured grading and prioritisation of risks At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Ethical dilemmas / decision-making – Unclear accountability – Automation bias / excessive dependence – Pretending to be human or unclear authorship – Lack of traceability of decisions	**Analysis risk** Limited risk analysis focusing on protection needs without quantifying probabilities: – Identification of potential risks – Assignment of risk responsibility – Estimation of impact – Structured grading and prioritisation of risks. At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Ethical dilemmas / decision-making – Unclear accountability – Automation bias – Pretending to be human or unclear authorship – Lack of traceability of decisions	**Analysis risk** Primarily qualitative assessment without probabilities: – Identification of potential hazards – Assignment of risk responsibility – Qualitative assessment of impact – Qualitative grading and prioritisation of risks. At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Ethical dilemmas / decision-making – Unclear accountability – Automation bias – Pretending to be human or unclear authorship – Lack of traceability of decisions	No risk analysis was conducted.		System	Analysis	Maximum value	TR1.1
MA1.3	**A**	**B**	**C**	**D**	**Additional information**	**Reference level**	**Type**	**Weighting**	**Link**
Affected persons and users must be able to exercise their rights relating to personal data, including rights relating to data management, erasure and use, as well as the right to information, even during the operation of the AI system.	DA2.6					System	Measure	Normal	TR1.7, MA1.4, MA1.6, MA2.3
MA1.4	**A**	**B**	**C**	**D**	**Additional information**	**Reference level**	**Type**	**Weighting**	**Link**
It must be possible to give feedback on the AI system, ask questions, and report problems.	**Technical measures – Operation** – There is at least one channel through which the provider can be contacted for the purpose of providing feedback. – Channels are clearly identifiable within the context of the AI system by anyone with a legitimate interest. – It is evident that the possible channels for providing feedback are provided. **Organisational measures – System-related processes** – Review and verification of feedback and assignment of responsibility for it – Guaranteed and individualized response to inquiries – Process for forwarding and incorporating feedback into the further development of the AI system	**Technical measures – Operation** – There is at least one channel through which the provider can be contacted for the purpose of providing feedback. – Channels are clearly identifiable within the context of the AI system by anyone with a legitimate interest. **Organisational measures – System-related processes** – Review and verification of the feedback and assignment of responsibility for it takes place	**Technical measures – Operation** – The contact information is attached to the AI system and can be used for feedback purposes.	There are no channels available to provide feedback on the AI system.	This could include affected individuals or deployers. Possible feedback channels include surveys/questionnaires, feedback forms, service hotlines, email addresses, social media presence, suggestion boxes, chatbots, or comment sections.	System	Measure	Normal	DA2.6
MA1.5		**B**	**C**	**D**	**Additional information**	**Reference level**	**Type**	**Weighting**	**Link**
They must have the opportunity to learn about the application of the AI system, how the AI system supports a decision, and the interpretation of its outputs.	**Organisational measures – User instructions** Provision of instructions and information tailored to the target audience, containing at least the following: – Contact information – Purpose of the AI system – Intended scope of application of the AI system – Expected performance and functionality of the AI system – Known risks and implications of using the AI system (risks arising from the use of the system and from its outputs) – Summary of the expected input data – Description of how to interpret the outputs of the AI system – Description of the degree of autonomy and the decision-making processes of the AI system – Necessary measures for human oversight of the system – Requirements for users (e.g. training) for using the AI system – Necessary maintenance measures (including software updates) **Technical measures – Operation** – Integration of notifications to natural persons interacting with the AI system, informing them that the results and any decision-making are based on the AI system – Provision of information on outputs or decisions of the AI system that enables correct interpretation – Warning about so-called “automation bias” and its potential consequences	**Organisational measures – User instructions** Provision of instructions and information tailored to the target audience, containing at least the following information: – Contact information – Purpose of the AI system – Intended application area of the AI system – Expected performance and functionality of the AI system – Description of how to interpret the outputs of the AI system – Necessary measures for human oversight of the system – Necessary maintenance measures (including software updates) **Technical measures – Operation** – Integration of notifications to individuals interacting with the AI system informing them that the results and potential decision-making are based on the AI system – Warning about so-called "automation bias" and its potential consequences	**Organisational measures – User instructions** Provision of instructions and information containing at least the following: – Contact information – Purpose of the AI system – Intended scope of application of the AI system – Description of how to interpret the outputs of the AI system – Necessary measures for human oversight of the system **Technical measures – Operation** – Integration of notifications to individuals interacting with the AI system informing them that the results and potential decision-making are based on the AI system – Warning about so-called "automation bias" and its potential consequences	No notifications or alerts are provided to users or affected persons	The provision of notifications and instructions can be done via a graphical interface.	System	Measure	Normal	TR2, MA1.6
MA1.6	**A**	**B**	**C**	**D**	**Additional information**	**Reference level**	**Type**	**Weighting**	**Link**
Mechanisms must be in place that enable users or affected persons to reject, challenge, correct or interrupt the decisions and outputs of the AI system.	**Technical measures – Operation** The appropriate mechanisms for challenging, correcting, or interrupting the AI system must be tailored to the identified risks, the purpose of the AI system, and the target audience of the action. Measures to be considered include at least: – An objection form through which decisions can be challenged; – Referral to a human review process where a qualified and authorised person must confirm a decision or process an output; – An emergency stop function that allows an authorised person to put the system into a safe state; – An interactive graphical user interface to illustrate the actions.	**Technical measures – Operation** The appropriate mechanisms for challenging, correcting, or interrupting the AI system must be tailored to the identified risks, the purpose of the AI system, and the target audience of the action. Measures to be considered include at least: – Appeal form for submitting challenges to decisions – Forwarding to a human review process where a qualified and authorised person must confirm a decision or process an output – Emergency stop function which puts the system into a safe state by an authorised person – Interactive graphical user interface to illustrate the actions	**Technical measures – Operation** Measures under consideration include at least: – Appeal form for submitting challenges to decisions – Forwarding to a human review process where a qualified and authorised person must confirm a decision or process an output – Emergency stop function which a natural person can use to put the system into a safe state – Interactive graphical user interface to illustrate the actions	There are no mechanisms in place to challenge, correct, or interrupt the AI system.	Both human-in-the-loop (HITL) and human-on-the-loop (HOTL) designs are possible.	System	Measure	Maximum value	MA1.4, VE2.2
MA1.7	**A**	**B**	**C**	**D**	**Additional information**	**Reference level**	**Type**	**Weighting**	**Link**
It must be assessed whether the measures taken have reduced the identified risks to an acceptable level and whether the quality of the AI system meets the set targets.	**Evaluation** – Summary review of the effects of implementing the technical and organisational measures – Assessment of the interactions of the measures taken, and consideration of the mitigation of risks in conjunction with the risks of other quality dimensions – Identification and description of the residual risk after implementation of the technical and organisational measures – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	**Evaluation** – Weighing the interactions of the measures taken, and weighing the mitigation of risks in conjunction with the risks of other quality dimensions – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	**Evaluation** – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	No evaluation was conducted.		System	Evaluation	Maximum value	

Human oversight and control

MA2	Human oversight								
MA2.1	**A**	**B**	**C**	**D**	**Additional information**	**Reference level**	**Type**	**Weighting**	**Link**
The degree of autonomy of the AI system must be analysed and documented.	MA1.1					System	Analysis	Maximum value	MA1.1
MA2.2	**A**	**B**	**C**	**D**	**Additional information**	**Reference level**	**Type**	**Weighting**	**Link**
The risks that the AI system could cause harm without human oversight (and control) must be analysed.	**Analysis – Risk** Detailed risk analysis including quantification of probabilities of occurrence and risks: – Identification of potential risks and their causes – Assignment of risk responsibility – Estimation of the probability of occurrence – Estimation of the probability of detection – Estimation of the impact – Structured grading and prioritisation of risks At least the following sources of risk must be considered, taking into account the intended use (see TR1.1): – Implications of (semi-)autonomous decisions made by the system without control, intervention, or correction by natural persons – Ethical dilemmas / decision-making – Lack of traceability of decisions – Liability for decisions made by the AI system – Authorship of outputs made by the AI system – Error susceptibility of the AI system without this being detected by humans – Incorrect/improper use (e.g., due to lack of qualification)	**Analysis – Risk** Limited risk analysis focusing on protection needs without quantifying probabilities: – Identification of potential risks – Assignment of risk responsibility – Estimation of impact – Structured grading and prioritisation of risks. At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Implications of (semi-)autonomous system decisions without control, intervention, or correction by natural persons – Ethical dilemmas / decision-making – Lack of traceability of decisions – Liability for decisions of the AI system – Authorship of outputs of the AI system – Error susceptibility of the AI system without this being detected by humans – Incorrect/improper use (e.g., due to lack of qualification)	**Analysis – Risk** Primarily qualitative assessment without probabilities: – Identification of potential hazards – Assignment of risk responsibility – Qualitative assessment of the impact – Qualitative grading and prioritisation of risks. At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Implications of (semi-)autonomous decisions by the system without control, intervention, or correction by natural persons – Ethical dilemmas / decision-making – Lack of traceability of decisions – Liability for decisions of the AI system – Authorship of outputs of the AI system – Error susceptibility of the AI system without this being detected by humans – Incorrect/improper use (e.g., due to lack of qualification)	No risk analysis was conducted.		System	Analysis	Maximum value	TR1.1
MA2.3	**A**	**B**	**C**	**D**	**Additional information**	**Reference level**	**Type**	**Weighting**	**Link**
The AI system must be able to be monitored during development and operation.	**Technical measure – Operation** Continuous monitoring must be provided for during the implementation of the AI system and be implemented based on appropriate logging (see TR1.7). This includes testing the following options: – Performance monitoring, including monitoring of the models and data (i.e., incoming production data and, if applicable, the expanding training database) – Bias monitoring, including monitoring of the data (i.e., incoming production data and, if applicable, the expanding training database) in the context of preventing unfair discrimination and biases – Conducting tests (e.g., sanity checks) used within the monitoring framework to detect model and concept drift or harmful input data – If applicable, conducting tests to detect misuse and harmful input data – If applicable, quality control of the expanding training database **Organisational measure – System-related processes** In addition to the technical implementation of monitoring, the following aspects should be prepared: – A concept for the (automatic) monitoring and testing of major changes to the AI system, including software and hardware components, but especially in the case of online learning – Recommended tests must be documented as part of a continuous testing plan, particularly in the case of online learning – Where possible, mechanisms in the form of meaningful threshold definitions or scenarios in which (human) review and mitigation measures should be implemented – Mechanisms for sharing new information about potential security-related incidents and their prevention	**Technical measure – Operation** Continuous monitoring must be provided for during the implementation of the AI system and be implemented based on appropriate logging (see TR1.7). This includes the following: – Performance monitoring, including monitoring of the models and data i.e., incoming production data and, if applicable, the expanding training database) – Bias monitoring, including monitoring of the data (i.e., incoming production data and, if applicable, the expanding training database) in the context of preventing unfair discrimination and biases – Conducting tests (e.g., sanity checks) used within the monitoring framework to detect model and concept drift, or harmful input data **Organisational measure – System-related processes** In addition to the technical implementation of monitoring, the following aspects should be prepared: – A concept for the (automatic) monitoring and testing of major changes to the AI system, including software and hardware components, but especially in the case of online learning – Recommended tests must be documented as part of a continuous testing plan, especially in the case of online learning	**Technical measure – Operation** Continuous monitoring must be provided for during the implementation of the AI system and be implemented based on appropriate logging (see TR1.7). This includes the following: – Performance monitoring, including monitoring of the models and data (i.e., incoming production data and, if applicable, the expanding training database) - Bias monitoring, including monitoring of the data (i.e., incoming production data and, if applicable, the expanding training database) in the context of preventing unfair discrimination and biases – Conducting tests (e.g., sanity checks) used within the monitoring framework to detect model and concept drift, or harmful input data			System	Measure	Maximum value	VE1.7, CY1.5 (attacks during the operational phase), DA1.4
MA2.4	**A**	**B**	**C**	**D**	**Additional information**	**Reference level**	**Type**	**Weighting**	**Link**
The persons responsible for the operation and oversight of the AI system must have access to an understandable description of the AI system and be adequately prepared to perform their duties.	**Organisational measures – User instructions** Provision of target-group-oriented instructions and information with at least the following content: – Contact information – Purpose of the AI system – Intended application area of the AI system – Expected performance and functionality of the AI system – Known risks and implications for the use (of the results) of the AI system – Summary of the expected input data – Description of how to interpret the outputs of the AI system – Description of the degree of autonomy and decision-making processes of the AI system – Necessary/possible measures for human oversight and control of the system, when these should/can be applied, and how it will be recognized that human intervention is required – Requirements for users (e.g., training) for using the AI system – Necessary maintenance measures (including software updates) **Organisational measures – training** To qualify and train the personnel responsible for operating and monitoring the AI system, target group-oriented training materials have been developed. These materials focus on proper use and increased security awareness. The materials cover at least the following aspects: – Proper handling of system components in the production environment – Proper handling of relevant data types (e.g., training and validation data, operational data, customer data), also in accordance with applicable guidelines and legal and regulatory requirements – Appropriate execution of training, validation, and testing during the operation of the AI system and its components – Monitoring the performance of the AI system and its components – Information on potential threat scenarios and associated mitigation measures (e.g., interventions) – Response in the event of security incidents	**Organisational measures – User instructions** Provision of instructions and information tailored to the target audience, containing at least the following information: – Contact information – Purpose of the AI system – Intended scope of application of the AI system – Description of how to interpret the outputs of the AI system – Necessary (and possible) measures for human oversight and control of the system, when these should/can be applied, and how it will be recognized that human intervention is required – Necessary maintenance measures (including software updates) **Organisational measures – training** To qualify and train the personnel responsible for operating and monitoring the AI system, target group-oriented training materials have been developed. These materials focus on proper use and increased safety awareness. The materials cover at least the following aspects: – Proper handling of system components in the production environment – Monitoring the performance of the AI system and its components – Information on potential threat scenarios and corresponding mitigation measures (e.g., interventions) – Procedures in the event of security incidents	**Organisational measures – User instructions** Provision of instructions and information with at least the following content: – Contact information – Purpose of the AI system – Intended scope of application of the AI system – Description of how to interpret the outputs of the AI system – Necessary (or possible) measures for human oversight and control of the system, when these should/can be applied, and how it will be recognized that human intervention is required – Necessary maintenance measures (including software updates)	There are no materials available to prepare those responsible for the operation and oversight of the AI system.		System	Measure	Normal	MA1.5

Human oversight and control

MA2.5	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
It must be assessed whether the measures taken have reduced the identified risks to an acceptable level and whether the quality of the AI system meets the set targets.	**Evaluation** – Summary review of the effects of implementing the technical and organisational measures – Assessment of the interactions of the measures taken, and consideration of the mitigation of risks in conjunction with the risks of other quality dimensions – Identification and description of the residual risk after implementation of the technical and organisational measures – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	**Evaluation** Weighing the interactions of the measures taken, and weighing the mitigation of risks in conjunction with the risks of other quality dimensions – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	**Evaluation** – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	No evaluation was conducted.		System	Evaluation	Maximum value	

AI-specific cybersecurity

CY1 General AI-specific cybersecurity

CY1.1	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
Cybersecurity risks to the AI system must be analysed taking into account its intended use.	**Analysis – Risk** Detailed cybersecurity risk analysis, including quantification of probabilities of occurrence and risks: – Identification of potential risks and their causes – Assignment of risk responsibility – Estimation of the probability of occurrence – Estimation of the probability of detection – Estimation of the impact – Structured grading and prioritisation of risks. The risk assessment should, taking into account the intended use and scope (see TR1.1), consider at least the following: – All classic cybersecurity protection goals, i.e., confidentiality (e.g., protection of training data from malicious disclosure or misuse), integrity (e.g., protection of the AI system from malicious manipulation of the output to the attacker's advantage), and availability (e.g., protection of the AI system from maliciously caused overload) – Threat modeling to determine probabilities of occurrence – Vulnerability analysis to determine probabilities of occurrence – Analysis of the system assets to be protected to determine the impact. The risk assessment framework should ideally be embeddable in a cybersecurity management system and should ideally follow established standards.	**Analysis – Risk** Limited risk analysis focusing on protection needs without quantifying probabilities: – Identification of potential risks – Allocation of risk responsibility – Qualitative estimation of impact – Qualitative grading and prioritisation of risks. The risk assessment should, taking into account the intended use and scope (see TR1.1), consider at least the following: – All classic cybersecurity protection goals, i.e., confidentiality (e.g., protection of personal training data from malicious disclosure or misuse), integrity (e.g., protection of the AI system from malicious manipulation of the output to the attacker's advantage), and availability (e.g., protection of the AI system from maliciously caused overload) – Threat modeling to determine probabilities of occurrence – Vulnerability analysis to determine probabilities of occurrence – Analysis of the system assets to be protected to determine the impact. The risk assessment framework should ideally be embeddable in a cybersecurity management system and should ideally follow established standards.	**Analysis – Risk** Primarily qualitative assessment without probabilities – Identification of potential risks –Allocation of risk responsibility – Qualitative estimation of impact – Qualitative grading and prioritisation of risks. The risk assessment should, taking into account the intended use and scope (see TR1.1), consider at least the following: – All classic cybersecurity protection goals, i.e., confidentiality (e.g., protection of personal training data from malicious disclosure or misuse), integrity (e.g., protection of the AI system from malicious manipulation of the output to the attacker's advantage), and availability (e.g., protection of the AI system from maliciously caused overload) – Threat modeling to determine probabilities of occurrence – Vulnerability analysis to determine probabilities of occurrence – Analysis of the system assets to be protected to determine the impact. The risk assessment framework should ideally be embeddable in a cybersecurity management system and should ideally follow established standards.	No risk analysis was conducted.	Important standards include, for example, BSI IT Baseline Protection (BSI IT-Grundschutz), ISO 27k series, IEC 62443, ISO/SAE 21434, ETSI EN 303 645, NIST AI RMF, …	System	Analysis	Maximum value	TR1.1 (purpose)

