Federal Office
for Information Security

Test Criteria Catalogue for AI Systems in Finance

Authors

This evaluation catalogue was developed by d-fine GmbH in cooperation with TÜV Informationstechnik GmbH and Fraunhofer AISEC on behalf of the Federal Office for Information Security (BSI). This publication was authored by the following persons.

d-fine GmbH:

Ulf Menzler

Henry Kleineidam

Cleo Matzken

Todor Dobrikov

Philipp Herrmann

Petra Gutjahr

Alexander Mehlmann

TÜV Informationstechnik GmbH:

Danos Vasilios

Fabian Langer

Thora Markert

Melanie Bohnen

Fraunhofer AISEC:

Pascal Debus

Sebastian Issel

Kilian Tscharke

Alexander Wagner

We would like to express our gratitude to all participants who contributed their expertise through interviews and discussions, and to those who took the time to work with us through questionnaires and workshops. Their insights, experiences, and recommendations were invaluable to the creation of this catalogue.

Publication Notes

Given the international relevance of trustworthy AI in the financial sector and the widespread applicability of the EU AI Act across member states and beyond, this publication was prepared in English to ensure broader accessibility and facilitate collaboration with international stakeholders. English serves as the standard language in technical, regulatory, and academic discourse on AI, making it the most appropriate choice for addressing a diverse audience, including researchers, industry professionals, and policymakers across Europe and globally.

Publisher

Federal Office for Information Security
`https://www.bsi.bund.de`

Version	Date	Change
1.0	27.05.2025	Provisional final version

Federal Office for Information Security
P.O. Box 200363, 53133 Bonn, Germany
E-mail: ki-kontakt@bsi.bund.de
Internet: https://www.bsi.bund.de

Contents

Part I

Catalogue Structure

Chapter 1

Introduction

The following criteria catalogue was developed as part the of BSI Project 587 **"Development and testing of test criteria, requirements and methods for AI systems in the financial sector"** in the following referred to as **"AICRIV Finance"**.
One of the priorities this project is the development of evaluation criteria and requirements for AI applications and systems in the aforementioned domain. Building on existing approaches and information sources, e.g., standards, guidelines and regulations, a set of criteria for secure and trustworthy AI systems in finance was created and summarized in this first prototype of the AICRIV Finance criteria catalogue.
The catalogue provides practical criteria for testing AI systems and suggests suitable test methods and tools for technical and document-based testing. In addition, a process for applying the catalogue is described. At a later project stage, the developed approach and the catalogue will be tested using specific use cases to demonstrate their practical feasibility. The insights gained from this, along with further project experiences, will be incorporated into the final design of the criteria catalogue and the application process.
The focus is on developing specific testing approaches that can be applied to both the selected use cases and AI models in finance in general. The evaluation criteria and requirements are motivated by specific threat scenarios and attack vectors and cover security-related risks that affect the trustworthy use of AI systems. In addition to traditional IT security, it also addresses AI-specific aspects such as data quality, functionality, security, transparency, and bias. The aim is to provide a comprehensive basis for evaluating and securing AI systems in the financial sector.
This document provides a comprehensive presentation of the developed criteria catalogue and its application. Part I of this document describes the development of the catalogue, including the resources used and the process for deriving relevant content such as generic criteria, procedures, and testing tools. In addition, the catalogue structure is explained in detail and the assignment of the selected criteria to the requirements of the EU AI Act (AI Regulation) is shown. Part II contains the first prototype of the AICRIV Finance catalogue with the evaluation criteria.

Chapter 2

Criteria Dimension

The catalogue is thematically divided into 10 dimensions, although there are thematic overlaps, each developed criterion is assigned to a single dimension:

- **AI Security & Robustness**: Robustness of the AI application against both natural and adversarial perturbations of the input, as well as the analysis and documentation of potential risks of the AI application.
- **IT Security**: Classical IT security of the overall system and its components, including access controls, firewalls, and security updates.
- **Monitoring**: Technical monitoring of the AI application, as well as monitoring the behavior of the AI application to analyze performance and ensure security.
- **Performance**: Correct functionality and efficiency of the AI application, i.e. quality of output, as well as reproducibility and reliability.
- **Governance**: Structural and organizational conditions for the responsible and safe use of AI applications.
- **Human Oversight**: Monitoring of the AI application by humans and participation of people in the development and use of the AI application.
- **Fairness**: Ensuring that the operation of an AI application complies with predetermined ethical values and that there is sufficient diversity in the data.
- **Transparency**: Transparency in the decisions of the AI application, especially towards the user, but also towards other stakeholders, i.e. disclosure of how the AI application works.
- **Data Quality & Management**: Ensuring data quality and responsible data management, as well as compliance with current data protection regulations.
- **Development**: Requirements for the development phase and corresponding documentation of the development of the AI application, ranging from architecture and training process to validation, testing, and deployment.

Chapter 3

Example Entry

This chapter provides an example entry showing the structure of a criteria entry from the catalogue. The individual information fields within the entry contain brief descriptions of the corresponding content.

Example Criteria Entry

EX-01 *Criterion ID*

Criterion name

Relevance based on use case parametrization

Relevant filter logic applicable for this criterion

Evaluation requirement

Brief description of what the criterion is about

Evaluation principle

Document-based *or* Test-based

Supportive guidance

Supplementary definitions: *Explanation or definition of terminology*

Exemplary [parameters]: *Examples or options for the variables in the []-brackets to be filled with*

Additional guidance regarding the requirement: *Guidance and extra information that help with understanding of the criterion*

Evaluation method

Instructions for the testing and evaluation of the criterion

Evaluation tools

Suggestions for evaluation tools and methods for testing of the evaluation criteria

Reference to EU AI Act

References to corresponding articles of the EU AI Act

Note: The fulfilling of a requirement of the AICRIV Finance criteria catalog is not automatically equivalent to an adherence to the requirements of the EU AI Act.

First of all, each evaluation criterion is assigned a unique Criterion ID and a Criterion name. The ID consists of an abbreviation for the dimension the criterion is assigned to and a sequence number, e.g., TR-07 for the seventh criterion in the transparency dimension. The name of the criterion is another means of identifying the criterion and provides a first indication of its topic.
The Relevance based on use case parametrization states the relevant questions and answers of the use case questionnaire that determine whether the criterion should be applied. The questionnaire, along with the general approach in which it is applied, is presented in Chapter 5.
The information given in the Evaluation requirement field specifically describes the test objective. The requirement is formulated generically, i.e., it may have to be adapted for the evaluation of a specific system. It does not contain any specific references to current methodological approaches, which helps to ensure that the requirements remain relevant in the long term. The evaluation requirement may contain parameters in square brackets that must be filled in based on the nature of the target of the evaluation and also the evolving state of the art.
The stated principle in the Evaluation principle field is directly related to the evaluation method and defines the type of procedure for the evaluation of the requirement. It is either document-based or test-based.
The Supportive guidance contains definitions that are helpful for understanding the evaluation requirement and additional information on fulfilling the requirement. This could be, for example, recommendations on what specific information should be included in a particular documentation. These notes support the user of the catalogue in evaluating the criteria. In addition, common examples for the placeholders in the evaluation requirement can be found.
The Evaluation method, on the other hand, gives specific instructions to the tester as to how the fulfillment of the evaluation requirement can be checked. The Evaluation tool field supplements the evaluation method with relevant technical tools that can be used to check the evaluation requirement.
The Related AI Act references field contains thematically related requirements from the AI Regulation and thus offers users a direct link between the requirements for AI systems applicable in the EU and the AICRIV criteria catalogue. Further information on the mapping is described in Chapter 4.

Chapter 4

Mapping of the Evaluation Criteria to the Requirements of the EU AI Act

The AI Act [1] sets out extensive provisions in connection with the development, placing on the market and deployment of AI systems, most of which are aimed at providers of such systems (see in particular Art. 4, 6-22, 50, 72 and 73 AI Act).
A natural or legal person is deemed to be a provider of an AI system if it develops an AI system and places it on the market or puts it into service for its own purposes. The mere deployment of (non-self-developed) AI systems is also subject to certain provisions (see in particular Art. 4, 26, 27, 49, 50 and 86 of the AI Act).
The majority of the regulations only apply to AI systems that are used in predefined high-risk areas (so-called high-risk AI systems). Outside the high-risk area, the development and operation of certain AI systems is subject to transparency provisions (e.g., when chatbots interact with natural persons). In addition, both providers and deployers of AI systems must ensure that their employees have a sufficient level of AI literacy (Art. 4 AI Act).
Most of the evaluation criteria in the criteria catalogue are thematically related to many of the provisions of the AI Act for AI systems. The criteria catalogue can therefore help to meet the provisions of the AI Act. For this purpose, the criteria catalogue contains a field with references to corresponding articles of the AI Act.
It should be noted that the provisions of the AI Act are formulated in a generic or principle-based manner and therefore leave some room for interpretation. European standards and guidelines, which are intended to specify the provisions, are still being developed. As a result, the criteria have been linked to the provisions through the use of a certain degree of interpretation.

Important: Therefore, the fulfillment of an evaluation criterion is **not** to be equated with the full compliance with the corresponding provision of the AI Act. Rather, the fulfillment of an evaluation criterion is to be regarded as a possible contribution to the compliance with the corresponding provision of the AI Act.

[1] https://eur-lex.europa.eu/eli/reg/2024/1689/

The assessment of whether and to what extent the provisions of the AI Act are met by fulfilling the evaluation criteria may be at the discretion of the auditor who checks compliance with the EU AI Act in the future. Some criteria are relevant both for providers and for deployers of AI systems. In such cases, both the provisions of the AI Act for providers and those for deployers were assigned to the criteria.

A central and overarching provision of the AI Act for providers of high-risk AI systems is the maintenance of a quality management system (QMS, Art. 17 AI Act) to ensure permanent conformity with the AI Act. It can be assumed that most of the aspects addressed in the criteria catalogue are part of good quality management of AI systems. For this reason, the QMS provision was assigned to the majority of the criteria. It should be noted that for financial institutions, the AI Act assumes that some essential components of a QMS are already in place, as they are already subject to extensive internal governance provisions under sector-specific regulation (see Art. 17 (4) EU AI Act).

In addition to the provisions for AI systems, the AI Act also formulates provisions for the development and placing on the market of "general purpose AI models" (so-called "GPAI models", see Art. 53 and Art. 55 AI Act), which are aimed at providers of such models. A natural or legal person is deemed to be a provider of a GPAI model if it places a GPAI model on the market (see Art. 3(4) AI Act). As providers of GPAI models are not included in the addressees of the criteria catalogue, the mapping is limited to the provisions for AI systems; the provisions for GPAI models were not considered further.

Chapter 5

A Risk-based, Holistic and Generalizable Audit Approach

Among other things, the criteria catalogue is also intended to serve as a self-assessment tool for market participants, e.g., as preparation for an upcoming audit. Not all evaluation criteria are equally relevant for all AI systems. To apply the catalogue, a corresponding generalized testing approach was therefore developed which can be applied to various scenarios and use cases in the financial industry. For this purpose, the evaluation process is adapted based on the risk and characteristics of the system under consideration. The aim is to create a flexible testing approach that can be transferred to different application areas and that meets the specific and diverse provisions of AI applications in the financial sector.
For the evaluation of the system, whether as part of a self-assessment or by external auditors, a certain level of technical understanding is generally required, particularly in the field of artificial intelligence. Basic knowledge of AI and the existing system is sufficient for answering the initial questionnaire to classify the use case and select suitable evaluation criteria. For the actual evaluation, however, more in-depth specialist knowledge is required in order to implement the described test procedures and the supportive guidance, to use recommended test tools if necessary, to interpret test results correctly and to make a well-founded final assessment.
Due to the heterogeneity of use cases, e.g., in terms of data foundation, area of application, architecture, etc., each evaluation of a use case requires an individual selection of relevant criteria in order to ensure a targeted and efficient testing. At the beginning, a questionnaire helps to identify the relevant criteria for the evaluation process. This allows relevant criteria to be filtered out specifically for the respective use case. For example, the questionnaire asks whether personal data is processed in order to evaluate the relevance of fairness-related criteria.
The process of an evaluation using the AICRIV Finance criteria catalogue begins, as shown in Figure 5.1, with the filtering or parametrization of the use case to be tested. The relevant criteria are then output, which must be adapted according to the system under consideration so that they can be applied directly in a (practical) evaluation. A test-based criterion requires a specific method or tool for testing that can be adapted using examples from the supportive guidance. Furthermore, thresholds must be defined according to the selected method or metric and in