CY1.2	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
A cybersecurity management system (CSMS) must be in place for the AI system, including all AI components and the embedded ICT system.	**Organisational measures – Governance & Technical measures – Operations** (Connection to cybersecurity management system) It must be ensured that the AI system can be embedded in a cybersecurity management system (CSMS), in particular: – Monitoring and logging must be implemented or planned and provide interfaces for a CSMS – the documentation of vulnerabilities, assets, risk analysis and the results of all mitigation measures should be fully available – a strategy for software and security updates should be in place. The AI system should provide interfaces to cover at least the following aspects of a CSMS: – A system lifecycle-based system for regular cybersecurity quality control, including planned reviews of security measures and protocols. – Fundamental cybersecurity asset management, including security controls and monitoring systems themselves – Vulnerability management (identification throughout the entire lifecycle, especially training/evaluation and deployment phases; penetration tests should be mandatory where necessary) – A cybersecurity monitoring system – Access management to the various assets and components of the system – Aspects of personnel management, such as training, education, and role assignment in the development and support of the AI product – An analysis of the expected and actual timeframe within which security updates are provided for the AI system – Strategies for regular review of security measures and protocols – Documentation of the CSMS. The information must be directly accessible from the AI application. – The CSMS can be demonstrated through compliance with relevant cybersecurity standards; corresponding proof for the developed AI system should automatically be rated as A (e.g., BSI IT Baseline Protection (BSI IT-Grundschutz), ISO 27k series, IEC 62443, ISO/SAE 21434, ETSI EN 303 645, NIST AI RMF, ...)	**Organisational measures – Governance & Technical measures – Operations** (Connection to cybersecurity management system) As much information as possible should be provided so that the AI system can be embedded in a cybersecurity management system (CSMS), in particular: – Monitoring and logging must be provided as a minimum and offer interfaces for a CSMS – the documentation of vulnerabilities, assets, risk analysis and the results of all mitigation measures should be fully available. The AI system should provide interfaces to cover at least the following aspects of a CSMS: – A system lifecycle-based system for regular cybersecurity quality control, including planned reviews of security measures and protocols. – Basic cybersecurity asset management, including security controls and monitoring systems themselves – Vulnerability management (identification throughout the entire lifecycle, especially training/evaluation and deployment phases; penetration tests should be mandatory where necessary) – Access management to the various assets and components of the system – An analysis of the expected and actual timeframe within which security updates are provided for the AI system/application – Strategies for regular review of security measures and protocols – Documentation of the CSMS. The information must be directly accessible from the AI application. – The CSMS can be demonstrated through compliance with relevant cybersecurity standards; corresponding proof for the developed AI system should automatically be rated as A (e.g., BSI IT Baseline Protection (BSI IT-Grundschutz), ISO 27k series, IEC 62443, ISO/SAE 21434, ETSI EN 303 645, NIST AI RMF, ...)	**Organisational measures – Governance & Technical measures – Operations** (Connection to cybersecurity management system) – As much information as possible on AI-specific cybersecurity should be provided in order to embed the AI system in larger system contexts, in particular the documentation of vulnerabilities, assets, risk analysis and the results of all mitigation measures.	Integration into a CSMS was in no way planned or facilitated.	Important standards include, for example, BSI IT Baseline Protection (BSI IT-Grundschutz), ISO 27k series, IEC 62443, ISO/SAE 21434, ETSI EN 303 645, NIST AI RMF, ...	System/Component	Measure	Normal	CY1.5 (data access), CY2.3, CY2.2 (AI-specific mitigation measures)

CY1.3	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
Measures must be in place to control the cybersecurity aspects of AI supply chain management, particularly with regard to vulnerabilities or harmful practices in the use of pre-trained models, open-source machine learning libraries, and third-party training data.	**Organisational measures – System-related processes** – A process should be in place for analyzing the cybersecurity of the software and hardware supply chains with regard to the development and production environment of the AI system. – The supply chain analysis includes pre-trained models, datasets, and software libraries used, as well as the hardware used for the training and, if applicable, production phases of the AI system. – A process should be in place for the continuous monitoring of the supply chain, including a plan for updates and reviews following the discovery of previously undiscovered vulnerabilities in the supply chain. **Technical measures – models, data** – A security audit of all identified, used pre-trained models, datasets, and software libraries from the supply chain must be demonstrated. **Technical measures – Operation** – Interfaces should be available that allow for the technical monitoring of the AI system's software and hardware, including vulnerabilities, through cybersecurity monitoring.	**Organisational measures – System-related processes** – A process should be in place for analyzing the cybersecurity of the software and hardware supply chains with regard to the development and production environment of the AI system. – The supply chain analysis includes pre-trained models, datasets, and software libraries used, as well as the hardware used for the training and, if applicable, production phases of the AI system. – A process should be in place for the continuous monitoring of the supply chain, including a plan for updates and reviews following the discovery of previously undiscovered vulnerabilities in the supply chain.	**Organisational measures – System-related processes** – A process should be in place to analyse the cybersecurity of the software and hardware supply chains with regard to the development and production environment of the AI system. – The supply chain analysis includes pre-trained models, datasets, and software libraries used.	No measures were taken to control the cybersecurity aspects of AI supply chain management.		System/Component	Measure	Normal	

AI-specific cybersecurity

CY1.4	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
Measures and security controls must be in place to prevent misuse and abuse of the AI system.	**Organisational measures – System-related processes** – Processes must be in place to address potential intentional misuse of the AI system during operation, including the integration of possible shutdown mechanisms into risk management. – Processes should be in place to address any misuse detected through monitoring systems and logging. – Processes should specifically ensure that, in addition to the functionality, integrity, and availability of the AI system, the objectives defined under the various quality dimensions are also protected in the event of misuse or abuse. This includes safeguarding personal or proprietary data, protection against discrimination, the degree of autonomy, and the associated human control mechanisms. **Technical measures – system / operation** – Security controls to prevent misuse and abuse should be in place and, if possible, implemented before the production phase. This includes prompt/output filtering methods for generative AI. For safeguards against data access, see also CY1.5. – Inference inputs, queries, and prompts must be recorded as part of the monitoring and logging to enable investigation in the event of misuse, compromise, or abuse. – Red teaming of the entire system regarding misuse or abuse of the AI system should be conducted.	**Organisational measures – System-related processes** – Processes must be in place to address potential intentional misuse of the AI system during operation, including the integration of possible shutdown mechanisms. – Processes should be in place to address any misuse detected through monitoring systems and logging. – Processes should specifically ensure that, in addition to the functionality, integrity, and availability of the AI system, the objectives defined under the various quality dimensions are also protected in the event of misuse or abuse. This includes safeguarding the protection of personal or proprietary data, protection against non-discrimination, the degree of autonomy, and the associated human control mechanisms. **Technical measures – Operation** Provision must be made to record inference inputs, queries and prompts as part of monitoring and logging in order to enable investigation in the event of misuse, compromise or abuse.	**Organisational measures – System-related processes** Processes must be in place to address potential intentional misuse of the AI system during operation, including the integration of possible shutdown mechanisms. – Processes should be in place to address any misuse detected through monitoring systems and logging. – Processes should specifically ensure that, in addition to the functionality, integrity, and availability of the AI system, the objectives defined under the various quality dimensions are also protected in the event of misuse or abuse. This includes safeguarding the protection of personal or proprietary data, protection against non-discrimination, the degree of autonomy, and the associated human control mechanisms.	There are no measures in place to prevent misuse and abuse of the AI system.		System	Measure	Normal	VE2.5 (same indicator)

CY1.5	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
Measures and security controls must be in place to regulate access to training, testing, validation, and all other AI-specific data.	**Organisational measure – Governance** Individual access rights to the data must be defined and documented and consistent with the purpose of the AI system (see also CY1.2). Appropriate access management should include at least the following: a) Granting and modifying (provisioning) access rights based on the principle of least privilege and the need-to-know principle; b) Separation of duties; c) Denying access to data to unauthorised individuals; d) Regularly reviewing granted permissions; e) Revoking permissions in the event of changes in the employee's employment status or role; f) In the case of personal data, ensuring data subjects have the opportunity to access their data (see also MA1.3); g) Data documentation should include details on access management (see DA1.1). **Technical measure – data** Technical measures such as an access management system must be defined and in place to prevent unauthorised access to personal and proprietary training, testing, validation, and other AI-specific data. – The transmission of personal or proprietary data between parties or to a cloud must be secured, at least by: a) using and documenting encryption methods for transmission (data in transit); b) implementing technical safeguards for communication security. **Technical measure – Operation** – Technical measures must be defined and implemented to prevent unauthorised access to personal and proprietary training, testing, validation, and other AI-specific data during the operation of the AI system, including input and output data of the AI system.	**Organisational measure – Governance** Individual access rights to the data must be defined and documented and consistent with the purpose of the AI system (see also CY1.2). Appropriate access management should include at least the following: a) Granting and modifying (provisioning) access rights based on the principle of least privilege and the need-to-know principle; b) Separation of duties; c) Denying access to data to unauthorised individuals; d) Regularly reviewing granted permissions; e) Revoking permissions in the event of changes in the employee's employment status or role; f) In the case of personal data, ensuring data subjects have the opportunity to access their data (see also MA1.3); g) Data documentation should include details on access management (see DA1.1). **Technical measure – data** – Technical measures such as an access management system must be defined and in place to prevent unauthorised access to personal and proprietary training, testing, validation, and other AI-specific data. – The transmission of personal or proprietary data between parties or to a cloud must be secured, at least by: a) using and documenting encryption methods for transmission (data in transit); b) implementing technical safeguards for communication security.	**Organisational measure – Governance** – Individual access rights to the data must be defined and documented and must be consistent with the purpose of the AI system (see also CY1.2) **Technical measure – data** – Technical measures such as an access management system must be defined and in place to prevent unauthorised access to personal and proprietary training, testing, validation, and other AI-specific data. – The transmission of personal or proprietary data between parties or to a cloud must be secured, at least by: a) using and documenting encryption methods for transmission (data in transit); b) implementing technical safeguards for communication security.	No measures were taken or security controls introduced to regulate access to AI-specific data.		Component	Measure	Maximum value	CY1.2 (CSMS), CY2.3 (attacks on training phase) DA2.1, DA2.2, DA3.2, MA1.3

CY1.6	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
It must be assessed whether the measures taken have reduced the identified risks to an acceptable level and whether the quality of the AI system meets the set targets.	**Evaluation** – Summary review of the effects of implementing the technical and organisational measures – Assessment of the interactions of the measures taken, and consideration of the mitigation of risks in conjunction with the risks of other quality dimensions – Identification and description of the residual risk after implementation of the technical and organisational measures – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	**Evaluation** – Weighing the interactions of the measures taken, and weighing the mitigation of risks in conjunction with the risks of other quality dimensions – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	**Evaluation** – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	No evaluation was conducted.		System	Evaluation	Maximum value	

CY2	Resilience against AI-specific attacks								
CY2.1	**A**	**B**	**C**	**D**	**Additional information**	**Reference level**	**Type**	**Weighting**	**Link**
Risks to the cybersecurity of the AI system from AI-specific attacks must be analysed taking into account the intended use.	**Analysis – Risk** Detailed cybersecurity risk analysis, including quantification of probabilities of occurrence and risks: – Identification of potential risks and their causes – Assignment of risk responsibility – Estimation of the probability of occurrence – Estimation of the probability of detection – Estimation of the impact – Structured grading and prioritisation of risks. The risk assessment should, taking into account the intended use and scope (see TR1.1), consider at least the following: – Analysis of all potential attack scenarios and attack vectors on the AI elements of the system, including AI-specific attacks (e.g., evasion attacks or data poisoning) or classic attacks (e.g., malware embedded in AI software libraries). The risk assessment framework should ideally be embeddable in a cybersecurity management system and should ideally follow established standards.	**Analysis – Risk** Limited risk analysis focusing on protection needs without quantifying probabilities – Identification of potential risks – Allocation of risk responsibility – Estimation of impact – Structured grading and prioritisation of risks. The risk assessment should, taking into account the intended use and scope (see TR1.1), consider at least the following: - analysis of all potential attack scenarios and attack vectors on the AI elements of the system, including AI-specific attacks (e.g., evasion attacks or data poisoning) or traditional attacks (e.g., malware embedded in AI software libraries). The risk assessment framework should ideally be embeddable into a cybersecurity management system and should ideally follow established standards.	**Analysis – Risk** Primarily qualitative assessment without probabilities – Identification of potential risks – Allocation of risk responsibility – Qualitative estimation of impact – Qualitative grading and prioritisation of risks. The risk assessment should, taking into account the intended use and scope (see TR1.1), consider at least the following: – Analysis of all potential attack scenarios and attack vectors on the AI elements of the system, including AI-specific (e.g., evasion attacks or data poisoning) or classic attacks (e.g., malware embedded in AI software libraries). The risk assessment framework should ideally be embeddable in a cybersecurity management system and should ideally follow established standards.	No risk analysis was conducted.	Important standards include, for example, BSI IT Baseline Protection (BSI IT-Grundschutz), ISO 27k series, IEC 62443, ISO/SAE 21434, ETSI EN 303 645, NIST AI RMF, …	System	Analysis	Maximum value	TR1.1 (purpose)

AI-specific cybersecurity

CY2.2	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
Measures and security controls must be in place to mitigate AI-specific attacks that may occur during operation.	**Technical measures – Testing** (metrics, AI models, attacks in the inference and deployment phases): Examination of the scope-based selection and number of metrics and tests of adversarial robustness with respect to attack scenarios during the inference and deployment phases that may target the confidentiality, integrity, or availability of the AI system. This includes testing with basic and advanced methods for at least the following attack types: a) evasion attacks, b) latency attacks, c) model extraction attacks, d) data reconstruction or property inference attacks, e) prompt injection attacks or jailbreaks. – Penetration tests of system aspects related to AI-specific attacks during operation should be performed. The selected metrics and tests should have at least the following characteristics: – Methods should be applied to each AI model within and for the AI system as a whole, if it generates an aggregated output. – If justified by the selection, the metrics and tests should include both important basic methods and advanced methods (regarding complexity, information content, implementation effort, e.g., comprehensive testing tools). – The selected metrics and tests should allow for a high degree of automation where possible. – If applicable, thresholds should be defined (as minimum requirements for functionality/performance) and the thresholds justified, taking into account the scope and intended use. – If applicable, the metrics and thresholds should be differentiated according to different use cases, with reference to the definition of the scope. **Technical measures – model** – Appropriate AI-specific security controls and technical measures to mitigate the identified risk of attacks during the inference phase and deployment of the AI system must be implemented at the model level (e.g., adversarial training, methods to strengthen model robustness, or uncertainty methods). – Robustness proofs should be provided for each AI model against various relevant attack types. **Technical measures – system** – Appropriate AI-specific or non-AI-specific security controls to mitigate the identified risk of attacks during the inference phase and deployment of the AI system must be implemented at the system level (e.g., monitoring of inputs, see MA2.3). – Classic security controls for the entire system during operation should be anticipated and planned, including integration into an existing CSMS. – Red teaming of the entire system with regard to AI-specific attacks during the operational phase should be carried out.	**Technical measures – Testing** (metrics, AI models, attacks in the inference and deployment phases): Examination of the scope-based selection and number of metrics and tests of adversarial robustness with respect to attack scenarios during the inference and deployment phases that may target the confidentiality, integrity, or availability of the AI system. This includes testing with basic methods for at least the following attack types: a) evasion attacks, b) latency attacks, c) model extraction attacks, d) data reconstruction or property inference attacks, e) prompt injection attacks or jailbreaks. Penetration tests of the entire system with respect to AI-specific attacks during operation should be performed. The selected metrics and tests should have at least the following characteristics: – Methods should be applied to each AI model within and for the AI system as a whole, if it generates an aggregated output. – If justified by the selection, the metrics and tests should include both important basic methods and advanced methods (regarding complexity, information content, implementation effort, e.g., comprehensive testing tools). – The selected metrics and tests should allow for a high degree of automation where possible. – If applicable, thresholds should be defined (as minimum requirements for functionality/performance) and the thresholds justified, taking into account the scope and intended use. -If applicable, the metrics and thresholds should be differentiated according to different use cases,with reference to the definition of the scope. **Technical measures – model** – Appropriate AI-specific security controls to mitigate the identified risk of attacks during the inference phase and deployment of the AI system must be implemented at the model level (e.g., adversarial training, methods to strengthen model robustness, or uncertainty methods). – Robustness proofs should be provided for each AI model against various relevant attack types. **Technical measures – system** – Appropriate AI-specific or non-AI-specific security controls to mitigate the identified risk of attacks during the inference phase and deployment of the AI system must be implemented at the system level. (e.g., monitoring of inputs, see MA2.3) – Classic security controls of the entire system during operation should be anticipated and planned, including integration into an existing CSMS.	**Technical measures – Testing** (metrics, AI models, attacks in the inference and deployment phases) Examination of the application-scope-justified selection and number of metrics and tests of adversarial robustness with respect to attack scenarios in the inference and deployment phases that may target the confidentiality, integrity, or availability of the AI system. This includes testing with basic methods, ideally for the following attack types: a) evasion attacks, b) latency attacks, c) model extraction attacks, d) data reconstruction or property inference attacks, e) prompt injection attacks or jailbreaks. The selected metrics and tests should ideally have the following characteristics: – Methods can be applied either to each AI model or to the AI system as a whole, depending on what is more sensible and feasible; – Basic methods are sufficient (regarding complexity, information content, implementation effort, e.g., a simple metric); – If necessary, definition of thresholds (as minimum requirements for functionality/performance) and justification of the thresholds, taking into account the application scope and intended use. **Technical measures – model** – Appropriate AI-specific security controls to mitigate the identified risk of attacks during the inference phase and during the deployment of the AI system must be implemented at the model level (e.g., adversarial training, methods to strengthen model robustness, or uncertainty methods). – Robustness proofs should be provided for each AI model against various relevant attack types. **Technical measures – system** – Appropriate AI-specific or non-AI-specific security controls to mitigate the identified risk of attacks during the inference phase and deployment of the AI system must be implemented at the system level. (e.g., monitoring of inputs, see MA2.3) – Classic security controls of the entire system during operation should be anticipated and planned, including integration into an existing CSMS.	No measures were taken to mitigate AI-specific attacks during operation.	Model-level security controls and measures can include, for example, measures in the model architecture such as model hardening through distillation, training measures such as adversarial training, or uncertainty methods for detecting adversarial inputs. System-level security controls can include, for example, access management, usage restrictions on the number of system calls, out-of-distribution detection, and prompt filtering. For details on the taxonomy of different attack types and mitigations, see, for example, MITRE ATLAS or NIST AI 100-2 E2023 on "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations."	System/Component	Measure	Maximum value	CY1.1 (risk), CY1.2 (CSMS), MA2.3 (monitoring), VE1.3 (tests and measures for general robustness), VE2.2 (measures to mitigate system errors and failures), TR1.4, TR1.3

CY2.3	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
Measures and security controls must be in place to mitigate AI-specific attacks that occur during the preparation and model training phase of the AI models.	**Technical measures – Testing** (metrics, AI models, attack scenarios, preparation and model training phase) Examination of the scope-based selection and number of metrics and tests of adversarial robustness with respect to attack scenarios from the preparation and model training phases that may target the confidentiality, integrity, or availability of the AI system. This includes testing with basic and advanced methods of at least the following attack types: a) data poisoning, b) label poisoning, c) AI-specific backdoor attacks. – Penetration tests of system aspects related to AI-specific attacks during the preparation and model training phases should be performed. The selected metrics and tests should have at least the following characteristics: – Methods should be applied to each AI model within and for the AI system as a whole, if it generates an aggregated output. – If justified by the selection, the metrics and tests should include both important basic methods and advanced methods (regarding complexity, information content, implementation effort, e.g., comprehensive testing tools). – The selected metrics and tests should allow for a high degree of automation where possible. – If applicable, thresholds should be defined (as minimum requirements for functionality/performance) and the thresholds justified, taking into account the scope and intended use. – If applicable, the metrics and thresholds should be differentiated according to different use cases, with reference to the definition of the scope. **Technical measures – model** – Appropriate AI-specific security controls to mitigate the identified risk of attacks during the preparation and model training phase of the AI system must be implemented at the model level (e.g., data filtering methods or backdoor search systems, uncertainty methods). – Robustness proofs should be provided for each AI model against various relevant attack types. **Technical measures – system** – Appropriate AI-specific or non-AI-specific security controls to mitigate the identified risk of attacks during the preparation and model training phase of the AI system must be implemented at the system level. – Traditional security controls for the entire system during operation should be anticipated and planned, including integration into an existing CSMS. – Red teaming of the entire system with regard to AI-specific attacks should be carried out during the preparation and model training phase.	**Technical measures – Testing** (metrics, AI models, attack scenarios, preparation and model training phase) Examination of the scope-based selection and number of metrics and tests of adversarial robustness with respect to attack scenarios from the preparation and model training phases that may target the confidentiality, integrity, or availability of the AI system. This includes testing with basic methods for at least the following attack types: a) data poisoning, b) label poisoning, c) AI-specific backdoor attacks. The selected metrics and tests should have at least the following characteristics: – Methods should be applied to each AI model within the system and to the AI system as a whole if it generates aggregated output. – If justified by the selection, the metrics and tests should include both important basic methods and advanced methods (in terms of complexity, information content, implementation effort, e.g., comprehensive testing tools). – if necessary, setting threshold values (as minimum requirements for functionality/performance) and justifying the threshold values, taking into account the scope and intended use. **Technical measures – model** – Appropriate AI-specific security controls to mitigate the identified risk of attacks during the preparation and model training phase of the AI system must be implemented at the model level (e.g., data filtering methods or backdoor search systems, uncertainty methods). – Robustness proofs should be provided for each AI model against various relevant attack types. **Technical measures – system** – Appropriate AI-specific or non-AI-specific security controls to mitigate the identified risk of attacks during the preparation and model training phase of the AI system must be implemented at the system level.	**Technical measures – Testing** (metrics, AI models, attack scenarios, preparation and model training phase): Examination of the application-scope-justified selection and number of metrics and tests of adversarial robustness with respect to attack scenarios from the preparation and model training phases that may target the confidentiality, integrity, or availability of the AI system. This includes testing with basic methods, ideally for the following attack types: a) data poisoning b) label poisoning c) AI-specific backdoor attacks. The selected metrics and tests should ideally have the following characteristics: – Methods can be applied either to each AI model or to the AI system as a whole, depending on what is more sensible and feasible. – Basic methods are sufficient (regarding complexity, information content, implementation effort, e.g., a simple metric). – If necessary, definition of thresholds (as minimum requirements for functionality/performance) and justification of the thresholds, taking into account the application scope and intended use. **Technical measures – model** – Appropriate AI-specific security controls to mitigate the identified risk of attacks during the preparation and model training phase of the AI system must be implemented at the model level (e.g., data filtering methods or backdoor search systems, uncertainty methods). – Robustness proofs should be provided for each AI model against various relevant attack types. **Technical measures – system** – Appropriate AI-specific or non-AI-specific security controls to mitigate the identified risk of attacks during the preparation and model training phase of the AI system must be implemented at the system level.	No measures were taken to mitigate AI-specific attacks during the preparation and model training phase.	Model-level/data-level security controls can include, for example, mitigations such as filtering for data poisoning or model backdoor detection systems. System-level security controls can include, for example, access management, data security management, and out-of-distribution detection. For details on the taxonomy of different attack types and mitigations, see, for example, MITRE ATLAS or NIST AI 100-2 E2023 on "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations."	System/Component	Measure	Maximum value	CY1.1 (risk), DA1.1 (risk), CY1.2 (CSMS), CY1.8 (data access), VE1.3 (tests and measures for general robustness), VE2.2 (measures to mitigate system errors and failures), TR1.4, TR1.3

CY2.4	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
It must be assessed whether the measures taken have reduced the identified risks to an acceptable level and whether the quality of the AI system meets the set targets.	**Evaluation** – Summary review of the effects of implementing the technical and organisational measures – Assessment of the interactions of the measures taken, and consideration of the mitigation of risks in conjunction with the risks of other quality dimensions – Identification and description of the residual risk after implementation of the technical and organisational measures – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	**Evaluation** – Weighing the interactions of the measures taken, and weighing the mitigation of risks in conjunction with the risks of other quality dimensions – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	**Evaluation** – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	No evaluation was conducted.		System	Evaluation	Maximum value	

Reliability

VE1 Performance and robustness

VE1.1	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
Risks that could lead to insufficient performance and robustness of AI components of the AI system must be analysed taking into account the intended use.	**Analysis – Risk** Detailed risk analysis including quantification of probabilities of occurrence and risks: – Identification of potential risks and their causes – Assignment of risk responsibility – Estimation of the probability of occurrence – Estimation of the probability of detection – Estimation of the impact – Structured grading and prioritisation of risks. At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Faulty model training, incorrect model selection, or unsuitable optimisation and validation strategies – Poorly chosen metrics and tests – Limitations in model training, e.g., due to insufficient or inadequate training, validation, or test data, overfitting (direct/indirect analysis) – Hardware limitations in training and inference – Impact of changes to the composition of the AI system (software and hardware levels), in particular risks from retraining or online learning	**Analysis – Risk** Limited risk analysis focusing on protection requirements without quantifying probabilities – Identification of potential risks – Allocation of risk responsibility – Estimation of impact – Structured grading and prioritisation of risks. At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Faulty model training, incorrect model selection, or unsuitable optimisation and validation strategies – Poorly chosen metrics and tests – Limitations in model training, e.g., due to insufficient or inadequate training, validation, or test data, overfitting – Hardware limitations in training and inference – Need for a well-calibrated uncertainty estimation – Impact of changes to the AI system's composition (software and hardware levels), particularly risks from retraining or online learning	**Analysis – Risk** Primarily a qualitative assessment of hazards without probabilities – Identification of potential hazards: – Assignment of risk responsibility – Qualitative assessment of impact – Qualitative grading and prioritisation of hazards. At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Faulty model training, incorrect model choice, or unsuitable optimisation and validation strategies – Poorly chosen metrics and tests – Limits in model training, e.g., due to insufficient or inadequate training, validation, or test data, overfitting – Hardware limitations in training and inference – Need for a well-calibrated uncertainty assessment – Impact of changes to the AI system's composition (software and hardware levels), particularly risks from retraining or online learning	No risk analysis or hazard assessment was carried out.	When using models from external providers, provide documentation and clear justifications for validation. Unlike VE2.1, this is not about the risks arising from an AI system with insufficient performance, but about the risks that lead to insufficient performance itself and can thus jeopardize the intended use. Risk = Hazard x Probability of occurrence.	Component	Analysis	Maximum value	TR1.1, TR1.3

VE1.2	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
Suitable metrics and tests must be defined to assess whether the performance of the AI system achieves its intended functionality.	**Analysis – Metrics / Thresholds [Performance]** The selection and number of metrics and tests are determined by the application area, with a focus on the usefulness of different categories. This includes: – statistical evaluation of the AI model or empirical evaluation of the system functionality across the entire AI system, e.g., through user experiments or surveys; – appropriate methods for determining uncertainty, chosen according to the AI model, or the use of probabilistic AI model architectures; Metrics for Calibrating Uncertainty Determination (e.g., via confidence levels) The selected metrics and tests should ideally have the following characteristics and parameters: – Methods should be applied to each AI model within and to the AI system as a whole if it generates a collected output. – If justified by the selection, the metrics and tests should include both basic methods (simple metrics) and some advanced methods (regarding information content, validity, implementation effort, e.g., comprehensive testing tools - see additional information). – Definition of thresholds (as minimum requirements for functionality/performance) and justification of the thresholds, taking into account the scope and purpose (if applicable: differentiation of metrics and thresholds according to different use cases with reference to the definition of the scope). The justification of the metrics, tests, and methods addresses the following aspects: – Scope, purpose of the AI system, – Model type, – Composition of the AI system (i.e., interaction of the components or relationship between the ML model and the overall system), scope of the AI systems (e.g., classification vs. regression)	**Analysis – Metrics / Thresholds [Performance]** The selection and number of metrics and tests are determined by the application area, with a focus on the usefulness of different categories. This includes: – statistical evaluation of the AI model or empirical evaluation of the system functionality across the entire AI system, e.g., through user experiments or surveys; – appropriate methods for determining uncertainty, chosen according to the AI model, or the use of probabilistic AI model architectures; Metrics for Calibrating Uncertainty Determination (e.g., via confidence levels) The selected metrics and tests should ideally have the following characteristics and parameters: – Methods should be applied to each AI model within and for the AI system as a whole, if it generates a collected output. – If justified by the selection, the metrics and tests should include both basic methods (simple metrics) and some advanced methods (regarding information content, validity, implementation effort, e.g., comprehensive testing tools – see additional information). – Definition of threshold values (as minimum requirements for functionality/performance) and justification of the thresholds, taking into account the scope and purpose. The justification of the metrics, tests, and methods addresses the following aspects: – Scope, purpose of the AI system, – Model type, – Composition of the AI system (i.e., interaction of the components or relationship between the ML model and the overall system), task area of the AI system (e.g., classification vs. regression).	**Analysis – Metrics / Thresholds [Performance]** Selection and number of metrics and tests based on the application area. Depending on the AI model, this includes at least one of the following methods: – statistical evaluation of the AI model or empirical evaluation of the system functionality across the entire AI system, e.g., through user experiments or surveys. The selected metrics and tests should ideally have the following characteristics and parameters: – Methods can be applied either to each AI model or to the AI system as a whole, depending on what is more practical and feasible. – Basic methods are sufficient (regarding information content, validity, and implementation effort, e.g., a comprehensive testing tool – see additional information). – Definition of threshold values (as minimum requirements for functionality/performance) and justification of these threshold values, taking into account the application area and intended use. The justification for the metrics, tests, and methods addresses at least one of the following aspects: – Application area, purpose of the AI system, – Model type, – Composition of the AI system (i.e., interaction of the components or relationship between the ML model and the overall system), task area of the AI system (e.g., classification vs. regression).	No metrics or tests were systematically defined to examine the system's performance.	When using models from external providers, provide documentation and clear justifications for validation. The intended functionality should be evident from the purpose and application area. Distinguish between basic and advanced methods: – The "Technical Test Method Collection.xlsx" lists some common methods and categorizes them as basic and advanced. – Self-developed test methods are recognized as advanced methods. The selected metrics and tests should allow for a high degree of automation where possible.	Component	Analysis	Maximum value	TR1.3, TR1.4

VE1.3	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
Suitable metrics and tests must be defined to assess whether the robustness of the AI system achieves the intended functionality.	**Analysis – Metrics / Thresholds [Robustness]** The application area dictates the selection and number of metrics, tests, and methods for analyzing the system's robustness, depending on the usefulness of as many different categories as possible. This includes: – statistical evaluation of the AI model's robustness, e.g., to determine edge cases, or empirical evaluation of system functionality, e.g., in the form of experiments under extreme conditions; depending on the AI model, appropriate methods for determining uncertainty or the use of probabilistic AI model architectures; Metrics for calibrating uncertainty determination, e.g., for evaluating outlier effects – detection of erroneous inputs or malfunctions at the model level. The selected metrics and tests should have at least the following characteristics: – Methods should be applied to each AI model within and to the AI system as a whole if it generates a collected output. – If justified by the selection, the metrics and tests should include both basic methods (simple metric) and some advanced methods (regarding information content, validity, implementation effort, e.g., comprehensive testing tools – see additional information). – Establishment of thresholds (as minimum robustness requirements) and justification of the thresholds, taking into account the scope and purpose (if applicable: differentiation of metrics and thresholds according to different use cases with reference to the definition of the scope). The justification of the metrics, tests, and methods addresses the following aspects: – Scope, purpose of the AI system, – Model type, – Composition of the AI system (i.e., interaction of the components or relationship between them). ML model and overall system), scope of the AI system (e.g. classification vs. regression)	**Analysis – Metrics / Thresholds [Robustness]** The selection and number of metrics and tests are determined by the application area, with a focus on the usefulness of different categories. This includes: – statistical evaluation of the AI model or empirical evaluation of the system functionality across the entire AI system, e.g., through user experiments or surveys; – appropriate methods for determining uncertainty, chosen according to the AI model, or the use of probabilistic AI model architectures; Metrics for calibrating uncertainty determination, e.g., for evaluating outlier effects – detection of erroneous inputs or malfunctions at the model level. The selected metrics and tests should ideally have the following characteristics: – Methods should be applied to each AI model within and to the AI system as a whole if it generates a collected output. – If justified by the selection, the metrics and tests should include both basic methods (simple metrics) and some advanced methods (regarding information content, validity, implementation effort, e.g., comprehensive testing tools – s ee additional information). – Definition of thresholds (as minimum requirements for functionality/performance) and justification of the thresholds, taking into account the application area and intended use. The justification of the metrics, tests, and methods addresses the following aspects: – Application area, purpose of the AI system, – Model type, – Composition of the AI system (i.e., interaction of the components or relationship between the ML model and the overall system), task area of the AI system (e.g., classification vs. regression).	**Analysis – Metrics / Thresholds [Performance]** The selection and number of metrics and tests are determined by the application area. Depending on the AI model, this includes at least one of the following methods: – statistical evaluation of the AI model or empirical evaluation of the system functionality across the entire AI system, e.g., through user experiments or surveys; – detection of erroneous inputs or malfunctions at the model level. The selected metrics and tests should ideally have the following characteristics: – Methods can be applied either to each AI model or to the AI system as a whole, depending on what is more sensible and feasible. – Basic methods are sufficient (regarding information content, validity, implementation effort, e.g., comprehensive testing tools – see additional information). – Definition of thresholds (as minimum requirements for functionality/performance) and justification of the thresholds, taking into account the scope and intended use. The rationale for the metrics, tests, and methods addresses the following aspects: – Application area, purpose of the AI system, – Model type, – Composition of the AI system (i.e., interaction of the components or relationship between the ML model and the overall system), task area of the AI system (e.g., classification vs. regression)	No metrics or tests were systematically defined to examine the robustness of the system.	The intended functionality should be derived from the intended use and application area. The rationale for the metrics addresses, for example, the following aspects: – Application area, purpose of the AI system, – Model type, – Composition of the AI system (i.e., interaction of the components or relationship between the ML model and the overall system), – ML task (Classification/Regression/Generation/ Unsupervised (e.g., Clustering, Anomaly Detection…)/Reinforcement Learning, etc.). For statistical evaluation: if available, on established benchmark datasets, e.g., with appropriate augmentations to cover robustness tests. Distinction between basic and advanced methods: – The "Technical Testing Methods Collection.xlsx" lists some common methods and categorizes them as basic and advanced. – Self-developed test methods are recognized as advanced methods. The selected metrics and tests should, if possible, allow for a high degree of automation.	Component	Analysis	Normal	TR1.3, TR1.4, CY2.2, CY2.3 (tests and mitigations of robustness against attacks)