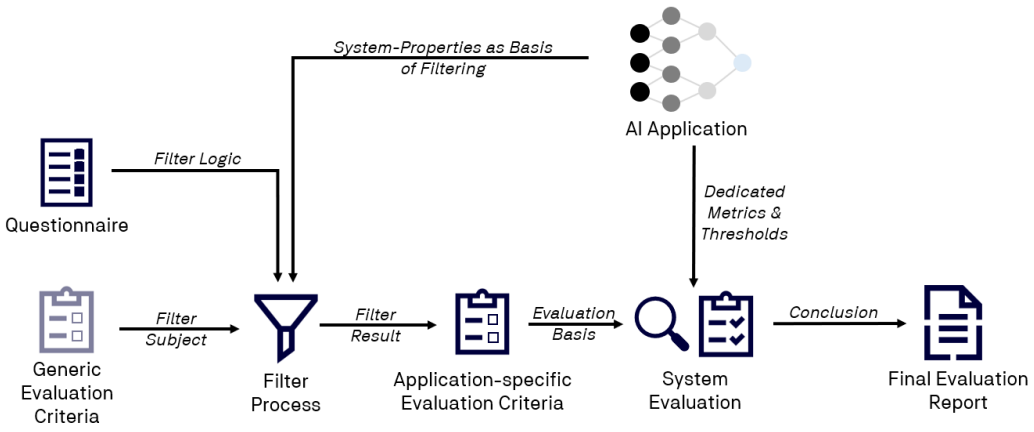

Figure 5.1: Application of the criteria catalogue as a process-based representation.

line with the use case requirements under consideration, that define when a test is considered to be passed. If necessary, the evaluation method is also specified accordingly and a suitable evaluation tool is selected to check the requirement. In this way, the generic criteria are tailored to the use case under consideration.
In the final step, the criteria catalogue is applied in practice. The evaluation method suggests how the criterion should be checked. The evaluation tool section provides examples for means of evaluation to investigate the fulfillment of the criterion. When testing the AI application using the specific criteria, particular focus should be placed on correct documentation in order to ensure transparency, traceability and consistency in the implementation and testing. A final evaluation report documents the procedure and findings for all evaluated criteria.

Questionnaire for Filtering Relevant Criteria

Q1 Does the system integrate Generative AI Foundation Models like Large Language Models?

- Yes: The system integrates at least one Generative AI Foundation Model.
- No: Other approaches such as Deep Learning or Shallow Learning Methods were used.

Q2 Is the architecture of involved AI models fully known?

- Yes: The architecture is fully known.
- No: The architecture is not or only partially known.

Q3 Are the parameters (e.g., weights or probability scores) of involved AI models known?

- Yes: There is full knowledge of parameters of the involved AI models.

- No: There is only partial or zero knowledge of parameters of the involved AI models (e.g. usage of third-party model with API access providing model results and probability scores or other meta-information / providing model results only).

Q4 Is the training data fully available for inspection?

- Yes: The training data is fully available and can be inspected.
- No: The training data is only partially known or completely unknown.

Q5 Has at least one integrated model been trained, even partially?

- Yes: Training of at least one integrated model was performed partially (e.g., with fine-tuning) or from scratch.
- No: Not at all (e.g. because of using pre-trained / off-the-shelf models).

Q6 Was personally identifiable or sensitive data used in the training of involved AI models?

- Yes: Personal data was used to train the involved AI models.
- No: Personal data was not used to train the involved AI models.

Q7 Does the AI system interact directly with external consumers (e.g., customers)?

- Yes: The model interacts directly with external consumers.
- No: The system does not interact directly with external consumers.

Q8 Is the model required to explain the reasons for certain conclusions?

- Yes: The model must explain reasoning for specific inference results.
- No: A black box result by the system is sufficient.

Q9 Does the AI system act completely autonomously (human-out-of-the-loop)?

- Yes: The AI system operates autonomously with no human oversight.
- No: The system integrates a human (either human-on-the-loop, human-in-the-loop, or human control).

Q10 Does the system have an impact on society?

- Yes: The model has a societal impact.
- No: The model has no societal impact.

Q11 Is the AI system built/operated on a hybrid infrastructure integrating on-premise as well as cloud components?

- Yes: The system is built/operated on a hybrid infrastructure.
- No: The system is built/operated on either an on-premise or cloud infrastructure.

Part II

Criteria Catalogue

Chapter 6

AI Security & Robustness

Contents

AISR-01

Handling of AI specific security incidents

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

The system shall comply with applicable (cloud) standards and regulations, ensuring necessary security controls, incident management, and auditing. Any AI model security incidents shall be promptly addressed, logged, and documented.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Security Controls:
Measures and mechanisms implemented to safeguard systems, data, and processes against threats, ensuring confidentiality, integrity, and availability as per applicable regulations and standards.
AI-Specific Security Incidents:
Events where AI systems are compromised, misused, or malfunction due to vulnerabilities, attacks, or breaches, potentially impacting model integrity, data privacy, or system behaviour.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Regulations & standards: e.g. Cloud Computing Compliance Criteria Catalogue (C5), NIST, NIS-2.
For the handling of security incidents for example the Policy for Security Incident Management (C5-SIM-01) from C5 can be consulted for more information.

Evaluation method

Verify that the system is compliant to necessary standards and regulations. Check whether an adequate process for handling security incidents are in place and previous identified security incidents related to the AI model were addressed and documented.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)
- 26 Obligations of deployers of high-risk AI systems
- 73 Reporting of serious incidents

AISR-02

Assessment of threats in third party models

Relevance based on use case parametrization

- The architecture is not or only partially known (Q2) OR
- There is only partial or zero knowledge of parameters of the involved AI models (e.g. usage of third-party model with API access providing model results and probability scores or other meta-information / providing model results only) (Q3) OR
- The training data is only partially known or completely unknown (Q4)

Evaluation requirement

If the training data and/or model originate from third party, they should be analyzed for potential attack vectors and the results shall be documented.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Third Party:
An external entity (not directly affiliated the system's owner), responsible for providing training data, models, or other services.
Attack Vectors:
Paths or methods leveraged by adversaries to exploit vulnerabilities in systems, such as data poisoning, model manipulation, or injection of malicious triggers or backdoors.

Exemplary [parameters]

Additional guidance regarding the requirement

Considerations should be at least given to:
- Triggers;
- Backdoors;
- Anomalies in the training data.

Evaluation method

Ensure that third-party components are analysed for potential attack vectors. Ensure adequate documentation of the analysis.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 10 Data and data governance
- 11 Technical documentation
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)
- 18 Documentation keeping
- 25 Responsibilities along the AI value chain

AISR-03

Risks from non-specified inputs

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

The behaviour of the system for, and the risk associated with, inputs that do not conform to the specified requirements shall be analysed with [methods] and documented.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Associated Risks:
The likelihood and potential impact of adverse effects, such as security vulnerabilities or functional failures, caused by handling non-conforming inputs.
Non-Conforming Inputs:
Inputs that deviate from defined standards, such as malformed, corrupted, or improperly formatted data, potentially affecting system integrity or performance.

Exemplary [parameters]

[methods]: Testing frameworks, Robustness assessment tools, American Fuzzy Lop ++

Additional guidance regarding the requirement

Damaged or manipulated inputs can be corrected if it is safely possible.

Evaluation method

Check that the behaviour of the system for inputs that do not meet the specified requirements is analysed and documented.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 9 Risk Management System
- 10 Data and data governance
- 11 Technical Documentation
- 13 Transparency and provision of information to deployers
- 17 Quality Management System (QMS)
- 18 Documentation keeping

AISR-04

Assessment of security threats

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

Procedures shall be implemented to continuously monitor and assess new threats related to the AI model(s).

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Not applicable.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

All potential threats to the AI system should be carefully considered, including design and technical faults, environmental risks, and cyber threats. A thorough review of data and features is essential to identify vulnerabilities and threats throughout the entire AI lifecycle.
These threats can include specific technical faults (e.g., hardware information leakage) and cyber threats (e.g., DDoS attacks, backdoors).
In relation to privacy threats, the analysis should focus on identifying and documenting vulnerabilities in the ML model that could lead to unintended personal inferences or potential side-channel attacks, which might result in the leakage of personal data.

Evaluation method

Verify that procedures are in place to continuously monitor and assess emerging threats. Ensure that the process covers all potential threats as defined in the supportive guidance.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 9 Risk Management System
- 11 Technical Documentation
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)
- 18 Documentation keeping

AISR-05

Documented security review protocol

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

The frequency, scope, and method(s) for security and robustness reviews of productive models shall be documented.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Security Reviews:
Systematic assessments to detect and address vulnerabilities in productive models, ensuring they are safeguarded against potential threats, breaches, and unauthorized access.
Robustness Reviews:
Complete evaluations of a model's ability to maintain functionality and performance under adversarial attacks, unexpected inputs, or challenging conditions.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Not applicable.

Evaluation method

Ensure adequate documentation of the frequency, scope and methodologies for independent safety reviews of product models.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 9 Risk Management System
- 11 Technical Documentation
- 13 Transparency and provision of information to deployers
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)
- 18 Documentation keeping

AISR-06

Formal verification

Relevance based on use case parametrization

- Other approaches such as Deep Learning or Shallow Learning Methods were used (Q1) AND
- The architecture is fully known (Q2) AND
- There is full knowledge of parameters of the involved AI models (Q3)

Evaluation requirement

The possibility of verifying the correctness of the output by formal verification, at least of modules of the full system shall be considered.

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Formal Verification:
The use of mathematical techniques to ensure that the system or its components operate as intended by checking against predefined rules and specifications.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Formal verification can be conducted by mapping functions to smaller AI models that can be individually checked and verified.

Evaluation method

Verify that formal verification has been considered to verify the correctness of the output. Evaluate the formal guarantees globally (for the whole system) or module-wise using the suggested evaluation tools from the supportive guidance.

Evaluation tools

- ReLUPlex - can be used to formally verify neural network properties by checking activation functions and constraints in individual AI model components.
- Abstract Domain Transformers - can be used to perform static analysis and verify AI model behavior against predefined specifications.

Reference to EU AI Act

- 9 Risk Management System
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)

AISR-07

White box robustness

Relevance based on use case parametrization

- The architecture is fully known (Q2) AND
- There is full knowledge of parameters of the involved AI models (Q3)

Evaluation requirement

The attack surface for white-box evasion attacks shall be assessed with appropriate [attack vectors].

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

White-Box:
Scenarios where attackers have full access to the internal structure, parameters, and algorithms of a system or model, allowing detailed analysis and exploitation.
Evasion Attacks:
Methods where attackers modify inputs to trick a system or model into making wrong or unexpected decisions without changing the system itself.

Exemplary [parameters]

[attack vectors]: Standard Adversarial Attacks - FGSM (Fast gradients sign method), PGD (Project gradient descent), BIM (Basic Iterative Method), CW (Carlini-Wagner).

Additional guidance regarding the requirement

Not applicable.

Evaluation method

Verify that the attack surface has been assessed for white box attacks. Verify that the system was tested against relevant white-box attacks and demonstrates adequate resistance.

Evaluation tools

- CleverHans - can be used to generate adversarial examples and evaluate the AI model's robustness against common white-box evasion attacks.
- Adversarial Robustness Toolbox - can be used to systematically test AI models with a range of adversarial attacks and measure their resistance to white-box evasion techniques.

Reference to EU AI Act

- 9 Risk Management System
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)

AISR-08

Gray and black box robustness via surrogate model

Relevance based on use case parametrization

- The architecture is not or only partially known (Q2) OR
- There is only partial or zero knowledge of parameters of the involved AI models (e.g. usage of third-party model with API access providing model results and probability scores or other meta-information / providing model results only) (Q3)

Evaluation requirement

The attack surface for gray and black box evasion attacks shall be assessed with appropriate [attack vectors].

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Gray Box:
A scenario where attackers have partial knowledge of the system or model, such as access to some parameters, architecture details, or limited input-output behaviour.
Black Box:
A scenario where attackers have no internal knowledge of the system or model and rely on observable input-output behaviour.
Evasion Attacks:
Methods where attackers modify inputs to trick a system or model into making wrong or unexpected decisions without changing the system itself.

Exemplary [parameters]

[attack vectors]: e.g., Standard gray and black box attacks - Transfer attack with similar (pre-trained) surrogate model, Training of a (fine-tuned) surrogate model if high bandwidth with original model is available, Z00 (Zero order attack), Boundary Attack, Single Pixel Attack.

Additional guidance regarding the requirement

The known information should be assumed realistically (obtainable by an attacker).

Evaluation method

Verify that the attack surface has been assessed for gray and black box attacks. Verify that the system was tested against relevant gray-/black-box attacks and demonstrates adequate resistance.

Evaluation tools

- CleverHans - can be used to simulate gray-box and black-box attacks using techniques such as transfer attacks, surrogate models, and boundary attacks.
- Adversarial Robustness Toolbox - can be used to conduct advanced black-box and gray-box evasion tests.