Reliability

VE1.4	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
A test plan must be developed and implemented that includes checking all planned metrics and tests, including testing the AI system under representative conditions of the application domain.	**organisational measures – System-related processes** Documented test plan for performance and robustness, defining at least the following points (e.g., in tabular form): – Test objects must be defined (i.e., model to be tested (with version number) or other component/algorithm to be tested, such as an uncertainty estimator for evaluating outlier effects) – Proposed test methods must include at least the metrics and tests previously defined in VE1.2 – VE1.3 and must, as far as possible, cover performance, robustness, and uncertainty estimation. Depending on the previously determined analysis, advanced methods may need to be included in the test plan (for example, regarding implementation effort and hardware requirements), and the characteristics of the test methods must be clearly and comprehensively defined with regard to complexity and information content. Additional specifications for the test parameters may need to be defined. – The test data used must be described. – The test environment must allow testing of the AI system under representative conditions of the application domain, closely resembling the later production environment. – The timing and frequency of the tests must be defined, including the preparation of test plans for a future operational phase of the AI system to continuously test its performance and robustness, particularly with regard to changes in system composition or changing training data (see VE1.6). – The impact of potentially significant changes to the AI system on robustness tests must be considered. – The test resources required for execution (software and hardware requirements) must be determined and documented. – The person responsible for conducting and documenting the tests must be designated. The test plan must be justified with reference to the application context (including, if applicable, an ODD) and to VE1.2-VE1.3 with an argument that the test plan covers all important aspects of performance and robustness (regarding all necessary, representative scenarios).	**Organisational measures – System-related processes** Documented test plan for performance and robustness, defining at least the following points (e.g., in tabular form): – Test objects must be defined (i.e., model to be tested (with version number) or other component/algorithm to be tested, such as an uncertainty estimator for evaluating outlier effects) – Proposed test methods must include at least the metrics and tests previously defined in VE1.2-VE1.3 and must, as far as possible, cover performance, robustness, and uncertainty estimation. Depending on the previously determined analysis, advanced methods may need to be included in the test plan (for example, regarding implementation effort and hardware requirements), and the characteristics of the test methods must be clearly and comprehensively defined with regard to complexity and information content. – Additional specifications for the test parameters may need to be defined. – The test data used must be described. – The test environment must allow testing of the AI system under conditions as representative as possible of the application domain, but does not necessarily have to exactly reflect the final production environment. – The timing and frequency of the tests must be defined, including the preparation of test plans for a future operational phase of the AI system, in order to continuously test the performance and robustness of the AI system, especially with regard to changes in the system composition or changing training data (see VE1.6). – The test resources required for execution (software and hardware requirements) must be determined and documented. The person responsible for conducting and documenting the tests must be designated. – The test plan must be justified with reference to the application context (including, if applicable, an ODD) and to VE1.2-VE1.3.	**Organisational measures – System-related processes** Documented test plan for performance and robustness, defining at least the following points (e.g., in tabular form): – Test objects must be defined (i.e., model to be tested (with version number) or other component/algorithm to be tested, such as an uncertainty estimator for evaluating outlier effects) – Proposed test methods must include at least the metrics and tests previously defined in VE1.2-VE1.3 and must, as far as possible, cover performance, robustness, and uncertainty estimation. The characteristics of the test methods must be clearly and comprehensively defined with regard to complexity and information content. – Additional specifications for the test parameters may need to be established. The test data used must be described. – The timing and frequency of the tests must be specified, including the preparation of test plans for a future operational phase of the AI system. This is necessary to continuously test the performance and robustness of the AI system, particularly with regard to changes in system composition or changing training data (see VE1.6). – The test resources required for execution (software and hardware requirements) must be determined and documented. – The person responsible for conducting and documenting the tests must be designated.	No systematic test plans were defined.	The classification of metrics and methods from simple to advanced depends on many details, but is broadly based on complexity and expected information content (e.g., simple metric, benchmark, up to expert-driven validation approaches such as systematic vulnerability scanning, visual exploration, application of XAI methods, etc.). The degree of test automation also plays a role; where possible, auto-matable methods should be preferred.	System/Component	Measure	Normal	TR1.6 (development process documentation)

VE1.5	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
The AI system must be tested according to the test plan with different inputs, conditions, and environments to ensure its performance and robustness.	**Technical measure – tests** – Formal description of the application scope as input for test methods (Input space, application area with distribution, application boundary, description of an ODD if applicable) – Following the test plan (see VE1.4), analysis of the coverage of the application area (or an ODD if applicable) by the available test data – Description of the given format, i.e., the inputs, conditions, and environments on which performance and robustness are tested directly or indirectly – Execution of the test according to the test planand documentation of all test results – documentation of weaknesses, including at least: a) Any unexpected reduction in the actual performance of the AI system due to the system design and in relation to its intended use; b) Limits of the input range (In which situations/with which inputs does the AI system only function to a limited extent or not at all?); c) Shortcuts in models (if none were identified, this must be documented).	**Technical measure – tests** – Formal description of the application scope as input for test methods (Input space, application area with distribution, application limit, possibly description of an ODD) – Execution of the test according to the test plan and documentation of all test results – documentation of weaknesses, including at least a) Any unexpected reduction in the actual performance of the AI system due to the system design and in relation to the intended use b) Limits of the input range (In which situations/with which inputs does the AI system only work to a limited extent or not at all?	**Technical measure – tests** – Formal description of the application scope as input for test methods (Input space, application area with distribution, application limit, description of an ODD if applicable) – Execution of the test according to the test plan and documentation of all test results	No documented tests were performed.	For technical details describing application areas, see, for example, the Fraunhofer AI test catalog, e.g., [VE-R-RE-RI-01], [VE-R-RE-KR-02], [VE-R-RO-RI-01], [VE-R-RO-KR-01], [VE-R-RO-KR-03], and generally the test measures in the chapter on reliability.	System/Component	Measure	Maximum value	TR1.6 (development process documentation)

VE1.6	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
Monitoring the performance and robustness of the AI system must be possible.	MA2.3					System/Component	Measure	Normal	VE1.6

VE1.7	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
It must be assessed whether the measures taken have reduced the identified risks to an acceptable level and whether the quality of the AI system meets the set targets.	**Evaluation** – Summary review of the effects of implementing the technical and organisational measures – Assessment of the interactions of the measures taken, and consideration of the mitigation of risks in conjunction with the risks of other quality dimensions – Identification and description of the residual risk after implementation of the technical and organisational measures – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	**Evaluation** – Weighing the interactions of the measures taken, and weighing the mitigation of risks in conjunction with the risks of other quality dimensions – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	**Evaluation** – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	No evaluation was conducted.		System	Evaluation	Maximum value	

Reliability

VE2 Fallback plans and functional safety

VE2.1	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
Risks for possible consequences of a malfunction or failure and for the functional safety of the AI system must be analysed taking into account the intended use.	**Analysis – Risk** The risk analysis regarding the possible consequences of a (partial) failure of the AI system, as well as for functional safety, should address hazards and harmful effects on the outside world caused by malfunction or failure due to insufficient performance or robustness. The risk analysis should include: Detailed risk analysis including quantification of probabilities of occurrence and risks: – Identification of potential risks and their causes – Assignment of risk responsibility – Estimation of the probability of occurrence – Estimation of the probability of detection – Estimation of the impact – Structured grading and prioritisation of risks. At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Insufficient performance and robustness of the AI components of the AI system for the intended use (see VE1) – Criticality of the intended use and application area with regard to possible damage and consequences due to insufficient performance and robustness – Possible functional reasons for the failure or partial failure of the AI system with their probabilities of occurrence – Faulty (systematic and random) or abusive use of the AI system outside the intended use – System components with direct interfaces to the AI system (e.g., providing input data or using output data of the AI system). Possible consequences and damage that may arise from accident risks should be considered at least for: – Life and physical health of people – fundamental rights – property and things	**Analysis – Risk** The risk analysis regarding the possible consequences of a (partial) failure of the AI system, as well as for functional safety, should address hazards and harmful effects on the outside world caused by malfunction or failure due to insufficient performance or robustness. The risk analysis should include: Limited risk analysis focusing on protection requirements without quantifying probabilities – Identification of potential risks – Assignment of risk responsibility – Impact assessment – Structured grading and prioritisation of risks. At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Insufficient performance and robustness of the AI components of the AI system for the intended use (see VE1) – Criticality of the intended use and application area with regard to potential damage and consequences due to insufficient performance and robustness Possible functional reasons for the failure or partial failure of the AI system with their probabilities of occurrence – Faulty (systematic and random) or abusive use of the AI system outside its intended use – System components with direct interfaces to the AI system (e.g., providing input data or using output data from the AI system). Potential consequences and damage that may arise from accident risks should be considered at least for: – Human life and limb and physical health – Fundamental rights – property and things	**Analysis – Risk** The risk analysis regarding the possible consequences of a (partial) failure of the AI system, as well as for functional safety, should address hazards and harmful effects on the outside world caused by malfunction or failure due to insufficient performance or robustness. The risk analysis should include: Primarily a qualitative assessment without probabilities – Identification of potential hazards – Assignment of risk responsibility – Qualitative assessment of the impact – Qualitative grading and prioritisation of hazards. At least the following risk sources must be considered, taking into account the intended use (see TR1.1): – Insufficient performance and robustness of the AI components of the AI system for the intended use (see VE1) – Criticality of the intended use and application area with regard to potential damage and consequences due to insufficient performance and robustness – Possible functional reasons for the failure or partial failure of the AI system with their probabilities of occurrence – Faulty (systematic and random) or abusive use of the AI system outside the intended use – System components with direct interfaces to the AI system (e.g., providing input data or using output data of the AI system). Potential consequences and damage that may arise from accident risks should at least be considered for: – Life and limb and physical health of people – Fundamental rights – Property and Things	No risk analysis was conducted.	Here, the protection requirements analysis can also be referenced. However, it should be noted that the auditor is not expected to conduct a complete risk analysis, but rather to assess the existence and scope of such an analysis. Here, it would be possible to refer to common standards, e.g., for functional safety and safety, such as ISO/IEC Guide 51, and to conduct an FMEA (which, if available, should actually be taken into account here). More details can be found in Chapter 8 of the Fraunhofer IAIS AI Testing Catalog. A classic reference on functional AI safety is Amodei 2006.	System	Analysis	Maximum value	

VE2.2	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
Suitable technical measures must be defined in the operation of the AI system to mitigate the risk of malfunction or failure of the AI system.	**Technical measure – Operation** – Systemic redundancy and fallback mechanisms should be provided in the AI system for operation (e.g., switching to a safe mode, emergency stop switch). – Safe failure mechanisms should be planned and, if possible, already implemented (e.g., tamper protection, safe mode). – Interfaces between the monitoring system provided for in VE1.6 and the fallback and failure systems should be ensured to enable the monitoring of AI components to be linked to that of a larger IT system. – Interfaces for an alarm system should be planned (end users, providers, responsible authorities). – Fail-safe logging of a product in operation must be supported by the AI system (e.g., black box), see also TR1.7. – Intervention measures ("incident response") for troubleshooting and system recovery from operation should be possible.	**Technical measure – Operation** – Systemic redundancy and fallback mechanisms should be provided in the AI system for operation (e.g., switching to a safe mode, emergency stop switch). – Safe failure mechanisms should be planned and, if possible, already implemented (e.g., tamper protection, safe mode). – Interfaces between the monitoring system provided for in VE1.6 and the fallback and failure systems should be ensured in order to connect the monitoring of AI components with that of a larger IT system; see also TR1.7.	**Technical measure – Operation** – Mechanisms for safe failure should be planned and, if possible, already implemented (e.g., tamper protection, safe mode).	No measures were taken to mitigate the risk of malfunctions and failures during operation.		System	Analysis	Normal	MA2.3 (monitoring), TR1.7(logging), CY2.2, CY2.3 (measures to mitigate attacks on the system)

VE2.3	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
Suitable technical measures in the form of metrics and tests must be defined to mitigate the risk of malfunction or failure of the AI system.	**Technical measures – metrics & thresholds** The selection and number of metrics and tests, justified by the application area, should include a variety of categories based on their usefulness. This includes: – Methods that promote/ensure the reliability of the AI components, e.g., a well chosen optimisation method, uncertainty determinations (e.g., confidence levels). – Detection methods and error handling in checks at the data, model, or output levels, including appropriate thresholds. – Detection methods and error handling in checks at the system level, including non – AI-specific measures and appropriate thresholds. The selected metrics and tests should have at least the following characteristics: – If justified by the selection criteria, the metrics and tests should include both basic methods (simple metrics) and some advanced methods (regarding information content, significance, implementation effort, e.g., a comprehensive testing tool – see additional information). – If necessary, the definition of thresholds (as minimum requirements for the functionality/performance of the entire system) and justification of these thresholds, taking into account the application area and intended use.	**Technical measures – metrics & thresholds** The selection and number of metrics and tests are justified by the application area, depending on the usefulness of different categories. This includes: – Methods that promote/ensure the reliability of the AI components, e.g., a well chosen optimisation method, uncertainty determination (e.g., confidence levels); – Detection methods and error handling in checks at the data, model, or output levels, including appropriate thresholds. The selected metrics and tests should ideally have the following characteristics: – If justified by the selection criteria, the metrics and tests should include both basic methods (simple metrics) and some advanced methods (regarding information content, significance, implementation effort, e.g., comprehensive testing tools – see additional information); – If necessary, the definition of thresholds (as minimum requirements for the functionality/performance of the entire system) and justification of these thresholds, taking into account the application area and intended use.	**Technical measures – metrics & thresholds** The selection and number of metrics and tests are justified by the application area. This includes at least one of the following methods: – Methods that promote/ensure the reliability of the AI components, e.g., a well chosen optimisation method, uncertainty determination (e.g., confidence levels) – Detection methods and error handling in checks at the data, model, or output levels, including appropriate thresholds. The selected metrics and tests should ideally have the following characteristics: – Basic methods are sufficient (regarding information content, significance, and implementation effort, e.g., a simple metric – see additional information) – If necessary, the definition of thresholds (as minimum requirements for the functionality/performance of the entire system) and justification of these thresholds, taking into account the application area and intended use.	No technical methods or metrics were implemented to mitigate the risk of malfunctions and failures.	Distinction between basic and advanced methods: – The test methods collection lists several common methods and categorizes them as basic and advanced. – Self-developed test methods are recognized as advanced methods. The selected metrics and tests should allow for a high degree of automation where possible.	System	Analysis	Normal	

VE2.4	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
Suitable organisational measures must be defined to mitigate the risk of malfunction or failure of the AI system.	**Organisational measures – Governance** – Planning of, or ,where possible, integration oforganisational or technical measures to mitigate errors in a management system, particularly with regard to human-caused systematic errors and faulty actions during the life cycle Planning measures to mitigate the second-order impact on stakeholders, including communication. – Establishing an incident response channel for future customers through which further mitigation measures can be taken immediately (without undue delay). – The mitigation measures should be part of a defined strategy that can be integrated into an ongoing governance system for managing security risks.	**Organisational measures – Governance** – Planning of, or, where possible, integration of organisational or technical measures to mitigate errors in a management system, particularly with regard to human – Caused systematic errors or faulty actions during the life cycle – the mitigation measures should be part of a defined strategy that can be integrated into an ongoing governance system for managing security risks.	**Organisational measures – Governance** – Planning of organisational or technical measures to mitigate random errors in a management system, particularly with regard to human-caused systematic errors and faulty actions during the life cycle	No organisational methods were provided to mitigate the risk of malfunctions and failures.		System	Analysis	Normal	