Reference to EU AI Act

- 9 Risk Management System
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)

AISR-09

White box attacks

Relevance based on use case parametrization

- The architecture is fully known (Q2) AND
- There is full knowledge of parameters of the involved AI models (Q3)

Evaluation requirement

The attack surface for white-box information extraction attacks shall be assessed with appropriate [attack vectors].

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

White-Box:
Scenarios where attackers have full access to the internal structure, parameters, and algorithms of a system or model, allowing detailed analysis and exploitation.
Extraction Attacks:
Methods where attackers try to obtain sensitive information from a model, such as its training data, parameters, or underlying logic, by analyzing its responses or internal structure.

Exemplary [parameters]

[attack vectors]: e.g., Membership inference attacks, Property/Attributed Inference attacks, Reconstruction Attacks (of training samples).

Additional guidance regarding the requirement

Not applicable.

Evaluation method

Verify that the attack surface has been assessed for white-box information extraction attacks. Verify that relevant attack vectors have been evaluated and that the sensitive information, such as training data, model parameters, or underlying logic, can not be extracted.

Evaluation tools

- Adversarial Robustness Toolbox - can be used to conduct membership inference, property inference, and reconstruction attacks to test the model's vulnerability to information extraction.
- TensorFlow Privacy - can be used to assess differential privacy protections and analyze the model's susceptibility to white-box extraction techniques.
- Privacy Meter - can be used to evaluate membership inference risks and determine if attackers can infer whether specific data was used in model training.

Reference to EU AI Act

- 9 Risk Management System
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)

AISR-10

Gray and black box attacks

Relevance based on use case parametrization

- The architecture is not or only partially known (Q2) OR
- There is only partial or zero knowledge of parameters of the involved AI models (e.g. usage of third-party model with API access providing model results and probability scores or other meta-information / providing model results only) (Q3)

Evaluation requirement

The attack surface for gray and black box information extraction attacks shall be assessed with appropriate [attack vectors].

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Gray Box:
A scenario where attackers have partial knowledge of the system or model, such as access to some parameters, architecture details, or limited input-output behaviour.
Black Box:
A scenario where attackers have no internal knowledge of the system or model and rely on observable input-output behaviour.
Extraction Attacks:
Methods where attackers try to obtain sensitive information from a model, such as its training data, parameters, or underlying logic, by analyzing its responses or internal structure.

Exemplary [parameters]

[attack vectors]: e.g., Label-Only Membership Inference Attacks, Model Extraction, Prompt Injection, Divergence Attack.

Additional guidance regarding the requirement

The known information should be assumed realistically (obtainable by an attacker).

Evaluation method

Verify that the attack surface has been assessed for grey-/and black-box information extraction attacks. Verify that relevant attack vectors have been evaluated and that the sensitive information, such as training data, model parameters, or underlying logic, can not be extracted.

Evaluation tools

- Adversarial Robustness Toolbox - can be used to conduct label-only membership inference, model extraction attacks, and divergence attacks to evaluate how much information can be extracted with limited access.
- TensorFlow Privacy - can be used to assess privacy-preserving techniques and analyze how effective they are against gray-box and black-box extraction attacks.
- Privacy Meter - can be used to evaluate the risk of membership inference and determine whether attackers can infer training data presence in black-box scenarios.

Reference to EU AI Act

- 9 Risk Management System
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)

AISR-11

Backdoor attacks

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

[Attack vectors] shall be carried out to exploit identified vulnerabilities considering the integrity of relevant data sets and their impacts.

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Backdoor attack: A backdoor attack is a type of data poisoning attack where an adversary implants a hidden pattern (trigger) in the training data so that the model behaves normally on clean inputs but produces incorrect or malicious outputs when the trigger is present in the input. The goal is to create a model that is covertly controlled under specific conditions while remaining undetected during regular use.

Exemplary [parameters]

[attack vectors]: e.g., Data poisoning or data tampering through backdoors.

Additional guidance regarding the requirement

Not applicable.

Evaluation method

Verify that the attack surface for data integrity attacks, e.g., data poisoning or tampering, has been assessed. Verify that relevant attack vectors have been evaluated and the system demonstrates adequate resistance.

Evaluation tools

- Adversarial Robustness Toolbox (ART) - can be used to conduct data poisoning and backdoor detection, simulating adversarial manipulations of training datasets.
- LabelFix - can be used to detect and analyze mislabeled or tampered data.

Reference to EU AI Act

- 9 Risk Management System
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)

AISR-12

Robustness on corner cases

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

The robustness of the system on corner cases and/or boundary values shall be assessed with suitable [measures] and documented.

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Robustness:
Robustness in AI systems refers to the ability to function reliably and consistently even under uncertain conditions, such as unexpected inputs or disturbances. A robust AI system delivers consistent and correct results, even when faced with noisy or manipulated data such as adversarial attacks. The goal is to ensure that the system is not easily disrupted by external influences.
Boundary Values:
Inputs at the edge of an allowed range or threshold, used to evaluate a system's performance and stability under extreme conditions.

Exemplary [parameters]

[measures]: Empirical statistical robustness tests for identified corner cases, testing on the boundary values of the system (if quantifiable).

Additional guidance regarding the requirement

Not applicable.

Evaluation method

Verify if the AI system maintains reliable and stable when exposed to extreme or unexpected inputs and assess that the robustness is compliant with the minimum acceptable robustness of the system.

Evaluation tools

- Adversarial Robustness Toolbox - can be used to generate adversarial perturbations and edge-case inputs to test model robustness.

Reference to EU AI Act

- 9 Risk Management System
- 11 Technical Documentation
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)
- 18 Documentation keeping

AISR-13

I/O security

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

Input and output shall be protected from tampering by classical [measures].

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Tampering:
The unauthorized alteration, manipulation, or interference with data, inputs, or outputs to compromise the integrity, functionality, or security of the system.

Exemplary [parameters]

[measures]: e.g., Data encryption, input validation and sanitization, authentication and authorization (see ITSEC-36), integrity checks, rate limiting and throttling, secure APIs and endpoints, audit logging and monitoring, human in the loop oversight.

Additional guidance regarding the requirement

Not applicable.

Evaluation method

Verify if classical security measures are in place to prevent unauthorized alteration or manipulation of data.

Evaluation tools

- Taint Analysis - can be used to track untrusted data sources and ensure that there is input validation, sanitization, and protection against tampering.
- Static Code Analysis - can be used to detect security vulnerabilities related to data integrity, authentication, authorization, and secure API handling.

Reference to EU AI Act

- 9 Risk Management System
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)

AISR-14

Countermeasures against evasion attacks

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

[Countermeasures] against evasion attacks shall be implemented and documented.

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Evasion Attacks:
Methods where attackers modify inputs to trick a system or model into making wrong or unexpected decisions without changing the system itself.

Exemplary [parameters]

[countermeasures]: e.g., Adversarial training, anomaly detection, employing multiple AI systems with different architectures or training data redundantly against adversarial attacks.

Additional guidance regarding the requirement

Not applicable.

Evaluation method

Verify that countermeasures against adversarial attacks are in place and work properly.

Evaluation tools

- CleverHans - can be used to generate adversarial examples and test the effectiveness of implemented countermeasures.
- Adversarial Robustness Toolbox - can be used to conduct structured adversarial attacks and evaluate the model's resistance to evasion attempts.

Reference to EU AI Act

- 9 Risk Management System
- 11 Technical Documentation
- 13 Transparency and provision of information to deployers
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)
- 18 Documentation keeping

AISR-15

Countermeasures against information extraction attacks

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

[Countermeasures] against information extraction attacks shall be implemented and documented.

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Extraction Attacks:
Methods where attackers try to obtain sensitive information from a model, such as its training data, parameters, or underlying logic, by analyzing its responses or internal structure.

Exemplary [parameters]

[countermeasures]: e.g.,
- Differential Privacy (epsilon-DP training (e.g., DP-SGD), Rényi Differential Privacy)
- Regularization Techniques (L2/L1 norm regularization, Dropout / DropConnect, Label smoothing)
- Data Augmentation (Synthetic data generation (GAN-based), Mixup and CutMix strategies)
- Knowledge Distillation (Use of student-teacher model architectures where only the student is exposed to queries)
- Model Compression/Pruning (Reduces memorization of individual training samples)
- Ensemble Models with Randomization (Stochastic selection among several models at inference time)
- Response Filtering (Limit to top-k or top-p outputs, Output obfuscation or rounding of probabilities)
- Access Restriction on Confidence Scores (Return only class labels instead of full confidence scores, Add noise to model outputs (especially probabilistic outputs))
- Rate Limiting / Throttling (Restrict frequency of API access or number of queries)

- Query Auditing and Anomaly Detection (Track query patterns to detect suspicious sequences indicative of extraction attempts)
- Dynamic Watermarking (Insert unique, hard-to-replicate patterns into model behavior to identify extraction or unauthorized use)
- Dataset Condensation or Sanitization (Remove outliers or memorization-prone samples from training datasets)

Additional guidance regarding the requirement

Not applicable.

Evaluation method

Verify that effective measures are in place to prevent unauthorized extraction of training data, model parameters, or internal logic

Evaluation tools

- Privacy Meter - can be used to evaluate the effectiveness of membership inference attack countermeasures.

Reference to EU AI Act

- 9 Risk Management System
- 11 Technical Documentation
- 13 Transparency and provision of information to deployers
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)
- 18 Documentation keeping

AISR-16

Countermeasures against data poisoning attacks

Relevance based on use case parametrization

- Traning of at least one integrated model was performed partially (e.g., with finetuning) or from scratch (Q5)

Evaluation requirement

[Countermeasures] against data poisoning attacks shall be implemented and documented.

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Poisoning Attacks:
A type of adversarial attack where malicious data is injected into the training dataset to corrupt the model's learning process, leading to degraded performance, biased outputs, or hidden malicious behaviours.

Exemplary [parameters]

[countermeasures]: e.g.,
- Data Collection & Ingestion (Access control to datasets (e.g., strict role-based access to data and version histories)
- Data provenance tracking using cryptographic methods (e.g., blockchain or signed datasets) Source validation and authentication, especially for third-party or crowd-sourced data, Differential access policies for sensitive training data
- Data sanitization and filtering (e.g., removing outliers, duplicates, or inconsistent entries)
- Use of trusted data sources and certified data providers)
- Poisoning Detection & Analysis (Anomaly detection models for spotting suspicious data points (e.g., clustering, distance-based, or graph-based methods)
- Influence function analysis to measure training point impact, Backdoor detection tools (e.g., activation clustering, spectral signatures)
- Statistical validation (e.g., comparing class distributions)

- Data slicing and canary examples to detect poisoning, Model behavior auditing after fine-tuning)
- Training-Time Defenses (Robust training algorithms (e.g., trimmed loss minimization, differential privacy, Deep k-NN), Noise injection in training,)
- Post-Training Monitoring & Evaluation (Model fingerprinting and output distribution tracking, Shadow model comparisons, Periodic re-evaluation against clean benchmarks, Interpretability tools for suspect behavior (e.g., SHAP, LIME))

Additional guidance regarding the requirement

Possible attacks: e.g., availability poisoning, targeted poisoning, backdoor poisoning, model poisoning, model skewing.

Evaluation method

Verify that suitable measures to detect and mitigate data poisoning are in place and work properly.

Evaluation tools

- LabelFix - can be used to detect mislabeled or manipulated training samples that may indicate poisoning attempts.
- Dioptra - can be used to simulate and evaluate poisoning attacks, enabling assessment of existing countermeasures and model robustness.

Reference to EU AI Act

- 9 Risk Management System
- 11 Technical Documentation
- 13 Transparency and provision of information to deployers
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)
- 18 Documentation keeping

AISR-17

Countermeasures against model theft attacks

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

[Countermeasures] against model theft attacks shall be implemented and documented, where the deployer is responsible for the intellectual property of the model or where the license terms of a third-party or open-source model require such protection.

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Model Theft Attacks:
Attempts by attackers to replicate or steal a machine learning model by exploiting access to its outputs, structure, or functionality.

Exemplary [parameters]

[countermeasures]: e.g., access control, model obfuscation, watermarking, homomorphic encryption, limited queries/output.

Additional guidance regarding the requirement

Not applicable.

Evaluation method

Verify that suitable countermeasures against model theft attacks are in place and work properly.

Evaluation tools

- Model Inversion Attack - those techniques can be used to assess whether sensitive model information can be extracted, evaluating the effectiveness of defenses.
- Adversarial Robustness Toolbox - can be used to simulate model theft attempts.
- Knockoff Nets - can be used to evaluate model extraction risks by replicating the behavior of a black-box model using query-based attacks.

Reference to EU AI Act

- 9 Risk Management System
- 11 Technical Documentation
- 13 Transparency and provision of information to deployers
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)
- 18 Documentation keeping

AISR-18

Countermeasures against output integrity attacks

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

[Countermeasures] against output integrity attacks shall be implemented and documented.

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Output Integrity Attack:
An attack that occurs when an adversary manipulates or alters the final results produced by the AI model without interfering with the input data or performing adversarial attacks on the model itself. In this type of attack, the input data remains untouched, but the generated output is modified or tampered with before it reaches the end user or decision-making system.

Exemplary [parameters]

[countermeasures]: e.g., Encryption, secure communication channels, output monitoring.

Additional guidance regarding the requirement

Not applicable.

Evaluation method

Verify that suitable countermeasures against output integrity attacks, i.e., attacks that manipulate the output of the AI system are in place and work properly.

Evaluation tools

- Penetration Testing Tools (e.g., OWASP ZAP) - penetration testing can be used to identify vulnerabilities in web applications and APIs where output data may be intercepted or manipulated.

Reference to EU AI Act

- 9 Risk Management System
- 11 Technical Documentation
- 13 Transparency and provision of information to deployers
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)
- 18 Documentation keeping

AISR-19

Residual risk mitigation

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

If countermeasures tested and derived against attacks do not lower the risk to an acceptable level, or if there are no specific countermeasures available, alternative measures not linked to specific threat scenarios shall be developed, implemented, and tested.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Not applicable.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Suitable alternative measures eventually need to be investigated for the individual combination of use case and residual risk. An initial literature research can be performed and based on the results a possible solution can be developed inhouse or by an external service provider.

Evaluation method

Analyse whether the countermeasures tested and derived against attacks reduce the risk to an acceptable level. Verify that alternative measures are implemented if not.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 9 Risk Management System
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)

Chapter 7

Data Quality & Management

Contents

DQM-01

Data planning

Relevance based on use case parametrization

- Traning of at least one integrated model was performed partially (e.g., with finetuning) or from scratch (Q5)

Evaluation requirement

The data planning process shall be documented

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Data planning: The process of strategically defining how data will be collected, processed, managed, stored, analyzed, and maintained to meet organizational objectives and compliance requirements.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

The data planning process should consider at least the following elements:
- The data model or data architecture necessary to achieve the data requirements;
- Plan for acquiring the necessary data as identified by the data requirements;
- Roles, skills and people necessary to execute the data quality process;
- IT and other resources necessary to execute the data quality process;
- Time and budget necessary to execute the data quality process;
- Acquiring the data according to the data requirements;
- Executing data quality measures according to the data quality model;
- Meeting legal requirements;
- Meeting other data requirements.

Evaluation method

Verify that the data planning was performed and documented accordingly. Considerations should be given to the listed elements in the supportive guidance.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 10 Data and data governance
- 11 Technical Documentation
- 17 Quality Management System (QMS)
- 18 Documentation keeping

DQM-02

Data acquisition

Relevance based on use case parametrization

- Traning of at least one integrated model was performed partially (e.g., with finetuning) or from scratch (Q5)

Evaluation requirement

The data acquisition shall be documented.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Data acquisition: The process of collecting, measuring or obtaining raw data from various sources to be used for training and validating machine learning models.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Considerations regarding data provenance should be given to:
- Compliance with the elements defined in the data planning process;
- Adherence to the principles of the data quality process;
- Data Ownership and Responsibility;
- Data Licensing and Usage Rights;
- Key data attributes identified during the data requirements process, such as provenance, bias, consistency, reliability, validity, data types, schema, format, and low data noise;
- The source of the data, whether it is internal or external;
- Verification of the trustworthiness of each data source used;
- The data collection process, specifying if it was conducted by humans, automated sensors, or

both;
- The method of data collection, such as surveys or streaming;
- In the case of personal data, the original purpose of data collection.

Evaluation method

Verify that the data acquisition is adequately documented, see the supporting guidance for an indication of what should be included in the documentation.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 10 Data and data governance
- 11 Technical Documentation
- 17 Quality management system (QMS)
- 18 Documentation keeping

DQM-03

Data quality requirements

Relevance based on use case parametrization

- Traning of at least one integrated model was performed partially (e.g., with finetuning) or from scratch (Q5)

Evaluation requirement

Data Requirements and data quality management shall be documented

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Data quality: The degree to which data meets the standards for (among others) consistency, informativeness, representativeness, timeliness, validity, reliability, completeness.
Data Management: The practice of organizing, storing, protecting and maintaining data to ensure its quality, accessibility, and usability throughout its lifecycle.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Data requirements should consider at least the following elements:
- Definition of quality standards, including accuracy, consistency, completeness, and timeliness of the data;
- Specification of necessary attributes or features that the data must contain to support model training;
- Determination of the volume of data needed to achieve reliable results and generalization;
- Assessment of statistical measures such as mean, variance, and distribution characteristics to understand dataset behavior;

- Assurance that the dataset accurately represents the population in terms of demographics, behaviors, and geographies relevant to the problem;
- Compliance with regulations (e.g., GDPR) and ethical guidelines regarding data privacy and usage;
- Clear guidelines on which data to include or exclude based on specific attributes or quality metrics;
- Uncertainty and variability assessment;
- Annotation Requirements for how data should be labelled or categorized, including quality assurance measures;
- Inclusion and exclusion criteria for data based on relevant attributes;
- Technical inclusion and exclusion criteria;
- Guidelines on whether and how synthetic data can be used to augment the dataset;
- Processes for evaluating, resolving, and closing data quality issues;
- Regular reviews and validations of the quality and relevance of the data throughout the model lifecycle.

Evaluation method

Verify that the data requirements were developed and documented accordingly. Considerations should be given to the listed elements in the supportive guidance.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 10 Data and data governance
- 11 Technical Documentation
- 13 Transparency and provision of information to deployers
- 17 Quality Management System (QMS)
- 18 Documentation keeping

DQM-04

Data quality assessment

Relevance based on use case parametrization

- Traning of at least one integrated model was performed partially (e.g., with finetuning) or from scratch (Q5)

Evaluation requirement

The quality of gathered data shall be assessed and documented with [measures] and [corrective measures] shall be in place to ensure stability.

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Stability: In the context of data, it refers to the degree to which data remains consistent, reliable and unchanged over time or under varying conditions. It is critical for maintaining trust in data-driven processes, ensuring consistency, and supporting robust decision-making.

Exemplary [parameters]

[measures]: Data Validation, Consistency Check, Bias Detection Tools, Statistical (Power) Analysis, cross-validation of labels
[corrective measures]: LabelFix, Bias Correction Tools, Trigger model retraining when data shift is detected.

Additional guidance regarding the requirement

Requirements according to DQM-03.

Evaluation method

Verify that a process for continuous assessment of data quality is in place and documented. Also verify that data subject to this process meets consistency, reliability, and bias-free standards, with corrective actions in place. The required data properties and key figures about the data that shall be validated are based on the data requirements (see DQM-03) directly or indirectly induced by performance/ functional requirements of the system. Examples for testing tools and corresponding data properties to test can be found in the column "Evaluation tool".

Evaluation tools

- TensorFlow Data Validation - can be used to analyze dataset distribution, detect anomalies..
- LabelFix - can be used to identify inconsistencies, and ensure labeling quality.
Statistical Power Analysis - can be used to assess data sufficiency and reliability.
- Power Analysis Tools - can be used to verify if the dataset is large enough to detect meaningful statistical effects and support robust decision-making.
Dimensionality Reduction Analysis (verify stability under varying data conditions):
- PCA (Principal Component Analysis) - can be used to analyze variance in datasets and detect instability due to redundant or highly correlated features.
- t-SNE - can be applied for high-dimensional data visualization to detect clustering inconsistencies that may indicate bias or data drift.
Fairness and Bias Detection Tools (verify data fairness and detect bias-related instability):
- Fairlearn - can be used to valuate the fairness of datasets by measuring disparate impact and mitigating algorithmic bias.
- IBM AI Fairness 360 - can be used for advanced bias detection, fairness metrics analysis.
Data Quality Assessment Tools (evaluate correctness, consistency, and reliability of labeled data):
- Talend Data Quality - can be used to detect anomalies, check data consistency, and validate dataset structure.

Reference to EU AI Act

- 10 Data and data governance
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality management system (QMS)

DQM-05

Data tracking procedures

Relevance based on use case parametrization

- Traning of at least one integrated model was performed partially (e.g., with finetuning) or from scratch (Q5)

Evaluation requirement

The traceability of data shall be ensured and documented.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Traceability: Refers to the ability to track, document, and verify the origin, history and transformation of data throughout its lifecycle.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Some factors to consider regarding traceability:
- Clear identification of all data records through indexing;
- Recording and authorization of all data changes;
- Placing the entire data processing software and libraries under version control;
- Consistency of metadata;
- Safe delivery or transfer of data.

Evaluation method

Verify that traceability of data is ensured regarding the factors stated in the supportive guidance.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 10 Data and data governance
- 12 Record-keeping
- 17 Quality Management System (QMS)
- 19 Automatically generated logs

DQM-06

Data integrity

Relevance based on use case parametrization

- Traning of at least one integrated model was performed partially (e.g., with finetuning) or from scratch (Q5)

Evaluation requirement

The integrity of the data shall be ensured through suitable [measures].

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Integrity: In terms of data, integrity ensures that data remains unchanged, uncorrupted, and true to its original form through collection, storage, transmission and processing.

Exemplary [parameters]

[measures]: e.g., Hashing, digital signatures, version control, backups.

Additional guidance regarding the requirement

Data integrity should be ensured through measures such as hashing, digital signatures, version control, and backups.
The following tools can be used for the implementation:
- MLOps can be used to evaluate whether data integrity checks, such as versioning, hashing, and digital signatures, are integrated into the machine learning lifecycle.
- DevOps can be used to assess if CI/CD pipelines incorporate data integrity measures, such as automated checks for corruption and rollback mechanisms.

Evaluation method

Verify that methods preserving the integrity of the used data are in place and are being used throughout the whole lifecycle.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 10 Data and data governance
- 17 Quality Management System (QMS)
- 19 Automatically generated logs
- 26 Obligations of deployers of high-risk AI systems

DQM-07

Data decommissioning

Relevance based on use case parametrization

- Traning of at least one integrated model was performed partially (e.g., with finetuning) or from scratch (Q5)

Evaluation requirement

The data decommissioning process shall be documented.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Data decommissioning: The process of securely and systematically retiring, archiving or disposing of data that is no longer needed.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Depending on the needs and (regulatory) requirements a plan for decommission of datasets shall be in place.
Every decommission plan shall include
-Clear descriptions and means for identification of the datasets that are covered by the decommision plan
-Requirements for the decommission of data, e.g., end of use, privacy/copyright infringements, etc.
-Involved personnel/roles and corresponding responsibilities
-Description of the decommissioning process with the main steps of identification of datasets, archiving of datasets (scope/extent of archiving, storage medium, encryption, etc.) and data

deletion (e.g., means of erasure such as digital override or physical destruction of media)
Every conducted data decommission shall be documented with a clear assignment to the datasets that were subject to the decommissioning.
Furthermore, a recovery plan for decommissioned (and archived) data should be implemented if necessary.

Evaluation method

Ensure that data can be decommissioned. Verify that the decommission plan is created, reviewed and met.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 10 Data and data governance
- 11 Technical Documentation
- 17 Quality Management System (QMS)
- 18 Documentation keeping

DQM-08

Users' consent to the use of their data

Relevance based on use case parametrization

- Personal data was used to train the involved AI models (Q6)

Evaluation requirement

It shall be ensured that individuals provide active informed consent for the AI system to process and store their data. It shall be ensured that individuals can perform their right to revoke this consent by allowing users to request the deletion of their data and to stop its processing.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Active informed consent: A type of consent where individuals actively and explicitly agree to the processing and storage of their data after being fully informed about the nature, purpose, and implications of that processing.
Revocation of consent: The ability of individuals to withdraw their previously granted consent at any time, making further processing or storage of their data unlawful unless another legal basis exists.
Right to deletion (Right to be forgotten): The right of individuals to have their personal data erased or deleted when it is no longer necessary for the purposes for which it was collected, or when consent has been withdrawn.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Individuals should be given clear, understandable information about:
- What data is being processed

- Why it is being processed
- How long it will be stored
- Who will have access to the data
- Any potential risk or consequences

The applicable law for this can be taken from the GDPR.

Evaluation method

Verify that individuals actively consent to processing the data and that it is logged appropriately. Ensure that data can be deleted if it is requested by a user.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 10 Data and data governance

DQM-09

Users' right to be forgotten

Relevance based on use case parametrization

- Personal data was used to train the involved AI models (Q6)

Evaluation requirement

Measures shall be in place to realize a user's right to be forgotten.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Right to be forgotten:
The right of individuals to have their personal data erased or deleted when it is no longer necessary for the purposes for which it was collected, or when consent has been withdrawn (GDPR Art. 17).