Reliability

VE2.5	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
Within the framework of the intended functionality and taking into account the scope of application, the defined measures must be tested before operation.	**Technical measure – test** The mitigation measures and strategies defined in VE2.2-VE2.4 must be tested and evaluated as far as possible, taking into account the scope and intended use. This includes: – Testing and evaluating monitoring and logging (see also MA2.3 und TR1.7) with regard to the risks identified in VE2.1, particularly concerning erroneous inputs, incorrect application environments, external changes, and extreme cases; – Testing fallback mechanisms, fail – safe modes, and alarm systems defined in VE2.2; at a minimum, the possibility of connecting such systems to the AI system must be tested; – Where applicable, checking metrics and tests to mitigate the risk of malfunction or failure of the AI system, as defined in VE2.3; – Documenting all test results	**Technical measure – test** The mitigation measures and strategies defined in VE2.2-VE2.4 must be tested and evaluated as far as possible, taking into account the scope and intended use. This includes: – Testing and evaluating monitoring and logging (see also MA2.3 und TR1.7) with regard to the risks identified in VE2.1, particularly concerning erroneous inputs, incorrect application environments, external changes, and extreme cases; – Where applicable, reviewing metrics and tests to mitigate the risk of malfunction or failure of the AI system, as defined in VE2.3; – Documenting all test results	**Technical measure – test** The mitigation measures and strategies defined in VE2.2-VE2.4 must be tested and evaluated as far as possible, taking into account the scope and intended use. This includes: – Evaluating monitoring and logging (see also MA2.3 und TR1.7) and the mitigation measures in VE2.2-VE2.4 with regard to the risks identified in VE2.1, particularly concerning erroneous inputs, incorrect application environments, external changes, and extreme cases – Documenting all test results	No tests of the technical measures were carried out.		System/Component	Measure	Maximum value	

VE2.6	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
Measures and security controls must be in place to prevent misuse or abuse of the AI system.	CY1.4					System/Component	Measure	Normal	

VE2.7	A	B	C	D	Additional information	Reference level	Type	Weighting	Link
It must be assessed whether the measures taken have reduced the identified risks to an acceptable level and whether the quality of the AI system meets the set targets.	**Evaluation** – Summary review of the effects of implementing the technical and organisational measures – Assessment of the interactions of the measures taken, and consideration of the mitigation of risks in conjunction with the risks of other quality dimensions – Identification and description of the residual risk after implementation of the technical and organisational measures – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	**Evaluation** – Weighing the interactions of the measures taken, and weighing the mitigation of risks in conjunction with the risks of other quality dimensions – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	**Evaluation** – Assessment by a qualified and authorised person as to whether the residual risk is tolerable – Justification of the tolerability of the residual risk	No evaluation was conducted.		System	Evaluation	Maximum value	

4.5 Process flow of the overall evaluation

These instructions describe the overall evaluation process in detail in five schematic steps (see Figure 6).

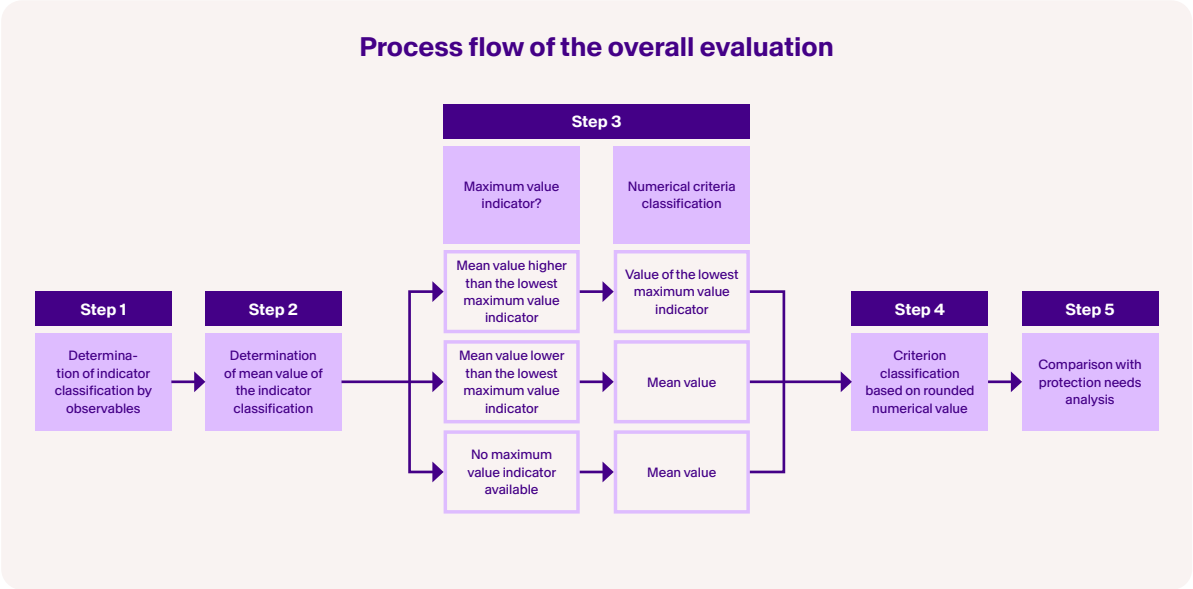


Figure 6: Process flow of the overall evaluation

To arrive at a rating at criteria level, the individual ratings of the respective indicators are summarised using an aggregation function ("rating"). The aggregation is based on an average of the equally weighted indicator ratings, taking into account individual maximum value indicators. Their rating represents a binding maximum value for the average due to their criticality for the determined protection needs. The result is a rating for each criterion, which shows the extent of the quality measures taken in the AI system by scoring from A (best) to C (lowest) or D (non-existent). A detailed description of the aggregation function, the selection and weighting by the maximum value indicators can be found in the following sections.

Step 1: Determination of indicator rating through observables
First, the ratings of the indicators are determined using the observables and categorised into the four ratings from D to A. This is a major process step in terms of content and is described conceptually in Figure 7.

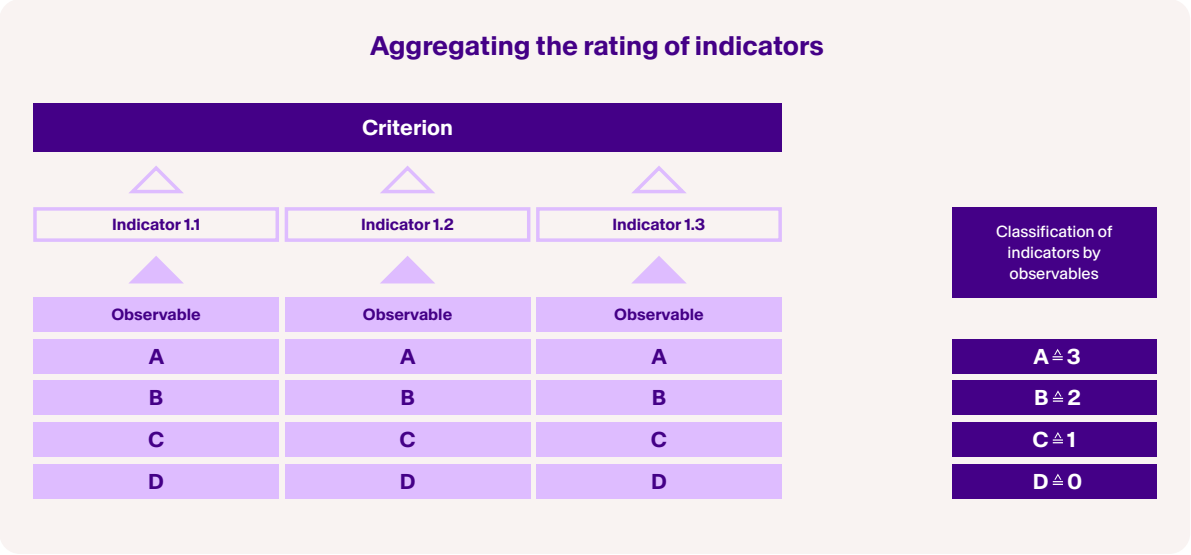


Figure 7: Illustration of the basic approach to aggregating the rating of indicators by observables to the criteria level.

Step 2: Determination of the mean value of the indicator ratings
In the next step, the arithmetic mean of the indicator ratings is determined. The exact formula for the determination can be found in Figure 9. The Figure 8 shows an exemplary case with three indicators.

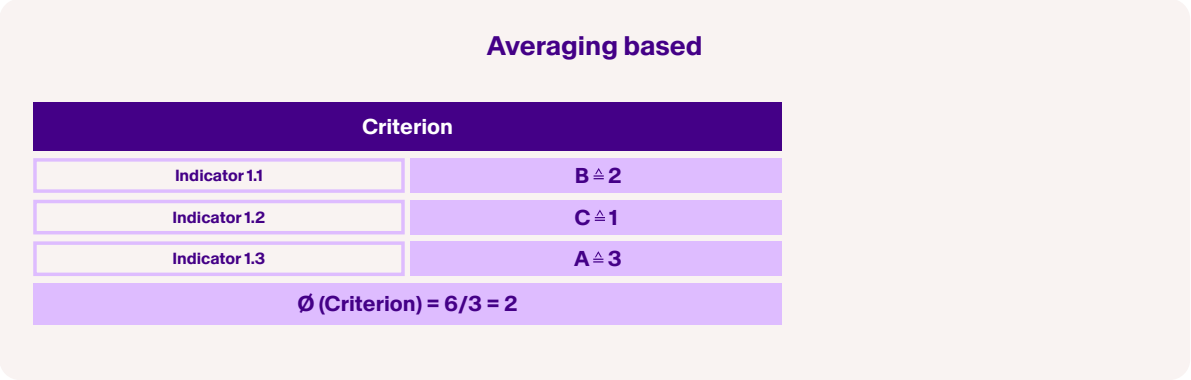


Figure 8: Example illustration of averaging based on the indicator rating

Step 3: Checking the maximum value indicators
Subsequently, the maximum value indicators must be checked and, if necessary, applied according to the descriptions in Figure 10 to change the result of the rating. As Figure 8 shows, there are three different conceptual options:

1. The mean value is greater than the smallest maximum value indicator. This means that the criteria rating is equal to the smallest maximum value indicator.

2. The mean value is less than or equal to the smallest maximum value indicator. In this case, the mean value determines the criteria rating.

3. There is no maximum value indicator. Here too, the average value determines the criteria rating.

Detailed description of the evaluation aggregation

Formal calculation
In detail, the aggregation from indicator to criteria level is carried out by assigning an integer between 3 and 0 to each possible rating in descending order (i.e., A=3, ..., D=0). This allows the calculation of an aggregated rating for each criterion L_Krit by a simple arithmetic mean over all associated indicator ratings L_(Ind, i):

$$L_{Krit} = \frac{\sum_{i=0}^{n} L_{Ind,i}}{n}$$

This is rounded normally, e.g., a "A" and a "B" at indicator level result in an overall rating of "A" (A=3, B=2 L_(Krit)= 5/2= 2,5 → 3). To allow individual indicators to be particularly critical for the protection needs of a criterion, such indicators are categorised as "extreme value indicators" for aggregation in the VCIO. As a result, their individual rating specifies a maximum value for the criteria rating L_Krit i.e. the calculation applies to the rating for maximum value indicator L_(Ind, j)^max:

$$L_{Krit} = \min\left(L_{Ind,j}^{max}, \frac{\sum_{i=0}^{n} L_{Ind,i}}{n}\right)$$

The maximum value indicators were identified in detail according to two principles; they are labelled as such in the assessment requirements:

1 Indicators that are considered particularly critical for the assigned protection categories of the protection needs analysis

2 Indicators that are not well balanced in terms of content by the measures in other indicators, for example, according to a principle of complementarity.

Figure 9: Detailed description of the evaluation aggregation from indicator to criteria level

Step 4: Determination of the criteria rating

Formally, a rating for the entire criterion is determined in the last step according to the results from step 3. See also Figure 10. Steps 1–4 are repeated for all criteria.

Rating of the criteria

Criterion classification based on the rounded numerical value	
Rounded numerical values	3 ≙ A
	2 ≙ B
	1 ≙ C
	0 ≙ D

Figure 10: Rating of the criteria based on the rounded numerical value.

Step 5: Comparison with the protection needs analysis
Once the rating of all criteria has been determined, the comparison with the protection ratings of the protection needs analysis can be carried out as described above (see Figure 11).

Depending on the comparison per criterion, the result is either passing or failing the requirements of the minimum standard.

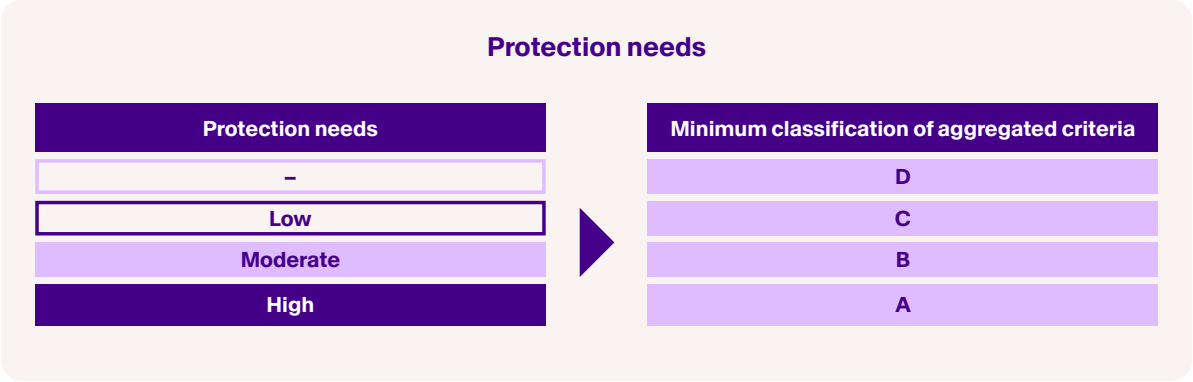


Figure 11: Protection needs as implicit minimum levels in the overall evaluation, which define the assessment depth of the minimum standard for an AI system in a specific use case.

4.6 Test Method Catalogue

Test Method Catalogue

This document is part of the MISSION KI quality standard. ©acatech – National Academy of Science and Engineering. This work is licensed under a Creative Commons Attribution-NoDerivs 4.0 International Licence (CC BY-ND 4.0). https://creativecommons.org/licenses/by-nd/4.0/deed.en

			See the test criteria catalog	Data required for using the metric (input data)	Level of detail of the metric	Required knowledge and access to the model	Verifiable tasks using the metric	Automated verification of metric results against limit values	The method's validity, scope, and information gain	Available frameworks, tools, or Git repositories	Examples of descriptive websites
Test/Method Name	**Brief description**	**Testing tools**	**Indi-cators**	**Data requirements**	**Result complexity**	**Model access**	**Task applicability**	**Automation option**	**Depth of the test method**	**Known implementations**	**Reference**
Accuracy	Measures the accuracy of a (binary) classifier as the ratio of correct predictions to total predictions	Amazon Sagemaker Suite, Azure ML, FairLearn, What If Tool, Fairness Indicators, Citadel RADAR, ML Privacy Meter, AI Qualify, DeepChecks, Zeno, Microsoft Responsible AI Dash-board, LangChain OpenEvals, The HELM Benchmark, GAIA Benchmarks, Robustness Gym, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Classification	Yes	Basic method	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html	https://en.wikipedia.org/wiki/Accuracy_and_precision
Accuracy Equality	Accuracy between groups		ND1.2 ND1.3 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Agnostic	Yes	Basic method		
Adversarial accuracy	Metric for the robustness of decisions	Adversarial-Attacks-PyTorch (Foolbox), AutoAttack, Alpha-Beta-CROWN, AI Qualify, AI4CYBER, IBM Adversarial robustness, Robustness Gym	VE1.3 VE1.4 VE1.5 CY2.2 CY2.3 MA2.3	Extended or new test data with GT and Input	Single/multiple real numbers	Inference access (inputs provided by the examiner)	Agnostic	Yes	Basic method	Empirical: https://github.com/fra31/auto-attack https://github.com/Harry24k/adversarial-attacks-pytorch Formal: https://github.com/Verified-Intelligence/alpha-beta-CROWN	
ALE (Accumulated Local Effects)	Measures the impact of one feature on the predicted outcome, while taking into account the average effect of other features.		TR2.2 TR2.3	existing test data with GT	complex results	Inference access (inputs provided by the examiner)	agnostic	probably	Advanced method		
Alpha Feature Importance	Number of features needed to obtain a goodness-of-fit explanation depending on alpha		TR2.2 TR2.3	Existing test data with GT and input	Single/multiple real numbers	Inference access (inputs provided by the examiner)	Agnostic	Yes	Basic method		
Amount of leaked information	Metric for the amount of information a system releases through attacks	Preamble	CY2.2 CY2.3 MA2.3	All available labeled data	Single/multiple real numbers	Model-independent	Agnostic	no	Basic method		https://dl.acm.org/doi/10.1145/1242572.1242598
Anonymity Set Size	Size of the input set that cannot be distinguished from a single input x based on the model outputs		DA1.2 DA1.3	All available labeled data	Single/multiple real numbers	Inference access	Agnostic	Yes	Basic method		https://link.springer.com/article/10.1007/BF00206326
APPS	Coding benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmark data set	Complex results	Inference access (inputs provided by the examiner)	GenAI	Yes	Basic method		
bAbI	Reasoning Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmark data set	Complex results	Inference access (inputs provided by the examiner)	GenAI	Yes	Basic method	https://github.com/facebookarchive/bAbI-tasks	
BLEU	Metric for the similarity of texts, used to evaluate free-text responses in benchmarks, to be used in combination with benchmarks	Azure Machine Learning, Moonshot, RAGAS, LangChain OpenEvals, Robustness Gym,	VE1.2 VE1.4 VE1.5 MA2.3	Extended or new test data with GT and Input	Single/multiple real numbers	Inference access (inputs provided by the examiner)	GenAI	Yes	Basic method	https://github.com/bangoc123/BLEU	https://dl.acm.org/doi/10.3115/1073083.1073135
BoolQ	Question-Answering Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmark data set	Single/multiple real numbers	Inference access (inputs provided by the examiner)	GenAI	Yes	Basic method	https://github.com/google-research-datasets/boolean-questions	https://arxiv.org/abs/1905.10044
Brier Score	Measures the mean squared difference between the predicted probability and the actual outcome	IBM UQ360, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Agnostic	Yes	Basic method	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.brier_score_loss.html	https://en.wikipedia.org/wiki/Brier_score
Calibration Error	Difference between calculated probabilities and accuracy	Truera, Zeno, The HELM Benchmark, IBM UQ360, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Agnostic	Yes	Basic method	https://torchmetrics.readthedocs.io/en/v0.8.0/classification/calibration_error.html	https://towardsdatascience.com/expected-calibration-error-ece-a-step-by-step-visual-explanation-with-python-code-c3e9aa12937d
CIDEr: Consensus-based Image Description Evaluation	Captures the consensus on the quality of image descriptions generated by genAI		VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	GenAI	Yes	Basic method	https://github.com/vrama91/cider	https://arxiv.org/abs/1411.5726
CLIP Image Quality Assessment	Measuring the visual content of images		VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	GenAI	Yes	Basic method	https://lightning.ai/docs/torchmetrics/stable/multimodal/clip_iqa.html	
CLIP score	Text-image similarity metric, e.g., quality of GenAI images		VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	GenAI	Yes	Basic method	https://lightning.ai/docs/torchmetrics/stable/multimodal/clip_score.html	https://huggingface.co/docs/diffusers/v0.21.0/conceptual/evaluation

Test Method Catalogue

This document is part of the MISSION KI quality standard. ©acatech – National Academy of Science and Engineering. This work is licensed under a Creative Commons Attribution-NoDerivs 4.0 International Licence (CC BY-ND 4.0). https://creativecommons.org/licenses/by-nd/4.0/deed.en

			See the test criteria catalog	Data required for using the metric (input data)	Level of detail of the metric	Required knowledge and access to the model	Verifiable tasks using the metric	Automated verification of metric results against limit values	The method's validity, scope, and information gain	Available frameworks, tools, or Git repositories	Examples of descriptive websites
Test/Method Name	**Brief description**	**Testing tools**	**Indi-cators**	**Data requirements**	**Result complexity**	**Model access**	**Task applicability**	**Automation option**	**Depth of the test method**	**Known implementations**	**Reference**
CNN/DailyMail	Summary Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmark data set	Complex results	Inference access (inputs provided by the examiner)	GenAI	Yes	Basic method	https://huggingface.co/datasets/abisee/cnn_dailymail	
Cohen Kappa score	A measure of the agreement between the raters	Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Classification	Yes	Basic method	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html	https://en.wikipedia.org/wiki/Cohen%27s_kappa
Combinatorial Testing	To evaluate the coverage of a dataset metric by checking whether all relevant combinations of feature values are ade-quately represented in the dataset.		MA2.3 DA1.2 DA1.3	all available labeled data	complex results	model-independent	agnostic	probably	Advanced method		https://openaccess.thecvf.com/content_CVPRW_2020/html/w20/Gladisch_Leveraging_Combinatorial_Testing_for_Safety-Critical_Computer_Vision_Datasets_CVPRW_2020_paper.html
Cosine Similarity	A measure of the similarity between two non-zero vectors defined in an inner product space	RAGAS, LangChain OpenEvals, Scikit-Learn	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Agnostic	Yes	Basic method	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html	https://en.wikipedia.org/wiki/Cosine_similarity
Counterfactual Explanations	Metric for the impact of changes in input characte-ristics on the prediction by generating alternative scenarios	What If Tool, FairLearn, Microsoft Responsible AI Dashboard	TR2.2 TR2.3	existing test data with GT and input	single/multiple real numbers	Inference access (inputs provided by the examiner)	agnostic	probably	Advanced method		
Coverage Error	Calculates how far we need to go through ranking points to capture all true designations	Revision	ND1.2 ND1.3 MA2.3 DA1.2 DA1.3	Existing test data with GT	Complex results	Inference access (inputs provided by the examiner)	Agnostic	probably	Basic method	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.coverage_error.html	
Data completeness (match rate)	Comparison of the entries of a data set to a reference data set	Citadel RADAR, Citadel Lens, Google ML Test Score, ScrutinAI	DA1.2 DA1.3 VE2.3 VE2.5	All available labeled data	Single/multiple real numbers	Model-independent	Agnostic	Yes	Basic method		
Data consistency	Check if all data from different sources are in the same format and if there are any contradictions between sources.	System Center Data Protection Manager (DPM), Preamble, AI4CYBER, Google ML Test Score, ScrutinAI	VE1.6 VE2.3 VE2.5 CY2.3 MA2.3 DA1.2 DA1.3	All available labeled data	Single/multiple real numbers	Model-independent	Agnostic	Yes	Basic method		
Data Coverage (Match Rate)	A measure of the coverage of all properties deemed relevant by the data.	System Center Data Protection Manager (DPM), AI4CYBER, Google ML Test Score	VE1.4 VE1.5 VE2.3 VE2.5 MA2.3 DA1.2 DA1.3	All available labeled data	Single/multiple real numbers	Model-independent	Agnostic	Yes	Basic method		
Data imputation	Benchmark for data completion		VE1.2 VE1.4 VE1.5 MA2.3	Benchmark data set	Single/multiple real numbers	Fixed test set (predictions only)	GenAI	Yes	Basic method		
Data timeliness	Check if all data is up to date.	System Center Data Protection Manager (DPM), Preamble, Google ML Test Score	VE1.6 VE2.3 VE2.5 CY2.3 MA2.3 DA1.2 DA1.3	All available labeled data	Single/multiple real numbers	Model-independent	Agnostic	Yes	Basic method		
Data Uniqueness	Metric for the occurrence of duplicates in a dataset	Citadel RADAR, Citadel Lens, Google ML Test Score	VE2.3 VE2.5 DA1.2 DA1.3	All available labeled data	Single/multiple real numbers	Model-independent	Agnostic	Yes	Basic method		
Data validity	Check if all data is in the required format.	Citadel RADAR, Citadel Lens, System Center Data Protection Manager (DPM), AI4CYBER, Google ML Test Score, ScrutinAI	VE1.6 VE2.3 VE2.5 CY2.3 MA2.3 DA1.2 DA1.3	All available labeled data	Single/multiple real numbers	Model-independent	Agnostic	Yes	Basic method		
DICE Score	Assess the similarity between a predicted Segmentation mask and the actual Segmentation mask	ScrutinAI	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Segmentation	Yes	Basic method	https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.dice.html	https://en.wikipedia.org/wiki/Dice-S%C3%B8rensen_coefficient

Test Method Catalogue

This document is part of the MISSION KI quality standard. ©acatech – National Academy of Science and Engineering. This work is licensed under a Creative Commons Attribution-NoDerivs 4.0 International Licence (CC BY-ND 4.0). https://creativecommons.org/licenses/by-nd/4.0/deed.en

			See the test criteria catalog	Data required for using the metric (input data)	Level of detail of the metric	Required knowledge and access to the model	Verifiable tasks using the metric	Automated verification of metric results against limit values	The method's validity, scope, and information gain	Available frameworks, tools, or Git repositories	Examples of descriptive websites
Test/Method Name	**Brief description**	**Testing tools**	**Indi-cators**	**Data requirements**	**Result complexity**	**Model access**	**Task applicability**	**Automation option**	**Depth of the test method**	**Known implementations**	**Reference**
Domain integrity	Check if all data lies within a meaningful predefined range.	Preamble	VE1.6 VE2.3 VE2.5 CY2.3 MA2.3 DA1.2 DA1.3	All available labeled data	Single/multiple real numbers	Model-independent	Agnostic	Yes	Basic method		
domino	Search for weak sections of unstructured data that are closer together in the embedding space and where DNNs exhibit low performance.		VE1.3 VE1.4 VE1.5	existing test data with GT	complex results	Fixed test set (predictions only)	agnostic	probably	Advanced method		https://arxiv.org/abs/2203.14960
Dyck	Reasoning Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmark data set	Complex results	Inference access (inputs provided by the examiner)	GenAI	Yes	Basic method		
Entity – and referential Integrity	Check for uniqueness of references in relational databases	Preamble	VE1.6 CY2.3 MA2.3 DA1.2 DA1.3	All available labeled data	Single/multiple real numbers	Model-independent	Agnostic	Yes	Basic method		
Equal Opportunity	Measures the true-positive rate across groups.	Astraea, Revise,Microsoft Responsible AI Dashboard, AIF360	ND1.2 ND1.3 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Agnostic	Yes	Basic method		
Equalized Odds	True positivity rate and false positivity rate between groups	AIF 360, FairLearn, FairTest, FairDM / FairSight, Fairness Indicators, Robuscope, Microsoft Responsible AI Dashboard, AIF360, CheckList	ND1.2 ND1.3 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Agnostic	Yes	Basic method	https://aif360.readthedocs.io/en/stable/modules/generated/aif360.sklearn.postprocessing.CalibratedEqualizedOdds.html	https://en.wikipedia.org/wiki/Equalized_odds
Error rate balance	Error rate between groups	NeMo Guardrails, AI Qualify, AIF360, Robustness Gym	ND1.2 ND1.3 MA2.3 DA1.2 DA1.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Agnostic	Yes	Basic method		
Error ratio	Errors between groups	AIF360	ND1.2 ND1.3 MA2.3 DA1.2 DA1.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Agnostic	Yes	Basic method		
Explainability Ease Score	Metric for the complexity of the input-output relationship	Truera, LIME, SHAP, AI4CYBER, ScrutinAI	TR2.2 TR2.3	existing test data with GT and input	single/multiple real numbers	Inference access (inputs provided by the examiner)	Agnostic	Yes	Advanced method		
F1 score	Harmonious average of Precision and Recall	Amazon Sagemaker Suite, Azure ML, FairLearn, What If Tool, Fairness Indicators, Citadel RADAR, Robuscope, DeepChecks, Zeno, Microsoft Responsible AI Dashboard, The HELM Benchmark, Robustness Gym, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Classification	Yes	Basic method	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html	https://en.wikipedia.org/wiki/F-score
False Negative Rate		Azure Machine Learning, ML Privacy Meter, TruEra, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Classification	Yes	Basic method	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html	https://en.wikipedia.org/wiki/False_positives_and_false_negatives#False_positive_and_false_negative_rates
False Omission Rate			VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Classification	Yes	Basic method	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html	https://en.wikipedia.org/wiki/Positive_and_negative_predictive_values
False Positive Rate		Azure Machine Learning, ML Privacy Meter, TruEra, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Classification	Yes	Basic method	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html	https://en.wikipedia.org/wiki/False_positive_rate
Feature importance spread	Measure of the dispersion of important features for a given output	Truera, LIME, SHAP, ELI5, ScrutinAI	TR2.2 TR2.3	Existing test data with GT and input	Single/multiple real numbers	Inference access (inputs provided by the examiner)	Agnostic	Yes	Basic method		
Feature importance stability	Variance of Feature Importance	LIME, ELI5	TR2.2 TR2.3	Existing test data with GT and input	Single/multiple real numbers	Inference access (inputs provided by the examiner)	Agnostic	Yes	Basic method		
Fraction of toxic output	Measures the proportion of toxic outputs	Citadel Lens, NeMo Guardrails, Llama Guard 3-8B, Guardrails AI, The HELM Benchmark	ND1.2 ND1.3 MA2.3	existing test data with GT	single/multiple real numbers	Fixed test set (predictions only)	GenAI	Yes	Advanced method	https://developers.perspectiveapi.com/s/about-the-api?language=en_US	https://arxiv.org/abs/2106.10328
Fréchet Inception Distance (FID)	Measures the quality of generative image models		VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	GenAI	Yes	Basic method	https://pytorch.org/ignite/generated/ignite.metrics.FID.html	https://en.wikipedia.org/wiki/Fr%C3%A9chet_inception_distance

Test Method Catalogue

This document is part of the MISSION KI quality standard. ©acatech – National Academy of Science and Engineering. This work is licensed under a Creative Commons Attribution-NoDerivs 4.0 International Licence (CC BY-ND 4.0). https://creativecommons.org/licenses/by-nd/4.0/deed.en

			See the test criteria catalog	Data required for using the metric (input data)	Level of detail of the metric	Required knowledge and access to the model	Verifiable tasks using the metric	Automated verification of metric results against limit values	The method's validity, scope, and information gain	Available frameworks, tools, or Git repositories	Examples of descriptive websites
Test/Method Name	**Brief description**	**Testing tools**	**Indi-cators**	**Data requirements**	**Result complexity**	**Model access**	**Task applicability**	**Automation option**	**Depth of the test method**	**Known implementations**	**Reference**
Fuzzy Testing	Fuzzy testing is an approach that uses uncertain or inaccurate input data to test the robustness and per-formance of machine learning models under varying con-ditions.	Astraea	VE1.2 VE1.3 VE1.4 VE1.5 MA2.3	existing test data with GT and input	single/multiple real numbers	full model access	agnostic	Yes	Advanced method		
Gender-based Illicit Proximity Estimate	Embedding-based linguistic discrimination detection		ND1.2 ND1.3 MA2.3	Existing test data with GT	Complex results	Inference access (inputs provided by the examiner)	Agnostic	probably	Basic method	https://github.com/vaibkumr/RAN-Debias	https://arxiv.org/abs/2006.01938
Grad-CAM	Creates heatmaps that highlight the areas of an image that contribute to the predictions of a neural network.	pytorch-grad-cam, Captum, Robuscope	TR2.2 TR2.3	existing test data with GT and input	complex results	Inference access (inputs provided by the examiner)	agnostic	probably	Advanced method		
GSM8K	Mathematics benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmark data set	Complex results	Inference access (inputs provided by the examiner)	GenAI	Yes	Basic method	https://huggingface.co/datasets/openai/gsm8k	
HellaSwag	Question-Answering Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmark data set	Single/multiple real numbers	Inference access (inputs provided by the examiner)	GenAI	Yes	Basic method	https://rowanzellers.com/hellaswag/	
Homogeneity Score	Homogeneity metric of a cluster label given a basic truth value	Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Clustering	Yes	Basic method	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.homogeneity_score.html	
HOTA (Higher Order Tracking Accuracy)	A standardized metric for comparing trackers in terms of the accuracy of detection, mapping, and localization		VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Complex results	Fixed test set (predictions only)	Recognition	Yes	Basic method	https://github.com/JonathonLuiten/TrackEval	https://autonomousvision.github.io/hota-metrics/
HumanEval	Coding benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmark data set	Complex results	Inference access (inputs provided by the examiner)	GenAI	Yes	Basic method		
IMDB	Sentiment analysis benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmark data set	Complex results	Inference access (inputs provided by the examiner)	GenAI	Yes	Basic method	https://huggingface.co/datasets/stanfordnlp/imdb	
Inception Score	Measures the realism of generated images		VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	GenAI	Yes	Basic method	https://lightning.ai/docs/torchmetrics/stable/image/inception_score.html	https://en.wikipedia.org/wiki/Inception_score
Integrated Gradients	Metric/visualization for the relationship between input features and prediction	Captum, Amazon Sagemaker Suite, Azure ML	TR2.2 TR2.3	existing test data with GT and input	complex results	Inference access (inputs provided by the examiner)	agnostic	probably	Advanced method		
IoU	Intersection over Union, a measure of the merging of two sets (e.g., Segmen-tation pixels)	ScrutinAI	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Segmentation	Yes	Basic method	https://lightning.ai/docs/torchmetrics/stable/detection/intersection_over_union.html	
Kolmogorov-Smirnov Test for Drift Distribution	Test whether two data sets correspond to the same distribution	Robuscope	VE1.6 MA2.3	All available labeled data	Single/multiple real numbers	Model-independent	Agnostic	Yes	Basic method		
k-Projection Coverage	Measures how well a generated dataset covers the distribution of a refe-rence dataset across multiple random subspaces.		MA2.3 DA1.2 DA1.3	all available labeled data	single/multiple real numbers	model-independent	agnostic	probably	Advanced method		https://arxiv.org/pdf/1805.04333
Legal Support	Legal Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmark data set	Complex results	Inference access (inputs provided by the examiner)	GenAI	Yes	Basic method		
LIME	Locally interpretable model for a specific output	LIME, Captum, Amazon Sagemaker Suite, Azure ML	TR2.2 TR2.3	existing test data with GT and input	complex results	Inference access (inputs provided by the examiner)	agnostic	no	Advanced method		https://arxiv.org/abs/1602.04938
LORE	Local rule-based explanations with black-box access		TR2.2 TR2.3	existing test data with GT and input	complex results	Inference access (inputs provided by the examiner)	agnostic	no	Advanced method		
LRP	"Explainable AI" method, which aims to interpret the decision-making process of models by assigning impor-tance points to the input features.		TR2.2 TR2.3	existing test data with GT and input	complex results	Inference access (inputs provided by the examiner)	agnostic	probably	Advanced method		https://www.tensorflow.org/tutorials/interpretability/integrated_gradients