Exemplary [parameters]

[measures]: e.g., distillation, retraining, federated learning, machine unlearning methods

Additional guidance regarding the requirement

Not applicable.

Evaluation method

Verify that a process for users to revoke their consent to usage of their personal data is in place, and verify that there is a process and adequate methods to delete the user's data from datasets and the model.
Additionally, it shall be verified that no information can be extracted by employing privacy attacks (e.g., membership inference, prompt injection) customized to a specific individual.

Evaluation tools

Not applicable.

Reference to EU AI Act

- Recital 10

DQM-10

Data sensitivity and necessity filtering

Relevance based on use case parametrization

- Personal data was used to train the involved AI models (Q6) AND
- Traning of at least one integrated model was performed partially (e.g., with finetuning) or from scratch (Q5)

Evaluation requirement

Prior to model training, sensitive data that are not essential for the intended purpose of the system shall be removed or (pseudo-)anonymized to ensure compliance with data privacy regulations.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Sensitive data: Data that, if disclosed, could harm or infringe on privacy rights.
Anonymization: The process of irreversibly transforming data so that an individual can no loner be identified, directly or indirectly.
Pseudonymization: The process of replacing identifiable data with pseudonyms or identifiers so that individuals cannot be identified without additional information.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Examples for sensitive data are among others:
- Personal data such as names, addresses
- Health information
- Financial information (e.g., credit card number)

- Biometric data (e.g., fingerprints, facial recognition)
- Racial or ethnic origin, political opinions, religious beliefs

Evaluation method

Verify that a functioning process is in place to remove all unneeded sensitive information from the data.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 10 Data and data governance

DQM-11

Documentation of development data splits

Relevance based on use case parametrization

- Traning of at least one integrated model was performed partially (e.g., with finetuning) or from scratch (Q5)

Evaluation requirement

A method for data splitting of the development data into training, validation and test data shall be defined and documented.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Data splitting: Process of partitioning data in portions for model training, validation and testing.
Training set: Used for the model training.
Validation set: Used for hyperparameter tuning.
Test set: Used for (final) evaluation of the model.
Data split ratios: Proportions of data allocated to training, validation, and test sets.
Random splitting: Random division into tain, validation and test set.
Stratification splitting: Splitting data in such a way that data property distributions are equal accross train, validation and test set.
Cross-validation: Splitting the data into multiple subsets and training and testing on diverse combinations of these subsets.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

A suitable data splitting method, e.g., random splitting, stratification splitting or cross-validation, shall be chosen and documented. A reasoning for the chosen method and the data split (ratios) shall be given. Defined data requirements that are relevant to the development data (see DQM-03 & DQM-04) shall also hold for the subsets for training, validation and testing resulting from the splitting process. Under all circumstances, for an unbiased evaluation with the ability to detect potential overfitting and validate the system's ability to generalize, the set of test data shall be distinct from the used training data.

Evaluation method

Review the documentation of the chosen method for data splitting, review the given reasoning and validate that the defined method was used to create the respective datasets. Validate that training and test dataset are distinct from each other.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 10 Data and data governance
- 11 Technical Documentation
- 17 Quality Management System (QMS)
- 18 Documentation Keeping

DQM-12

Documentation of data preprocessing steps

Relevance based on use case parametrization

- Traning of at least one integrated model was performed partially (e.g., with finetuning) or from scratch (Q5)

Evaluation requirement

The data preparation process shall be documented.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Data Preparation: The process of cleaning, transforming, and organizing raw data into a format suitable for analysis, modeling, or machine learning. This step ensures that data is of high quality and ready for use.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Data preparation process should consider:
- Dataset composition;
- Data labelling;
- Data annotation;
- Data quality assessment relative to the data quality measure targets established in the data requirements process;
- Data quality improvement such as data cleaning; data standardization, data normalization, data imputation
- Data de-identification to protect privacy;

- Data encoding;
- Individual processing steps such as conversions, transformations, aggregations, normalization, format conversions, feature calculation, conversion of numerical data into categories;
- Filtering, e.g.: detection and processing values with different measurement scales;
- Documentation of assumptions made, in particular with regard to the information that the data should measure and represent;

Evaluation method

Verify that there is documentation describing the process of the (pre)processing of data. Ensure that it is adequately documented, see supporting guidance.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 10 Data and data governance
- 11 Technical Documentation
- 17 Quality Management System (QMS)
- 18 Documentation Keeping

DQM-13

Data provisioning

Relevance based on use case parametrization

- Traning of at least one integrated model was performed partially (e.g., with finetuning) or from scratch (Q5)

Evaluation requirement

The data provisioning process shall be documented.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Data provisioning: The process of making data available to users, applications, or systems in a secure, efficient, and controlled manner. This can involve extracting, preparing, and delivering data from various sources to its intended destination.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Considerations should be given to:
Data requirements should be met before provisioning.
Data provisioning includes:
- Ensuring data quality when transferring or moving the data from one system to another;
- Making data available for the model.

Evaluation method

Ensure that the data requirements were met. Verify that the data is available for the model and that the data quality is consistent when transferring data.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 10 Data and data governance
- 11 Technical Documentation
- 17 Quality Management System (QMS)
- 18 Documentation Keeping

DQM-14

Data characteristics

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

Relevant data characteristics shall be documented.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Data characteristics: Attributes or properties that describe the nature, structure, quality, and behavior of data. Documenting these characteristics helps ensure that data is well-understood, appropriately used, and suitable for its intended purpose.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Some factors to consider in the documentation:
- Its dynamics: e.g., static, dynamic with occasional updates, dynamic real-time updates;
- Its scale: e.g., very large (Exabyte), Large (tens of petabyte), Medium (hundreds of GB), Small (tens of GB or smaller);
- Its structure: e.g., unstructured, semi-structured, structured, complex structured data;
- Its format: e.g. CSV, JSON, XML etc.;
- Its privacy risks: e.g., identified data, pseudonymised data, unlinked pseudonymised data, anonymised data, aggregated data;
- Its sensitivity, e.g. protected, confidential, public;

- Its degree of standardization: standardized data format, non-standardized data format, standardized dataset metadate, non-standardized dataset;
- The feature's data types: e.g. Tabular static numerical data, Temporal (longitudinal) data, image data, text data etc.;
- Dependencies between features, especially in tabular data;
- Its metadata.

Evaluation method

Verify that data characteristics are documented in accordance to the supportive guidance.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 10 Data and data governance
- 11 Technical Documentation
- 13 Transparency and provision of information to deployers
- 17 Quality Management System (QMS)
- 18 Documentation keeping
- 26 Obligations of deployers of high-risk AI systems

Chapter 8

Development

Contents

DEV-01

Technical model documentation

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

A description of the AI system shall be provided including relevant technical conditions.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Not applicable.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Aspects to consider in the documentation:
- Initial description of the AI system including relevant technical conditions and circumstances;
- Objective and scope definition;
- Intended usage environment for the system;
- Expected output ranges of the system;
- Training approach (central or federated);
- Type of decision making (probabilistic, deterministic);
- System architecture;
- System requirements;
- Relevant developer choices;
- Input data requirements;
- Expected ranges of the output;

- Development and lifecycle process;
- Scale of current deployment: pilot, narrow, broad, widespread;
- Relevant dependencies of the system on other models not directly developed or data not directly used;
- The versions of relevant software or firmware, and any requirements related to version updates;
- Information about the model(s) used in the system;
- The description of the hardware on which the AI system is intended to run.

Evaluation method

Ensure that a description of the AI system is provided, including the aspects stated in the supportive guidance.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 10 Data and data governance
- 11 Technical Documentation
- 13 Transparency and provision of information to deployers
- 17 Quality Management System (QMS)
- 18 Documentation keeping

DEV-02

Documentation of pretrained model usage

Relevance based on use case parametrization

- The architecture is not or only partially known (Q2) OR
- There is only partial or zero knowledge of parameters of the involved AI models (e.g. usage of third-party model with API access providing model results and probability scores or other meta-information / providing model results only) (Q3)

Evaluation requirement

If a pretrained model is used, its usage shall be justified and documented.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Not applicable.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

The documentation should at least cover:
- Why pretraining is appropriate for the task,
- An evaluation of the Service Level Agreement (if applicable) focusing on Availability (Ensuring the systems are reliably accessible as per agreed-upon terms), Scalability (Confirming the infrastructure can handle increased loads and scale as needed), Performance (Assessing the efficiency and responsiveness of the systems under various conditions).
- An exit and continuity strategy outlining how critical functionality will be maintained if the pretrained-model provider experiences a prolonged outage or ends the service.

Evaluation method

The documentation must justify, or reference, why in general a pretrained model is chosen for the use case and a justification for the selected pretrained model shall be given. This reasoning shall include the evaluation of the Service Level Agreement as stated in the supportive guidance.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 10 Data and data governance
- 11 Technical Documentation
- 13 Transparency and provision of information to deployers
- 17 Quality Management System (QMS)
- 18 Documentation keeping

DEV-03

Model development policy

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

Model development policies and instructions shall be documented.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Not applicable.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

The development policies and instructions should at least cover:
- Requirements for training material (datasets, guides needed for system training);
- System robustness;
- Design;
- Development;
- Deployment;
- Verification (application tests, code review etc.);
- Validation;
- Relevant subsystems;
- Review and approval procedures by management.

Evaluation method

Verify completeness of the documentation for model development policies and instructions.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 11 Technical Documentation
- 13 Transparency and provision of information to deployers
- 17 Quality Management System (QMS)
- 18 Documentation keeping

DEV-04

Runtime environment specification

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

The necessary runtime environment shall be determined and documented regarding hardware and software, to ensure compatibility and performance.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Runtime Environment:
The combination of hardware and software configurations required to execute a system or application, including operating systems, frameworks, libraries, and physical devices essential for proper functionality and performance.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Aspects to consider regarding:
- Hardware: Screen size, screen resolution, storage, network connection;
- Software: Operating system, browser, run-time environments such as Java Run-time Environment or .NET.
Tool which could be used during implementation:
Hardware Detection & enumeration:
- HWInfo
- OS-tools (msinfo32, lspci)

SBOM generation tools:
- Syft
- SPDX
- CycloneDX
Code dependency Tracker:
- dependency-track

Evaluation method

Verify that the hardware and software runtime environment has been identified, as described in the supportive guidance and documented accordingly.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 11 Technical Documentation
- 13 Transparency and provision of information to deployers
- 17 Quality Management System (QMS)
- 18 Documentation keeping

DEV-05

Model logging and versioning

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

Experiments, associated metadata and SOUP (Software of Unknown Provenance, including libraries and frameworks) shall be tracked with appropriate measures for traceability and reproducibility.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Not applicable.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Considerations should be given to:
- Test and validation results shall be reproducible;
- Development shall be tracked by a version control system (model & dataset);
- A new version should be assigned to every trained model, using qualifiers to distinguish between production and pre-production models
- The development process should be tracked.
Tool which could be used during implementation:
SBOM generation tools:
- Syft
- SPDX

- CycloneDX

Code dependency Tracker:
- dependency-track

Evaluation method

Verify that experiments, associated metadata and SOUP are adequately tracked, for exemplary measures, see supportive guidance.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 11 Technical Documentation
- 13 Transparency and provision of information to deployers
- 17 Quality Management System (QMS)
- 18 Documentation keeping

DEV-06

Model selection process

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

Where appropriate, multiple model types shall be trained and compared and the final model choice shall be justified.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Multiple Models:
The practice of training and evaluating different machine learning algorithms or architectures on the same task to compare their performance, robustness, and suitability before selecting the final model.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Define metrics and compare e.g., regarding:
- Correctness;
- Robustness;
- Explainability.

Evaluation method

Verify that a set of possible model candidates was defined. The different models should be trained. Check that the final choice of model is justified by comparing, among other things, correctness, robustness and explainability.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 12 Record-keeping
- 19 Automatically generated logs

Chapter 9

Fairness

Contents

FAIR-01

Potential bias and impact assessment

Relevance based on use case parametrization

- Personal data was used to train the involved AI models (Q6) OR
- The model has a societal impact (Q10)

Evaluation requirement

Sensitive features shall be identified and potential biases against individuals or groups of individuals shall be assessed and documented.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

A selection of potential biases:
- Individual bias = The tendency of an AI system to make skewed decisions based on patterns learned from biased data related to individual preferences or characteristics, e.g. when a person with a university degree from a less known university applies to a job that is not recognized by an AI system.
- Group bias = The tendency of an AI system to favor or disadvantage certain groups based on patterns learned from biased data associated with group characteristics such as gender, ethnicity, or age.
- Automation bias = Occurs when a human decision-maker prefers recommendations made by an automated decision-making system, even when the system makes errors, over non-automated information.
- Group attribution bias = Occurs when a human assumes that what is true for one individual or object is also true for all individuals or objects in the group.
- Sampling bias = Occurs when data records are not collected randomly from the intended population.
- Coverage bias = Happens when the population represented in a dataset does not match the population that the machine learning model is making predictions about.
- Non-normality bias = Arises when statistical methods are applied to non-normally distributed data.
- Simpson's paradox = Manifests when a trend indicated in individual groups of data reverses

when the groups of data are combined.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

In order to identify potential biases, the context, purpose and relevant protected attributes for the use case should be considered.