Test Method Catalogue

This document is part of the MISSION KI quality standard. ©acatech – National Academy of Science and Engineering. This work is licensed under a Creative Commons Attribution-NoDerivs 4.0 International Licence (CC BY-ND 4.0). https://creativecommons.org/licenses/by-nd/4.0/deed.en

			See the test criteria catalog	Data required for using the metric (input data)	Level of detail of the metric	Required knowledge and access to the model	Verifiable tasks using the metric	Automated verification of metric results against limit values	The method's validity, scope, and information gain	Available frameworks, tools, or Git repositories	Examples of descriptive websites
Test/Method Name	**Brief description**	**Testing tools**	**Indi-cators**	**Data requirements**	**Result complexity**	**Model access**	**Task applicability**	**Automation option**	**Depth of the test method**	**Known implementations**	**Reference**
LSAT	Legal Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmark data set	Complex results	Inference access (inputs provided by the examiner)	GenAI	Yes	Basic method	https://github.com/zhongwanjun/AR-LSAT	
MAE	Mean absolute error	Azure Machine Learning, DeepChecks, Zeno, Microsoft Responsible AI Dashboard, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Regression	Yes	Basic method	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html	https://en.wikipedia.org/wiki/Mean_absolute_error
Mahalanobis Distance	Distance between a sample and a distribution	Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Agnostic	Yes	Basic method	https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.mahalanobis.html	https://en.wikipedia.org/wiki/Mahalanobis_distance
mAP	Average precision	ScrutinAI	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Recognition	Yes	Basic method	https://lightning.ai/docs/torchmetrics/stable/detection/mean_average_precision.html	
MAPE	Percentage absolute error	DeepChecks, Zeno	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Regression	Yes	Basic method	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_percentage_error.html	https://en.wikipedia.org/wiki/Mean_absolute_percentage_error
MATH	Mathematics benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmark data set	Single/multiple real numbers	Inference access (inputs provided by the examiner)	GenAI	Yes	Basic method		
Matthew's correlation coefficient		TruEra	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Classification	Yes	Basic method	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.matthews_corrcoef.html	https://en.wikipedia.org/wiki/Phi_coefficient
Maximum Mean Discrepancy	Metric for the agreement of the distribution of a data set with a referenced data set description	Robuscope	VE1.6 MA2.3 DA1.2 DA1.3	all available labeled data	single/multiple real numbers	model-independent	agnostic	probably	Advanced method		https://jmlr.org/papers/volume13/gretton12a/gretton12a.pdf
Minimum Distortion	Metric for adversarial attacks	AI Qualify, IBM Adversarial robustness, Robustness Gym	VE1.3 VE1.4 VE1.5 CY2.2 CY2.3 MA2.3	Extended or new test data with GT and Input	Single/multiple real numbers	Inference access (inputs provided by the examiner)	Agnostic	Yes	Basic method		
mIoU	IoU values averaged over a dataset.	ScrutinAI	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Segmentation	Yes	Basic method		
MMLU	Question-Answering Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmark data set	Single/multiple real numbers	Inference access (inputs provided by the examiner)	GenAI	Yes	Basic method	https://github.com/hendrycks/test	https://en.wikipedia.org/wiki/MMLU
Monte Carlo Dropout	Method for uncertainty estimation in neural networks		VE1.2 VE1.3 VE1.4 VE1.5 VE2.5 MA2.3	existing test data with GT and input	single/multiple real numbers	full model access	agnostic	Yes	Advanced method		
Mutual information score	A measure of the mutual dependence between the two variables	Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Clustering	Yes	Basic method	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mutual_info_score.html	https://en.wikipedia.org/wiki/Mutual_information
NarrativeQA	Question-Answering Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmark data set	Single/multiple real numbers	Inference access (inputs provided by the examiner)	GenAI	Yes	Basic method	https://huggingface.co/datasets/deepmind/narrativeqa	
NaturalQuestions	Question-Answering Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmark data set	Single/multiple real numbers	Inference access (inputs provided by the examiner)	GenAI	Yes	Basic method	https://github.com/google-research-datasets/natural-questions	
Neuron Coverage	A metric for evaluating the test coverage of neural networks, measuring how many neurons in a network were activated during testing, in order to understand and improve the model's functionality	DeepChecks	VE1.2 VE1.4 VE1.5 MA2.3	existing test data with GT and input	single/multiple real numbers	full model access	agnostic	Yes	Advanced method		
OpenBookQA	Question-Answering Benchmark	Citadel Lens	VE1.2 VE1.4 VE1.5 MA2.3	Benchmark data set	Single/multiple real numbers	Inference access (inputs provided by the examiner)	GenAI	Yes	Basic method	https://huggingface.co/datasets/allenai/openbookqa	https://arxiv.org/abs/1809.02789

Test Method Catalogue

This document is part of the MISSION KI quality standard. ©acatech – National Academy of Science and Engineering. This work is licensed under a Creative Commons Attribution-NoDerivs 4.0 International Licence (CC BY-ND 4.0). https://creativecommons.org/licenses/by-nd/4.0/deed.en

			See the test criteria catalog	Data required for using the metric (input data)	Level of detail of the metric	Required knowledge and access to the model	Verifiable tasks using the metric	Automated verification of metric results against limit values	The method's validity, scope, and information gain	Available frameworks, tools, or Git repositories	Examples of descriptive websites
Test/Method Name	**Brief description**	**Testing tools**	**Indi-cators**	**Data requirements**	**Result complexity**	**Model access**	**Task applicability**	**Automation option**	**Depth of the test method**	**Known implementations**	**Reference**
Out-of-distribution (OOD) generalization	Evaluation of the model's capabilities on unseen data using new synthetic or real-world test data	Astraea, AI Qualify	VE1.2 VE1.4 VE1.5 MA2.3	Extended or new test data with GT and Input	Single/multiple real numbers	Inference access (inputs provided by the examiner)	Agnostic	probably	Basic method		[2403.01874] A Survey on Evaluation of Out-of-Distribution Generalization (arxiv.org)
Page-Hinkley Test	Metric for changes in the mean of a time series, used for drift detection	Robuscope	VE1.6 MA2.3	All available labeled data	Single/multiple real numbers	Model-independent	Agnostic	Yes	Basic method		
Panoptic Quality	A metric that combines Segmentation quality and Recognition quality		VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Segmentation	Yes	Basic method	https://lightning.ai/docs/torchmetrics/stable/detection/panoptic_quality.html	
Perceptual evaluation of Speech Quality (PESQ)	A recognized industry standard for audio quality that takes various characteristics into account.		VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Audio	Yes	Basic method	https://lightning.ai/docs/torchmetrics/stable/audio/perceptual_evaluation_speech_quality.html	https://en.wikipedia.org/wiki/Perceptual_Evaluation_of_Speech_Quality
Population Stability Index	Comparison of the distribution in a variable across two datasets, for drift detection		VE1.6 MA2.3	All available labeled data	Single/multiple real numbers	Model-independent	Agnostic	Yes	Basic method		
Precision	Number of correctly identified positive cases (measured against the total number of positive predictions)	Amazon Sagemaker Suite, Azure ML, FairLearn, What If Tool, Fairness Indicators, Citadel RADAR, Robuscope, DeepChecks, Zeno, Microsoft Responsible AI Dashboard, RAGAS, GAIA bench-marks, Robustness Gym, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Classification	Yes	Basic method	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html	https://en.wikipedia.org/wiki/Precision_and_recall
Predictions Groups Contrast	Difference in features when explaining a subgroup relative to the overall average		TR2.2 TR2.3	existing test data with GT and input	single/multiple real numbers	Inference access (inputs provided by the examiner)	Agnostic	Yes	Advanced method		
Predictive Rate Parity	Measures the accuracy rate of positive predictions across groups between groups	AIF 360, FairLearn, FairTest, FairDM / FairSight, Fairness Indicators, Microsoft Responsible AI Dashboard, AIF360	ND1.2 ND1.3 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Agnostic	Yes	Basic method		
QuAC	Question-Answering Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmark data set	Single/multiple real numbers	Inference access (inputs provided by the examiner)	GenAI	Yes	Basic method	https://quac.ai/	
R2	Measure of the variance in the dependent variable that is explained by an independent variable	Amazon SageMaker Suite, Azure Machine Learning, Deepchecks, Zeno, Microsoft Responsible AI Dashboard, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Regression	Yes	Basic method	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html	https://en.wikipedia.org/wiki/Coefficient_of_determination
RAFT	Text classification		VE1.2 VE1.4 VE1.5 MA2.3	Benchmark data set	Complex results	Inference access (inputs provided by the examiner)	GenAI	Yes	Basic method	https://huggingface.co/datasets/ought/raft	
Recall	Number of correctly identified positive cases (measured against the total number of actual positive cases)	Amazon Sagemaker Suite, Azure ML, FairLearn, What If Tool, Fairness Indicators, Citadel RADAR, Robuscope, DeepChecks, Zeno, Microsoft Responsible AI Dashboard, RAGAS, Robustness Gym, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Classification	Yes	Basic method	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html	https://en.wikipedia.org/wiki/Precision_and_recall
RMSE	Root mean square error	Amazon SageMaker Suite, Azure Machine Learning, Zeno, Microsoft Responsible AI Dashboard, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Regression	Yes	Basic method	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html	https://en.wikipedia.org/wiki/Root_mean_square_deviation
Robustness radius	Metric for the robustness of decisions	Adversarial Attacks PyTorch (Foolbox), Alpha-Beta-CROWN, AI Qualify, AI4CYBER, IBM Adversarial robustness	VE1.3 VE1.4 VE1.5 CY1.5 CY1.6	extended or new test data with GT and Input	single/multiple real numbers	Inference access (inputs provided by the examiner)	Agnostic	Yes	Advanced method		
ROC-AUC	Receiver-operator curve	Amazon Sagemaker Suite, Azure ML, FairLearn, What If Tool, Fairness Indicators, Citadel RADAR, ML Privacy Meter, Robu-scope, DeepChecks, Zeno, Micro-soft Responsible AI Dashboard, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Complex results	Fixed test set (predictions only)	Classification	probably	Basic method	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html	https://en.wikipedia.org/wiki/Receiver_operating_characteristic
ROUGE	Metric for the similarity of texts, used to evaluate free-text responses in benchmarks, to be used in combination with benchmarks	Azure Machine Learning, Moonshot, RAGAS, LangChain OpenEvals, HELM Benchmark, Robustness Gym	VE1.2 VE1.4 VE1.5 MA2.3	Extended or new test data with GT and Input	Single/multiple real numbers	Inference access (inputs provided by the examiner)	GenAI		Basic method		

Test Method Catalogue

This document is part of the MISSION KI quality standard. ©acatech – National Academy of Science and Engineering. This work is licensed under a Creative Commons Attribution-NoDerivs 4.0 International Licence (CC BY-ND 4.0). https://creativecommons.org/licenses/by-nd/4.0/deed.en

			See the test criteria catalog	Data required for using the metric (input data)	Level of detail of the metric	Required knowledge and access to the model	Verifiable tasks using the metric	Automated verification of metric results against limit values	The method's validity, scope, and information gain	Available frameworks, tools, or Git repositories	Examples of descriptive websites
Test/Method Name	**Brief description**	**Testing tools**	**Indi-cators**	**Data requirements**	**Result complexity**	**Model access**	**Task applicability**	**Automation option**	**Depth of the test method**	**Known implementations**	**Reference**
SHAP	Provides consistent and interpretable features to explain model predictions based on Shapley values.	SHAP, Captum, Amazon Sagemaker Suite, Azure ML, Robuscope,Microsoft Responsible AI Dashboard	TR2.2 TR2.3	existing test data with GT	single/multiple real numbers	Inference access (inputs provided by the examiner)	agnostic	probably	Advanced method		
Short-Time Objective Intelligibility (STOI)	Evaluation of speech signals based on an intelligibility measure that correlates highly with the intelligibility of degraded speech signals.		VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Audio	Yes	Basic method	https://lightning.ai/docs/torchmetrics/stable/audio/short_time_objective_intelligibility.html	
Signal-to-Noise Ratio (SNR)	A measure that compares the level of a desired signal with the level of background noise		VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Audio	Yes	Basic method	https://lightning.ai/docs/torchmetrics/stable/audio/signal_noise_ratio.html	https://en.wikipedia.org/wiki/Signal-to-noise_ratio
Silhouette Score	Mean value of the measure that indicates how similar an object is to its own cluster (cohesion) compared to other clusters (separation)	TrojanZoo, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Complex results	Fixed test set (predictions only)	Clustering	no	Basic method	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html	https://en.wikipedia.org/wiki/Silhouette_(clustering)
Sliceline	Search for weak sections of structured data that are semantically coherent and where DNN performs poorly		VE1.3 VE1.4 VE1.5	existing test data with GT	complex results	Fixed test set (predictions only)	agnostic	probably	Advanced method		https://mboehm7.github.io/resources/sigmod2021b_sliceline.pdf
Spearman's rank correlation coefficient	Measures how well two variables, e.g., input and output, can be represented by a monotonic function.	SciPy	TR2.2 TR2.3	Existing test data with GT and input	Single/multiple real numbers	Inference access (inputs provided by the examiner)	Agnostic	Yes	Basic method		
Spotlight	Searching for weak sections in unstructured data using the embedding space of DNNs		VE1.3 VE1.4 VE1.5	existing test data with GT	complex results	Fixed test set (predictions only)	agnostic	probably	Advanced method		https://dl.acm.org/doi/abs/10.1145/3531146.3533240
Statistical Parity	Measures whether different groups have the same probability of being predicted as positive, independent of the actual class.	AIF 360, FairLearn, FairTest, FairDM / FairSight, Fairness Indicators, Robuscope,Microsoft Responsible AI Dashboard, AIF360	ND1.2 ND1.3 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Agnostic	Yes	Basic method	https://aif360.readthedocs.io/en/latest/modules/generated/aif360.sklearn.metrics.statistical_parity_difference.html	https://dataplatform.cloud.ibm.com/docs/content/wsj/model/wos-stat-parity-diff.html?context=cpdaas
Structural similarity index measure (SSIM)	Method for assessing the perceived quality of images and videos (image similarity or generation)		VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	GenAI	Yes	Basic method	https://pytorch.org/ignite/generated/ignite.metrics.SSIM.html	https://en.wikipedia.org/wiki/Structural_similarity_index_measure
Success Rate of Label Poisoning Attacks	Proportion of successful label poisoning attacks	IBM Adversarial robustness	CY2.3 MA2.3	Extended or new test data with GT and Input	Single/multiple real numbers	Inference access (inputs provided by the examiner)	Agnostic	no	Basic method		
Success Rate of Backdoor Attacks	Proportion of successful backdoor attacks	TrojanZoo, IBM Adversarial robustness	CY2.3 MA2.3	Extended or new test data with GT and Input	Single/multiple real numbers	Inference access (inputs provided by the examiner)	Agnostic	no	Basic method		
Success Rate of Data Poisoning Attacks	Proportion of successful data poisoning attacks	auto-attack, DeepChecks, IBM Adversarial robustness	CY2.3 MA2.3	Extended or new test data with GT and Input	Single/multiple real numbers	Inference access (inputs provided by the examiner)	Agnostic	no	Basic method		
Success Rate of Membership Inference Attacks	Percentage of successful membership inference attacks	ml_privacy_meter, DeepChecks, IBM Adversarial robustness	CY2.2 CY2.3 MA2.3	Extended or new test data with GT and Input	Single/multiple real numbers	Inference access (inputs provided by the examiner)	Agnostic	no	Basic method	https://github.com/privacytrustlab/ml_privacy_meter	
Surrogacy Efficacy Score	Proportion of expenses that are easily explained	SHAP, ELI5	TR2.2 TR2.3	Existing test data with GT and input	Single/multiple real numbers	Inference access (inputs provided by the examiner)	Agnostic	Yes	Basic method		
SVM Failure Directions	Searching for weak sections in unstructured data using the CLIP embedding space and an SVM model		VE1.3 VE1.4 VE1.5	existing test data with GT	complex results	Fixed test set (predictions only)	agnostic	probably	Advanced method		https://arxiv.org/abs/2206.14754
Synthetic reasoning	Reasoning Benchmark	Citadel Lens	VE1.2 VE1.4 VE1.5 MA2.3	Benchmark data set	Complex results	Inference access (inputs provided by the examiner)	GenAI	Yes	Basic method		
Systematic weakness search	Searching for weak slices of unstructured data by converting unstructured data into structured data aligned with ODD and applying sliceline to find regions where DNN performs poorly.	ELI5	VE1.3 VE1.4 VE1.5	existing test data with GT	complex results	Fixed test set (predictions only)	agnostic	probably	Advanced method		https://openaccess.thecvf.com/content/CVPR2023W/SAIAD/html/Gannamaneni_Investigating_CLIP_Performance_for_Meta-Data_Generation_in_AD_Datasets_CVPRW_2023_paper.html
Time until Adversary's Success	Metric for adversarial attacks	AI Qualify	VE1.3 VE1.4 VE1.5 CY2.2 CY2.3 MA2.3	Extended or new test data with GT and Input	Single/multiple real numbers	Inference access (inputs provided by the examiner)	Agnostic	no	Basic method		