Evaluation method

Review the documentation whether the user demographic is correctly identified.
Verify that all of the potential biases were considered during the bias analysis and that sensitive features were identified.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 9 Risk Management System
- 10 Data and data governance
- 11 Technical documentation
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)
- 18 Documentation keeping

FAIR-02

Effective bias assessment

Relevance based on use case parametrization

- Personal data was used to train the involved AI models (Q6) OR
- The model has a societal impact (Q10)

Evaluation requirement

The level of effective bias with respect to identified sensitive features in the data shall be assessed with [measures].
The selection of [measures] and reason for their selection shall be included in the system description.

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Sensitive features: features that were identified during the assessment in FAIR-01, e.g., gender, ethnicity, religion etc.

Exemplary [parameters]

[measures]: e.g., Absolute Standardized Mean Difference (ASMD), Adverse Impact Ratio

Additional guidance regarding the requirement

Not applicable.

Evaluation method

Assess whether sensitive features have been identified and are complete.
Check the system description and/or the source code whether appropriate measures have been identified and correctly used to measure and/or mitigate the potential bias or fairness problems.

Evaluation tools

- Aequitas - can be used as an audit tool to check if fairness reports match expected demographic group fairness assessments.
- IBM AI Fairness 360 can be used for independent bias analysis to cross-check whether the AI system's bias detection aligns with industry standards.

Reference to EU AI Act

- 9 Risk Management System
- 10 Data and data governance
- 11 Technical documentation
- 13 Transparency and provision of information to deployers
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality management system (QMS)
- 18 Documentation keeping

FAIR-03

Bias mitigation

Relevance based on use case parametrization

- Personal data was used to train the involved AI models (Q6) OR
- The model has a societal impact (Q10)

Evaluation requirement

If a critical level of bias is detected, the system shall test and implement a [mitigation measure]. Results should be compared using both [bias measures] and [performance measures]. Bias that is considered critical for functionality shall be documented in the system description.

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Critical level of bias:
A threshold or degree of bias that, if exceeded, poses a significant risk to fairness, functionality, or compliance with ethical and legal standards.

Exemplary [parameters]

[mitigation measure]: e.g., Post-Stratification, Covariate Balancing Propensity Score, Inverse Propensity Score Weighting.
[bias measure]: see FAIR-02.
[performance measures]: see MON-01.

Additional guidance regarding the requirement

Not applicable.

Evaluation method

The detection mechanism must be in place and it must trigger the mitigation strategy, once the threshold is reached. The mitigation must now adjust aspects of the system to balance the output.

Evaluation tools

- Fair Training methods, e.g. ”Learning Certified Individually Fair Representations” - can be used to verify whether fairness training methods work by evaluating formal fairness guarantees on specific test samples.
- balance python package - the implemented diagnostics and evaluation methods such as statistical summaries of weight and covariates distributions can be used to indentify potential biases.
- AWS Sagemaker Clarify - the implemented methods and reporting can be used for pre- and post-training data and model training Bias detection.

Reference to EU AI Act

- 9 Risk Management System
- 10 Data and data governance
- 11 Technical documentation
- 13 Transparency and provision of information to deployers
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)
- 18 Documentation keeping

FAIR-04

Accessibility

Relevance based on use case parametrization

- The model interacts directly with external consumers (Q7) OR
- The system integrates a human (either human-on-the-loop, human-in-the-loop or human control) (Q9)

Evaluation requirement

The AI system shall ensure equal accessibility and inclusivity for relevant user groups.

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

All user groups:
All people including marginalized groups, people with disabilities (e.g., accessible to screen readers, including alt text for images, colour-blind friendly palettes, etc.).
Accessibility:
The system should be operable by people with a wide range of impairments, including visual impairments, hearing impairments and physical impairments; it should support multiple languages.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Not applicable.

Evaluation method

Assess whether the potential user groups have been identified and whether any potential accessibility or inclusivity challenges have been recognized.
Evaluate whether the system's design ensures accessibility for all identified user groups, including those with disabilities and diverse language needs.

Evaluation tools

- WAVE - can be used to check if the system is compliant by evaluating screen reader compatibility, contrast ratios, and missing alt text.
- Google Lighthouse - can be used to assess accessibility performance using automated audits, focusing on usability, contrast, ARIA attributes, and keyboard navigation.

Reference to EU AI Act

- 9 Risk Management System
- 11 Technical Documentation
- 13 Transparency and provision of information to deployers
- 18 Documentation keeping

Chapter 10

Governance

Contents

GOV-01

Periodic review of the (internal) AI policy

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

An AI policy shall be established and reviewed periodically to ensure its continuing suitability, adequacy and effectiveness.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

AI policy:
A set of guidelines and principles designed to govern the development, deployment, and use of artificial intelligence systems.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

The goal of the AI policiy is to ensure ethical, legal, and operational compliance, aligning AI practices with organizational goals and societal standards. Regular reviews help maintain its relevance and effectiveness in addressing emerging risks, technologies, and regulatory changes.

Evaluation method

Verify that AI policies are established and communicated effectively.
Evaluate the process for periodic reviews and assess its effectiveness

Evaluation tools

Not applicable.

Reference to EU AI Act

- 9 Risk Management System
- 17 Quality Management System (QMS)

GOV-02

Establishment of AI management system and regular audits

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

A Quality Management System shall be established and audited (internal or external) on a regular basis.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Quality Management System (QMS):
A structured framework that is concerned with the implementation, monitoring, and governance of AI systems.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

The QMS shall include at least:
- The name and address of the provider and, if the application is lodged by an authorised representative, also their name and address;
- The list of AI systems covered under the same quality management system;
- The technical documentation for each AI system covered under the same quality management system
- A description of the procedures in place to ensure that the quality management system remains adequate and effective
The review of the QMS shall include at least:

- The changes, if any, since the last review;
- Changes in the AI system that are relevant to the AI management system;
- Changes in the needs and expectations of the AI system that are relevant to the AI management system;
- Information on the performance of the AI system.

Evaluation method

Verify whether an AI Management System or a QMS that includes the AI system is established. Check that the review process of this Q/MS is effective and appropriate.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 17 Quality Management System (QMS)

GOV-03

Attacks with GenAI

Relevance based on use case parametrization

- The system integrates at least one Generative AI Foundation Model (Q1)

Evaluation requirement

The potential risks associated with the use of General Purpose AI (GPAI) / Generative AI (GenAI) shall be assessed and incorporated into the design of incident response and defense strategies.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Generative AI (GenAI):
A type of artificial intelligence that creates new content, such as text, images, or audio, by learning patterns from existing data.
General Purpose AI (GPAI):
Artificial intelligence systems designed to perform a wide range of tasks across various domains. These systems are adaptable and can solve diverse problems using the same core model, making them versatile for multiple applications.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

It could be considered, e.g., how GenAI could be used for attacks on the business's customers or clients through spoofing or GenAI generated content etc.
The OWASP LLM AI Cybersecurity & Governance Checklist can be consulted for more information.

Evaluation method

Verify that the threat of GPAI/GenAI has been considered in the design of incidents and defence strategies.
Ensure that the considerations have been properly documented.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 9 Risk Management System
- 17 Quality Management System (QMS)

GOV-04

Compliance of the AI service with general cloud computing compliance criteria

Relevance based on use case parametrization

- The system is built/operated on a hybrid infrastructure (Q11)

Evaluation requirement

The system should be compliant with general cloud computing compliance criteria, as set out in the Cloud Computing Compliance Criteria Catalogue (C5)

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

The Cloud Computing Compliance Criteria Catalogue (C5) by BSI outlines minimum security requirements for cloud applications.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Complying with the C5 Catalogue enhances trust and credibility in your system by demonstrating adherence to high-security standards. It strengthens market competitiveness. Additionally, compliance facilitates alignment with international frameworks, such as ISO 27001 and SOC 2, simplifying global standardization efforts.
By adhering to the C5 requirements, organizations can effectively mitigate risks associated with cyberattacks, operational disruptions, and reputational damage. Conversely, non-compliance may indirectly lead to violations of GDPR or industry-specific regulations, resulting in significant financial penalties and legal consequences.

A Tool which could be used during implementation is clouditor, an assurance tool that checks whether cloud-based services and applications are securely configured

Evaluation method

Verify that the system is compliant to necessary standards and regulations.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 17 Quality Management System (QMS)

GOV-05

Impact on IP rights

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

All possibilities for potential impact on IP rights shall be considered and documented.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

IP (Intellectual property):
Concerns content that is subject to copyright, trademark, or patent protection.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Considerations for potontial impact on IP rights are ownership, licensing, copyright, patents, trademarks, and trade secrets throughout the lifecycle of the system.

Evaluation method

Verify that an analysis for possible IP right infringements has been performed.
Evaluate if appropriate measures against the possible infringements have been taken.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 9 Risk Management System
- 17 Quality Management System (QMS)

GOV-06

System description/User guidance

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

A system description shall be created and made available to the user.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

System description:
A detailed document ensuring transparency and ease of understanding.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

The system description should include e.g.,:
- System overview: brief introduction to the system, purpose and intended use cases, key features and functionalities
- Technical Architecture: Description of the used infrastructure, components of the system, technologies and frameworks used
- User interface description
- User instructions: instructions on how to use the system, examples of input data and expected outputs
- Troubleshooting and FAQs
- Data Management: information on data input formats and sources, data storage, security measures, and privacy considerations

- System limitations
- System evaluations: Levels of performance, Performance metrics used, Bias detection metrics used, detected Biases that are considered critical for functionality
- Compliance and ethical considerations

Evaluation method

Verify that a system description including relevant aspects regarding the system is in place and made available to the user.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 11 Technical Documentation
- 13 Transparency and provision of information to deployers
- 17 Quality Management System (QMS)
- 18 Documentation keeping

GOV-07

Qualitative description of system

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

The AI system and it's intended purpose shall be described including relevant qualitative aspects.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Intended purpose:
The specific goal or function that the system is designed to achieve. It defines why something was created or implemented and what outcomes are expected from its use.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

The AI system and its intended purpose are clearly documented to ensure a comprehensive understanding.
Relevant qualitative aspects to consider in the documentation:
- Intended function/purpose;
- User group;
- The user's existing skills;
- The user's needs;
- Needed user's competency;
- The user's potential interaction with the system;
- Business function;

- The system's impacts on critical functions and activities of the organization;
- Specific risks of harm that can have an impact on natural persons;
- Implemented human oversight measures;
- Arrangements for internal governance;
- Established complaint mechanisms.

Evaluation method

Verify that a complete and comprehensive description of the AI system is provided. Including its intended function/purpose, the user group, user's competences, etc.. See supportive guidance for more information.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 11 Technical Documentation
- 13 Transparency and provision of information to deployers
- 17 Quality Management System (QMS)
- 18 Documentation keeping

GOV-08

Alternatives to ML approach

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

The alternatives to AI shall be listed and evaluated regarding utility, security, and performance. The superiority of ML methods shall be justified while taking its associated risks into account.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Machine Learning (ML):
A type of AI where systems learn from data to make predictions or decisions without explicit programming. It includes methods like supervised, unsupervised, and reinforcement learning.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Alternatives to ML can, e.g., be rule-based systems, statistical methods, optimization algorithms, hardcoded algorithms, heuristic methods, simulation methods, human decision-making.
Considerations should be given at least to:
- Utility;
- Security;
- Performance.

Evaluation method

Verify that an analysis was performed to identify and compare possible alternatives to AI.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 9 Risk Management System
- 17 Quality Management System (QMS)

GOV-09

Catalogue of AI services and tools

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

An asset inventory including all internally developed, external or third-party solutions shall be conducted for the overall system.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Asset inventory:
A detailed list of all resources, components, or items owned or used within a system. It includes physical assets, software, data, and third-party solutions.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

The information provided for each module should include at least:
- Its core application area;
- Its use;
- Its appearance
in the overall system.
Tools which could be used for implementation:
SBOM generation tools:
- Syft
- SPDX

- CycloneDX
Code dependency Tracker:
- dependency-track

Evaluation method

Verify that the asset inventory contains all relevant information, see supportive guidance.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 11 Technical Documentation
- 13 Transparency and provision of information to deployers
- 17 Quality Management System (QMS)
- 18 Documentation keeping

GOV-10

Impact Analysis for ML model inventory

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

For each asset in the ML model inventory, an impact analysis shall be performed to determine the impact if the model is compromised.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Impact analysis:
An assessment of how changes or risks affect a system's functionality, security, and performance, helping identify risks and plan mitigations.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Possible impacts include:
- Data breaches;
- Model manipulation;
- Operational disruption;
- Financial loss

Evaluation method

Verify that each asset in the ML model inventory has been assessed with an impact analysis for the event of a compromised model.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 9 Risk Management System
- 17 Quality Management System (QMS)

GOV-11

Human resource competencies

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

Information on all persons involved in the life cycles of the AI system shall be documented. Their competencies shall be ensured and documented through the necessary qualifications and trainings.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Life cycle:
The process from design and development through deployment, operation, and eventual decommissioning.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Involved personnel includes, among others, the people who
- have participated in the various stages of development, deployment, operation, change management, maintenance, transfer and decommissioning, continuous improvement of the AI system, as well as verification and integration of the AI system;
- are involved in the decision-making process (e.g. in the form of human-in-the-loop, human-on-the-loop, human-in-command).

Evaluation method

The necessary (personnel) roles for the AI lifecycle are defined including their necessary qualification.
Assess whether the personnel has the necessary qualifications corresponding to their role, e.g. by reviewing certificates, project experiences or evidence for related trainings.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 4 AI literacy
- 17 Quality Management System (QMS)
- 26 Obligations of deployers of high-risk AI systems

GOV-12

Allocation of responsibilities among all involved parties

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

It shall be ensured that and documented how responsibilities within the life cycle of the AI system are allocated between the organization, its partners, suppliers, customers and third parties.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Life cycle:
The process from design and development through deployment, operation, and eventual decommissioning.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Defining roles and responsibilities is critical for ensuring accountability throughout the organization for its role with respect to the AI system throughout its life cycle.
Examples of areas that can require defined roles and responsibilities can include:
- Risk management;
- AI system impact assessments;
- Asset and resource management;
- Security;
- Safety;
- Privacy;

- Development;
- Performance;
- Human oversight;
- Supplier relationships;
- Demonstrate its ability to consistently fulfil legal requirements;
- Data quality management (during the whole life cycle).

Evaluation method

Verify that all involved stakeholders in the full AI lifecycle are identified.
Assess whether their responsibilities are clearly described and communicated.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 17 Quality Management System (QMS)
- 25 Responsibilities along the AI value chain

GOV-13

AI security awareness training

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

Educational initiatives and practical exercises to enhance cybersecurity skills and raise awareness about risks along the entire lifecycle and supply chain of the AI systems for all AI stakeholders shall be conducted.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Supply chain:
The network of entities and processes involved in creating, delivering, and maintaining a system.
Life cycle:
The process from design and development through deployment, operation, and eventual decommissioning.

Exemplary [parameters]

Not applicable-

Additional guidance regarding the requirement

Relevant parties should have been identified in GOV-12.
Educationak initiatives include, e.g., Workshops and training sessions (internal or external), certification programs and cybersecurity awareness campaigns.
In addition, practical exercises can be, e.g., phishing simulations, incident response drills, penetratin testing labs, data recovery scenarios and secure coding challanges.

Evaluation method

Verify that AI stakeholders are educated with initiatives and practical exercises to enhance cybersecurity skills throughout the AI lifecycle.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 4 AI literacy
- 26 Obligations of deployers of high-risk AI systems

Chapter 11

Human Oversight

Contents

HO-01

Human oversight

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

The system shall be monitored through human oversight.

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Human oversight:
The process in which humans closely monitor AI systems and intervene when necessary to ensure safe, ethical, and compliant operations.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Especially meaningful decisions that have a direct impact on people should include effective human oversight.
Human oversight can help with detecting anomalies beyond automation, responding in unforeseen scenarios, contextual decision-making, improving system performance and providing accountability.

Evaluation method

Verify that a process is in place for a human to monitor the system.

Evaluation tools

- Simul8 - can be used to simulate real-world AI system operations and assess the effectiveness of human oversight in responding to anomalies and unforeseen scenarios.

Reference to EU AI Act

- 13 Transparency and provision of information to deployers
- 14 Human oversight
- 26 Obligations of deployers of high-risk AI systems

HO-02

Modification of automated decisions

Relevance based on use case parametrization

- The AI system operates autonomously with no human oversight (Q9)

Evaluation requirement

In the case of automated decision-making, [procedures] shall be put in place to allow authorized personnel and/or users of the AI service to override the system.

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Override the system:
Update or modify decisions or completely stop the operation of the system.

Exemplary [parameters]

[Procedures]: e.g.,
- Clear Escalation Paths;
- Incident Reporting Systems;
- Manual Override Buttons;
- Fail-Safe Triggers

Additional guidance regarding the requirement

Ensuring authorized personnel can override AI systems is critical to mitigate risks and prevent harm caused by incorrect or unforeseen decisions. Overrides enable human judgment in complex or high-stakes scenarios, ensuring fairness, safety, and compliance with regulations like GDPR. They also build trust and transparency by providing a safeguard against biases, malfunctions, or failures in automated systems.

Evaluation method

Verify that a procedure is in place to override the system if necessary. Ensure that only authorized persons are able to perform the procedure.

Evaluation tools

- OWASP ZAP - can be used to test access control vulnerabilities and ensure that override mechanisms cannot be misused or bypassed by unauthorized users.
- Metasploit - can be used to perform penetration testing on manual override systems, escalation paths, and fail-safe triggers to identify security gaps.)
- Burp Suite - can be used to analyze and validate API security, ensuring that override commands function correctly and securely.
- Gremlin - chaos testing can be used to simulate system failures and validate whether override mechanisms function effectively under real-world stress conditions.

Reference to EU AI Act

- 13 Transparency and provision of information to deployers
- 14 Human oversight
- 26 Obligations of deployers of high-risk AI systems

HO-03

Monitoring of oversight and overriding processes

Relevance based on use case parametrization

- The AI system operates autonomously with no human oversight (Q9)

Evaluation requirement

There should be clear specification and documentation of individuals that are authorized to override decisions made by the system.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Override the system:
Update or modify decisions or completely stop the operation of the system.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Considerations should be given to:
- Potential biases that may result from this arrangement;
- Only authorized persons should have the ability to correct outputs based on their roles and responsibilities;
- Observation mechanisms to observe project team members who can oversee and override AI system decisions;
- Access removal controls should be in place to quickly revoke individual access.

Evaluation method

Verify that the documentation contains the specification about people who are authorized to override the decision made by the system.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 13 Transparency and provision of information to deployers
- 14 Human oversight
- 26 Obligations of deployers of high-risk AI systems

HO-04

Attribution of ethical and legal responsibility in AI system lifecycle

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

At all stages of the AI system's lifecycle, the physical persons or existing legal entities holding ethical and legal responsibility shall be clearly identified, listed and documented.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Life cycle:
The process from design and development through deployment, operation, and eventual decommissioning.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Specific individuals or entities bearing such responsibility within the project team shall be clearly identified.
A suitable way to document could be a RACI chart (who is responsible, who is accountable, who should be consulted, and who should be informed)

Evaluation method

Verify that the physical persons or legal entities are determined and documented throughout all stages of the AI cycle.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 17 Quality Management System (QMS)

Chapter 12

IT-Security

Contents

ITSEC-01

Common standards and frameworks for privacy impact

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

Common standards, frameworks and best practices for PIA (privacy impact analysis) should be considered.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Privacy Impact Analysis (PIA):
A systematic process to identify, assess, and mitigate risks to individuals' privacy posed by a system. It evaluates how personal data is collected, stored, processed, and shared, ensuring compliance with data protection laws and safeguarding individuals' rights

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Relevant standards (among others) are:
- ISO/IEC 29134
- ISO/IEC 27001
- NIST Privacy Framework
- General Data Protection Regulation (GDPR)

Relevant frameworks and practices (among others) could be:
- GDPR DPIA Guidance
- OECD Guidelines on Privacy

Evaluation method

Verify that a PIA is carried out taking into account common standards, frameworks and best practices, see supportive guidance for relevant standards.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 10 Data and data governance
- 17 Quality Management System (QMS)
- 26 Obligations of deployers of high-risk AI systems

ITSEC-02

Supply chain security

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

The supply chain involved in the development shall be analyzed for potential security weaknesses, ensuring all components and dependencies are secure.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Supply chain:
The network of entities and processes involved in creating, delivering, and maintaining a system.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

This should also include a verification process for all SOUP (Software of Unknown Provenance) or OTS (Off-The-Shelf) components.
Tools which could be used during implementation:
- Snyk
- OWASP Dependency-Check
- Black Duck
- WhiteSource

Evaluation method

Verify that the supply chain is analyzed for potential security weaknesses. Ensure that the analysis includes all SOUP or OTS components.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 9 Risk Management System
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)

ITSEC-03

Risks from unspecified usage environments and users

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

The risks arising from unspecified usage environments and non-specified users shall be analyzed.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Unspecified Usage Environments:
Physical or operational settings that are not explicitly defined.
Non-Specified Users:
Individuals who are either untrained or lack proper authorization to access the system.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Potential risks are:
- Misuse of the system;
- Security risks;
- Safety concerns;
- Regulatory non-compliance;
- Bias or discrimination

Evaluation method

Verify that the different types of risk that could arise from unspecified usage environments are analysed.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 9 Risk Management System
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)

ITSEC-04

Secure GenAI-App integration

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

Connections with existing systems and databases shall be safeguarded with secure integrations at all AI system interfaces.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

System interfaces:
Points of interaction where a system communicates with other systems, applications, or users.
Secure Integrations:
The process of safely connecting systems or components to ensure data integrity, confidentiality, and functionality.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

System interfaces can be:
- LLM trust boundaries
- User Interface
- Hardware Interfaces
- Webhooks
- Input/Output Interfaces

Secure integration could be:

- Encryptied data transfer
- API security
- Firewalls and network segmentation
- Endpoint security

Evaluation method

Verify that secure integrations are in place to protect connections between the AI system and other existing systems or databases.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 9 Risk Management System
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)

ITSEC-05

Conventional network security protection

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

The network shall be protected by conventional IT security [measures].

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Conventional IT security:
Traditional methods and practices for protecting systems, data, and infrastructure from unauthorized access, misuse, or harm.

Exemplary [parameters]

[measures]: e.g., Firewalls, virtual private networks, network segmentation, access control.

Additional guidance regarding the requirement

A protected network is important to maintain data confidentioality and garnatee operational continuity along with cost reduction in the sense of mitigation financial loss due to e.g., data breachers, ransomware or downtime.
Possible attacks and threats include e.g. Malware, Phishing attacks, ransomware, (distributed) denial of service, unauthorized acess.
Tools that can be used for the implementation:
- Intrusion Detection System (IDS)
- Firewall
- Network Access Control (NAC)
- Security Information and Event Management (SIEM)

Evaluation method

Verify that the network is protected by classical IT security measures, e.g., via access control, maleware detection mechanisms, security incident management, etc.

Evaluation tools

- nmap - can be used to scan for open ports, detect misconfigurations, and identify potential attack vectors in the network.
- Nessus - can be used for vulnerability scanning to detect security gaps, outdated software, and misconfigured security controls.
- Wireshark - can be used to analyze network traffic and detect anomalies that could indicate security threats such as unauthorized access or malware communication.

Reference to EU AI Act

- 9 Risk Management System
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)

ITSEC-06

Gateway controls

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

[Mechanisms] shall be sufficiently implemented and documented to enable a graduated set of controls.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Graduated set of controls:
Security measures that were implemented in layers or tiers.

Exemplary [parameters]

[mechanisms]: e.g. Secure gateway, VPN, and routing for machine learning systems.

Additional guidance regarding the requirement

Possible levels are:
- risk levels
- sensitivity of the system or data
- criticality of the system or data

Evaluation method

Verify that the security mechanisms specified in the supportive guidance are adequately implemented and documented.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 11 Technical Documentation
- 13 Transparency and provision of information to deployers
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)
- 18 Documentation keeping
- 26 Obligations of deployers of high-risk AI systems

ITSEC-07

Availability and disaster recovery

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

The AI system shall comply with the defined availability standards for its IT infrastructure, applications and data. It should also have suitable [emergency plans] in place to restore the AI system and the model in the event of a failure.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Event of a failure:
An incident where a system, component, or process does not perform as intended, leading to disrupted functionality, reduced performance, or complete inoperability. Common types of failures include hardware failures, software failures, security failures, network failures or data failures. Emergency plans shall explicitly cover scenarios in which externally hosted or third-party AI model providers become unavailable or terminate their service. Contingency measures may include hot-swap to an alternative provider, activation of a locally hosted fallback model.