Test Method Catalogue

This document is part of the MISSION KI quality standard. ©acatech – National Academy of Science and Engineering. This work is licensed under a Creative Commons Attribution-NoDerivs 4.0 International Licence (CC BY-ND 4.0). https://creativecommons.org/licenses/by-nd/4.0/deed.en

			See the test criteria catalog	Data required for using the metric (input data)	Level of detail of the metric	Required knowledge and access to the model	Verifiable tasks using the metric	Automated verification of metric results against limit values	The method's validity, scope, and information gain	Available frameworks, tools, or Git repositories	Examples of descriptive websites
Test/Method Name	**Brief description**	**Testing tools**	**Indi-cators**	**Data requirements**	**Result complexity**	**Model access**	**Task applicability**	**Automation option**	**Depth of the test method**	**Known implementations**	**Reference**
True Negative Rate		Azure Machine Learning, ML Privacy Meter, TruEra, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Classification	Yes	Basic method	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html	https://en.wikipedia.org/wiki/Sensitivity_and_specificity
True Positive Rate		Azure Machine Learning, ML Privacy Meter, TruEra, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Classification	Yes	Basic method	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html	https://en.wikipedia.org/wiki/Sensitivity_and_specificity
TruthfulQA	Question-Answering Benchmark	Citadel Lens	VE1.2 VE1.4 VE1.5 MA2.3	Benchmark data set	Single/multiple real numbers	Inference access (inputs provided by the examiner)	GenAI	Yes	Basic method	https://huggingface.co/datasets/truthfulqa/truthful_qa	
T-SNE	Visualizes high-dimensional data in a low-dimensional representation to identify patterns or clusters.		TR2.2 TR2.3	existing test data with GT and input	complex results	model-independent	agnostic	no	Advanced method		
unc	Uncertainty estimation through model ensemble		VE2.3 VE2.5	Existing test data with GT and input	Single/multiple real numbers	Inference access (inputs provided by the examiner)	Agnostic	Yes	Basic method		
Wasserstein metric	Also called "Earth mover's distance" or optimal transport distance, it is a similarity metric between two probability distributions		VE1.2 VE1.4 VE1.5 MA2.3	Existing test data with GT	Single/multiple real numbers	Fixed test set (predictions only)	Agnostic	Yes	Basic method	https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wasserstein_distance.html	https://en.wikipedia.org/wiki/Wasserstein_metric
WikiFact	Knowledge Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmark data set	Complex results	Inference access (inputs provided by the examiner)	GenAI	Yes	Basic method	https://github.com/google-research-datasets/wikifact	
XSUM	Summary Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmark data set	Complex results	Inference access (inputs provided by the examiner)	GenAI	Yes	Basic method	https://huggingface.co/datasets/EdinburghNLP/xsum	

4.7 Assessment report template

MISSION KI

Assessment report

Results of the self-assessment of
[name of company] "[name of AI system]"
AI system based on the MISSION KI
quality standard

Assessment report

[Name of the company]
“[Name of the AI system]“ AI system

Self-assessment according to MISSION KI quality standard

Date: DD.MM.YYYY
Assessment ID: XXXXXX
Type: Initial assessment/follow-up assessment

MISSION KI
c/o acatech
German National Academy of Science and Engineering
Karolinenplatz 4, 80333 Munich, Germany

mission-ki.de

Gefördert durch:

Bundesministerium
für Digitales und
Staatsmodernisierung

aufgrund eines Beschlusses
des Deutschen Bundestages

Contents

Abbreviation for German:

1 DA = "Datenqualtität, -schutz und Governance"
2 ND = "Nicht-Diskriminierung"
3 TR = "Transparenz"
4 MA = "Menschliche Aufsicht und Kontrolle"
5 VE = "Verlässlichkeit"
6 CY = "KI-spezifische Cybersicherheit"

1. General information on the assessment

1.1. Objectives, scope and responsibility

This report presents the results of a voluntary self-assessment conducted to determine the scope and effectiveness of the quality assurance measures of the (name of AI system) according to the **MISSION KI** quality standard. The evaluation of the AI system is based on a comparison of the identified protection needs in the context of the specific intended purpose and the measures taken, which are supported by evidence. The organisation to be audited is responsible for ensuring that all relevant information is provided completely and truthfully and has been validated in accordance with the requirements.

1.2. Validation

As a general rule, the results are validated internally. Only in cases where it is explicitly stated, has an external validation of the results taken place to increase the reliability. Depending on the rating selected, the following requirements apply to the auditor:

Protection needs	Rating	Assessment approach	Qualification requirement
–	D	No specific requirement	No specific qualification required.
low	C	No requirement. Evidence and ratings can be checked by people who were involved in the development.	(Professional) experience in developing/deploying AI systems or comparable qualification
moderate	B	The persons are not involved in the development of the AI system	(Professional) experience in developing/deploying AI systems or comparable qualification
high	A	The four-eyes principle is applied, with organisationally independent persons	At least one person with auditing experience At least one person with (professional) experience in developing/deploying AI systems or comparable qualification

1.3. Validity of the assessment

The assessment statement is therefore only valid for the clearly defined version of the AI system that was available at the time of the test. It becomes invalid as soon as significant changes to the intended purpose, technical implementation or external conditions occur. As long as there are no such changes and this is ensured by regular checks or monitoring, the assessment statement remains valid.

2. Overview of the AI system and its component(s)

Industry:	**[X]**
Role:	**[Provider/deployer]**
Deployment location:	**[X]**
Contact:	**[X]**
Use case:	1.1 [Title]
Purpose:	1.2 [What the system is used for, in what context]
Task:	1.3 [Which decision/prediction/classification]
Inputs:	1.4 [Data types/situations]
Limits:	1.5 [Non-covered cases/quality limits]
Output/utilisation:	1.6 [Output form]
Users/Affected persons	1.7 [User group/affected persons]
Measure of human controls	1.8. [Participation in the deployment/oversight of the AI system, e.g. HIC, HITL, HOTL]
Deployment	1.9 [Cloud/on-prem/hybrid]
Regulatory requirements	1.10 [Special requirements for the usage context]
Changes on deployment	1.11 [Further training during deployment]

3. Evaluation of quality measures and precautions to safeguard the AI system

The test process for AI systems begins with a detailed description of the use case, followed by a protection needs analysis that determines which quality dimensions and criteria are most relevant and how critical their protection is depending on the intended purpose and context of the system.

Each criterion is first checked for applicability – if a criterion is rated as not applicable, a reason must be given in the *"Comment on protection needs"* field as to why it does not apply; no further evaluation of this criterion is required. For applicable criteria, the protection need is categorised into three levels: low, moderate or high, reflecting the severity of potential damage.

The system is then rated, with each applicable criterion being assessed on the basis of the measures and safeguards implemented. This evaluation is supported by evidence such as technical tests, documentation and certifications. The degree of validation of the evidence implemented must be listed. Measures that are particularly noteworthy in terms of the quality of the AI system can also be listed in this assessment report.

The results of the protection needs analysis determine the minimum level required for each criterion and are systematically compared with the actual rating of the system. The evaluation therefore checks whether the implemented measures fulfil the required protection needs.

Data quality, protection and governance (DA)[1]		
Data quality		
Protection needs:		**Low**
		Always applicable. A low protection need requires at least level C measures.
Level achieved:		**C**
		Evidence was validated internally.
Selected Measures	Organisational measures:	e. g., user instruction, training, specialised and technical documentation, reports
	Technical measures	e. g., test results, logging, monitoring, data processing steps, emergency stop switch
Comment:		
Criterion Fulfilled		

1 DA = abbreviation for German "Datenqualtität, -schutz und Governance"

Protection of personal data		
Protection needs:		**Choose an item.**
		Choose an item.
Level achieved:		**Choose an item.**
		Choose an item.
Selected Measures	Organisational measures:	e. g., user instruction, training, specialised and technical documentation, reports
	Technical measures:	e. g., test results, logging, monitoring, data processing steps, emergency stop button
Comment:		
Criterion Fulfilled		

Protection of proprietary data		
Protection needs:		**Choose an item.**
		Choose an item.
Level achieved:		**Choose an item.**
		Choose an item.
Selected Measures	Organisational measures:	e. g., user instruction, training, specialised and technical documentation, reports
	Technical measures:	e. g., test results, logging, monitoring, data processing steps, emergency stop button
Comment:		
Criterion Fulfilled		

Non-discrimination (ND)[2]		
Avoidance of unjustified discrimination and bias		
Protection needs:		**Choose an item.**
		Choose an item.
Level achieved:		**Choose an item.**
		Choose an item.
Selected Measures	Organisational measures:	e.g., user instruction, training, specialised and technical documentation, reports
	Technical measures:	e.g., test results, logging, monitoring, data processing steps, emergency stop button
Comment:		
Criterion Fulfilled		

2 ND = abbreviation for German “Nicht-Diskriminierung”

Transparency (TR)[3]

Traceability and documentation		
Protection needs:		**Choose an item.**
		Choose an item.
Level achieved:		**Choose an item.**
		Choose an item.
Selected Measures	Organisational measures:	e.g., user instruction, training, specialised and technical documentation, reports
	Technical measures:	e.g., test results, logging, monitoring, data processing steps, emergency stop button
Comment:		
Criterion Fulfilled		

Explainability and interpretability		
Protection needs:		**Choose an item.**
		Choose an item.
Level achieved:		**Choose an item.**
		Choose an item.
Selected Measures	Organisational measures:	e.g., user instruction, training, specialised and technical documentation, reports
	Technical measures:	e.g., test results, logging, monitoring, data processing steps, emergency stop button
Comment:		
Criterion Fulfilled		

3 TR = abbreviation for German "Transparenz"

Human Oversight and Control (MA)[4]		
Human capacity to act		
Protection needs:		**Choose an item.** Choose an item.
Level achieved:		**Choose an item.** Choose an item.
Selected Measures	Organisational measures:	e.g., user instruction, training, specialised and technical documentation, reports
	Technical measures:	e.g., test results, logging, monitoring, data processing steps, emergency stop button
Comment:		
Criterion Fulfilled		

Human oversight		
Protection needs:		**Choose an item.** Choose an item.
Level achieved:		**Choose an item.** Choose an item.
Selected Measures	Organisational measures:	e.g., user instruction, training, specialised and technical documentation, reports
	Technical measures:	e.g., test results, logging, monitoring, data processing steps, emergency stop button
Comment:		
Criterion Fulfilled		

4 MA = abbreviation for German “Menschliche Aufsicht und Kontrolle”

Reliability (incl. robustness and performance) (VE) [5]		
Performance and robustness		
Protection needs:		**Choose an item.**
		Choose an item.
Level achieved:		**Choose an item.**
		Choose an item.
Selected Measures	Organisational measures:	e.g., user instruction, training, specialised and technical documentation, reports
	Technical measures:	e.g., test results, logging, monitoring, data processing steps, emergency stop button
Comment:		
Criterion Fulfilled		

Fallback plans and functional safety		
Protection needs:		**Choose an item.**
		Choose an item.
Level achieved:		**Choose an item.**
		Choose an item.
Selected Measures	Organisational measures:	e.g., user instruction, training, specialised and technical documentation, reports
	Technical measures:	e.g., test results, logging, monitoring, data processing steps, emergency stop button
Comment:		
Criterion Fulfilled		

5 VE = abbreviation for German "Verlässlichkeit"

AI-specific cybersecurity (CY) [6]

General AI-specific cybersecurity		
Protection needs:		**Choose an item.**
		Choose an item.
Level achieved:		**Choose an item.**
		Choose an item.
Selected Measures	Organisational measures:	e.g., user instruction, training, specialised and technical documentation reports
	Technical measures:	e.g., test results, logging, monitoring, data processing steps, emergency stop button
Comment:		
Criterion Fulfilled		

Resilience against AI-specific attacks		
Protection needs:		**Choose an item.**
		Choose an item.
Level achieved:		**Choose an item.**
		Choose an item.
Selected Measures	Organisational measures:	e.g., user instruction, training, specialised and technical documentation, reports
	Technical measures:	e.g., test results, logging, monitoring, data processing steps, emergency stop button
Comment:		
Criterion Fulfilled		

6 CY = abbreviation for German "KI-spezifische Cybersicherheit"

General comments on testing of the AI system

[Please enter in this field any general observations, methodological notes or other particularities that were noted during the test. This may include, for example, comments on test procedures performed or not performed, the availability and quality of evidence, methodological limitations or general strengths and weaknesses of the AI system. Where appropriate, you can also add aspects such as "Changes since last evaluation", "Gaps identified" and "Planned improvements".]

[Template for "Changes since last evaluation", "Gaps identified" and "Planned improvements":]

Changes since the last evaluation

[Briefly state the areas in which evaluation or evidence has improved.]

Identified gaps (only if there are unfulfilled criteria)

Individual criteria, such as [criteria names], could not fully achieve the required level of protection. The reasons for non-compliance are documented in the respective sections or in the "Comment" field and should be taken into account when making improvements. These gaps must be addressed as a matter of priority.

Planned improvements

[In addition to measures to close identified gaps, this section could also document further planned improvements that are intended to contribute to the optimisation of the AI system or the testing process, regardless of existing gaps.]

- [If gaps have been identified:] Targeted measures are planned to close the identified gaps. This includes further technical tests, process optimisation and extended documentation. The implementation of these improvements will be tracked and verified in the next review.
- [Provided **no** gaps have been identified:] No gaps were identified during the test. Accordingly, no further improvement measures are planned to close gaps.

Final conclusion

For the AI system [name of AI application], measures have been taken [for all/not all] applicable criteria that are suitable to address the identified protection needs.

Criteria [criteria names] were not met. Criteria [criteria names] were exceeded.

4. Confirmations

The persons listed below are responsible for carrying out this evaluation. They confirm with their signature that the test procedures described in this assessment report have been properly carried out in accordance with their respective responsibilities (see section 1.2) and that the statements made are true. They also confirm for their respective areas of responsibility that relevant and accurate evidence has been provided to show that the scope and effectiveness of the measures are sufficient to plausibly cover the identified protection needs. However, they do not accept any responsibility for statements made beyond their responsibility.

Several signatures are required if different ratings are used. This is the case, for example, if different minimum levels have to be met to fulfil different levels of protection needs. In this case, all relevant persons named in section 1.2 with testing responsibility must sign.

Date. [Date]	Date. [Date]	Date. [Date]
Name: [Name of the auditor]	Name: [Name of the auditor]	Name: [Name of the auditor]
Department: [Name of the department]	Department: [Name of the department]	Department: [Name of the department]
Involved in the development of the evaluated AI tool: yes/no	Involved in the development of the evaluated AI tool: yes/no	Involved in the development of the evaluated AI tool: yes/no
Test/audit experience: yes/no	Test/audit experience: yes/no	Test/audit experience: yes/no
_____ Signature	_____ Signature	_____ Signature

Disclaimer
The self-assessment offered serves exclusively as a voluntary internal evaluation by the participating organisation. It does not constitute an official test, certification or legally binding assessment. acatech assumes no liability for the completeness, accuracy or legal effect of the results of this self-assessment. Use is at the participating company's own risk and responsibility. acatech is only liable in cases of intent or gross negligence and in cases of legal liability.

List of authors

acatech
Carolin Anderson
Simon Boffen
Dr. Philipp Heß
Adrian Meisner

AI Quality & Testing Hub
Dr. Simone Amoroso
Paul Luca Palupski
Dr. Cord Schlötelburg
Hosei Halim
Paula Hoffmann

Fraunhofer IAIS
Dr. Maram Akila
Dr. Daniel Becker
Rebekka Görge
Dr. Henrik Junklewitz
Dr. Michael Mock
Dr. Maximilian Poretschkin
Anna Schmitz
Sebastian Schmidt

PricewaterhouseCoopers
Lina Antje Gühne
Alina Kudanova
Nan-Hee Kang
Laszlo Kühl
Jan-Niklas Nieland
Hendrik Reese

TÜV AI.Lab
Dr.-Ing. Marc P. Hauer
Leonie Löbenberg
Matthias König
Dr. Christoph Poetsch
Franziska Weindauer

VDE
Nora Dörr
Andreas Hauschke
Dr. Thorsten Prinz