Exemplary [parameters]

[emergency plans]: e.g. Backup plans.

Additional guidance regarding the requirement

Relevant standards (among others) could be:
- ISO/IEC 27001
- NIST Cybersecurity Framework
- CIS Controls
- Risk Management Framework

Evaluation method

Verify that the AI system meets the defined availability standards for its IT infrastructure. Ensure that a process is in place to restore the AI system and model in the event of a failure, such as a backup plan.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 9 Risk Management System
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)

ITSEC-08

System availability assessment

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

The availability of the AI system shall be constantly evaluated with [suitable methods] and documented.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Availability:
Ability of a system, to remain accessible and operational for authorized users whenever needed.

Exemplary [parameters]

[suitable methods]: Uptime monitoring, latency monitoring, throughput monitoring, error rate monitoring, stress testing, scalability checks, maintenance scheduling, system health checks.

Additional guidance regarding the requirement

A high availability rate will improve business reputation, increase user satisfaction and build competitive advantage. Downtime of the system can lead to disrupt workflows, delay processes, and halt critical services, which can directly or indirectly lead to financial losses. A high availability rate is also required to be compliant with many legal and regulatory standards.
Tools which could be used for the implementation:
Uptime Monitoring Tools to check logs, historical uptime data, and alerts to ensure that monitoring is in place and functioning correctly: Nagios
Chaos Engineering Tools to inject faults into the system to test the suitable methods: Gremlin, Metasploit

Evaluation method

Verify that a procedure is in place to evaluate the availability of the system. This can include monitoring tools that monitor system properties related to the relevant business assets related to the system. In order to test the procedure, a stress test, causing the system to be unavailable shall be conducted and the reaction shall be evaluated.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 9 Risk Management System
- 11 Technical documentation
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)
- 18 Documentation keeping

ITSEC-09

Infrastructure security evaluation

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

Resilience testing shall be performed with [suitable techniques] to ensure that systems can withstand and recover from disruptions.

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Disruptions:
For example server crashes, network outages, high traffic or load, hardware failures, power failures etc.

Exemplary [parameters]

[suitable techniques]: e.g., Fault injection, failover testing, load and stress testing.

Additional guidance regarding the requirement

A measurement for resilience can be Performance.

Evaluation method

Verify that resilience testing is carried out. Check that appropriate techniques, see supportive guidance, are implemented correctly by ensuring if the system can withstand and recover from disruptions.

Evaluation tools

- Chaos Monkey - can be used to randomly terminate system instances and evaluate the system's ability to handle unexpected failures.
- Gremlin - can be used to conduct controlled fault injection and assess system resilience under various failure scenarios.
- Failure Injection Testing - can be used to simulate infrastructure-level failures and analyze system response and recovery strategies.

Reference to EU AI Act

- 9 Risk Management System
- 10 Data and data governance
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)

ITSEC-10

Data classification and protection based on sensitivity

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

Data must be classified and protected based on its sensitivity. User permissions must be managed effectively, and appropriate safeguards must be in place to ensure that access controls and defense-in-depth measures are robust.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Sensitive data:
Information that requires extra protection due to its potential to cause harm if accessed or disclosed without authorization, e.g. presonal or biometric data.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Appropriate safeguards can be:
- Role-based access crontrol (RBAC)
- Session management
- Encryption
- Physical Security

Evaluation method

Verify if appropriate measures for the classification of sensitive data are in place. Furthermore, verify that protective measure to safeguard the sensitive data are in place and test their functionality.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 10 Data and data governance
- 15 Accuracy, Robustness, Resilience, Cybersecurity

ITSEC-11

Ensuring data authenticity

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

The system shall ensure the authenticity of data.

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Authenticity of data:
The quality of being genuine, unaltered/altered with traces, and verifiably from a trusted source.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Data authenticity can be secured through the use of public key infrastructure (PKI) and digital certificates.

Evaluation method

Verify that every method to add or manipulate data can only be used with sufficient authentication of the user.
Also ensure that data is genuine, unaltered or traceably altered, and originates from a trusted source

Evaluation tools

- Wireshark - can be used to analyze network traffic for verifying data integrity, checking if data exchanges include valid digital signatures and certificates.
- PKI Compliance Audit Tools - compliance tools like X.509 certificate validators can be used to check if the system correctly implements PKI for authenticity verification.
- Scapy - can be used to craft and send inauthentic or tampered data packets, testing whether the system correctly detects and rejects altered data.

Reference to EU AI Act

- 10 Data and data governance
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality management system (QMS)

ITSEC-12

Authorization and access control

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

An access control concept for the AI System should be created, defining roles, authorizations and responsibilities along the MLOps process and its feedback loops by logging and auditing all activities.

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Access control concept:
A framework that defines how access to resources, data, or systems is managed and restricted based on user roles, permissions, and security policies.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Access control mechanisms can include least privilege access controls or defence-in-depth measures.

Evaluation method

Verify that an access control concept is in place. The AI system must withstand unauthorised access attempts using simulated attack scenarios. Ensure that least-privilege access controls and defence-in-depth measures effectively restrict unauthorised access by attempting to access resources and data with different privilege levels.

Evaluation tools

- OpenSCAP - can be used to audit and assess compliance with access control policies, ensuring that security configurations follow defined standards.
- Metasploit Framework - can be used to simulate unauthorized access attempts and assess how well the system enforces access restrictions.
- BloodHound - can be used to analyze permission structures and identify privilege escalation paths in the AI system's access control.

Reference to EU AI Act

- 12 Record-keeping
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)
- 19 Automatically generated logs

Chapter 13

Monitoring

Contents

MON-01

Performance development & operation

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

The performance of the system shall be continuously monitored and documented with [Key Performance Indicators]. Deviations from defined performance requirements shall be made transparent.

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Performance:
Indicator that shows how effectively and efficiently an AI system accomplishes its tasks.
KPIs (Key Performance Indicators):
Metrics to evaluate the performance.

Exemplary [parameters]

[key performance indicators]:, e.g. F1 Score, Mean Squared Error (MSE), response times, resource consumption.

Additional guidance regarding the requirement

The system shall demonstrate the same KPIs in development respectively testing and validation and after the deployment in its operating state.

Evaluation method

Verify that appropriate measure to continuously monitor and document the system performance are in place (exeplary tools are given in the supportive guidance). Verify the monitored variables and the defined KPIs and test the mechanims to detect (and document) performance deviations from the defined performance requirements.

Evaluation tools

- Google Lighthouse Performance Auditor - can be used to verify if system response times meet defined KPI thresholds and whether Performance degrades under load.
- Apache JMeter - can be used to validate if system Performance remains consistent across different loads and if deviations from expected KPIs are detected.
- ML Benchmark Tools - benchmarking Tools can be used to compare AI model Performance against predefined KPIs and industry standards.

Reference to EU AI Act

- 11 Technical documentation
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)
- 18 Documentation keeping
- 72 Post-market monitoring by providers for high-risk AI systems

MON-02

Self-error detection

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

The system should have [mechanisms] to allow self-detection of errors and fail-safe methods to mitigate entire system failures.

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Self-error detection:
Criteria and mechanisms for automatic error detection.

Exemplary [parameters]

[mechanisms]: e.g., Internal consistency checks, anomaly detection algorithms.

Additional guidance regarding the requirement

Entire system failures are where the AI system becomes completely non-functional or unable to perform its intended tasks due to Hardware Failures, Software Failures, Security Failures, Network Failures or Data Failures

Evaluation method

Verify if the AI system has mechanisms for self-detection of errors and fail-safe methods, e.g. by actively triggering fault conditions, analyzing the system's response, testing for internal consistency or anomalies.

Evaluation tools

- Chaos Monkey - controlled failures can be injected into the system, such as network disruptions or service crashes, to verify whether self-error detection mechanisms identify and mitigate the issue.
- Apache JMeter - can be used to validate if system performance remains consistent across different loads.
- Failure Injection Testing - can be used to simulate infrastructure-level failures and analyze system response and recovery strategies.

Reference to EU AI Act

- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)

MON-03

Logging

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

The AI system shall have consistent logging across all components. Policies and instructions with technical and organizational safeguards for the logging process shall be documented.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Consistent logging:
Systematic recording of events, actions, and decision made by the system.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

The log files should contain at least:
- Type of request ;
- Processing times including time stamps and metadata on the user requesting the AI service.
The log files should be kept for an appropriate amount of time (suggestion: at least 3 months) (depending on the system and the amount of data to be logged).
Contents to log shall be specified: e.g., Input, model (state), output, timestamps, system access.
Tools which could be used for implementation:
- Log Analysis and Monitoring Tools to check if the logs include required information and verify that logging is consistent: Loggly
- Log File Integrity Tools to verify that logs are secure, consistent, and meet the technical and organizational safeguard: Tripwire

Evaluation method

Verify that appropriate monitoring and logging capabilities (tools, storage, etc.) are implemented on a technical and organizational level. Evaluate their effectiveness by targeted test, e.g., by inducing certain events that shall then be verified via the resulting system logs.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 11 Technical documentation
- 12 Record-keeping
- 17 Quality Management System (QMS)
- 18 Documentation keeping
- 19 Automatically generated logs

MON-04

Incident response procedures

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

Formal incident response procedures shall be implemented to report AI systems incidents throughout the whole lifecycle. The procedures should be tested on a periodic basis, tracking with [Key Performance Indicators].

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Formal incident response procedures:
Structured guidelines and processes for identifying, managing, and mitigating security incidents or breaches.
KPIs (Key Performance Indicators):
Metrics to evaluate the performance.

Exemplary [parameters]

[KPIs]: e.g. F1 Score, Mean Squared Error (MSE)

Additional guidance regarding the requirement

AI system incidents could be for example loss of service, loss of equipment, loss of facilities, system malfunctions, system overloads, human errors, and non-compliances with policies or guidelines, breaches of physical security, uncontrolled system changes, software malfunctions, hardware malfunctions, and access violations.

Evaluation method

Verify that system states that qualify as incidents are defined. Verify that appropriate formal incident response procedures are implemented and documented.
Assess whether the response procedures are tested before deployment and an appropriate periodic plan of repeated testing is established.Verify the incident response procedures by simulating incidents, analyzing response effectiveness, and measuring key performance indicators.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 9 Risk Management System
- 12 Record-keeping
- 17 Quality Management System (QMS)
- 19 Automatically generated logs
- 26 Obligations of deployers of high-risk AI systems
- 72 Post-market monitoring by providers for high-risk AI systems
- 73 Reporting of serious incidents

Chapter 14

Performance

Contents

PERF-01

Definition of performance requirements

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

The performance requirements shall be defined and included in the system description with at least the [Key Performance Indicators] used.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Performance Requirements:
Criteria that define the expected level of functionality and efficiency of the system, often measured through specific metrics like accuracy, speed, reliability, or resource usage, to ensure it meets its intended goals.
KPIs (Key Performance Indicators):
Metrics to evaluate the performance.

Exemplary [parameters]

[key performance indicators] (KPIs): type of metric to evaluate the performance, e.g. F1 Score, Mean Squared Error (MSE).

Additional guidance regarding the requirement

Not applicable.

Evaluation method

Verify that the key performance indicators, see supportive guidance, are defined and included in the system description.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 11 Technical documentation
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)
- 18 Documentation keeping
- 72 Post-market monitoring by providers for high-risk AI systems

PERF-02

Generalization performance verification via XAI

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

The ability of the AI system to generalize shall be verified by appropriate [XAI methods].

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

XAI = Explainable AI
Generalize:
The ability of an AI system to perform well on new, unseen data. This demonstrates adaptability beyond examples used in training datasets.

Exemplary [parameters]

[XAI methods]: e.g., Post-hoc explanations, model explanations or data explanations can help to understand the AI system and its performance regarding generalization. For bigger systems, a partitioning into subsystems that are explained individually is possible.

Additional guidance regarding the requirement

Insufficient generalizability can, e.g., lead to model collapse in LLMs.

Evaluation method

Verify that an appropriate XAI method was implemented to ensure that the AI system effectively generalizes. This can be done by testing its ability to handle unseen data distributions through adversarial and explainability-driven evaluations.

Evaluation tools

- Foolbox - can be used to perform adversarial robustness testing by applying small, realistic perturbations to input data and verifying if the AI model maintains performance or collapses due to lack of generalization.
- Robustness Gym - can be used to test the AI system under different real-world data perturbations, including paraphrased text, altered image resolutions, and noise injections. Evaluate whether accuracy remains stable across these transformations.
Model Testing Frameworks with XAI Capabilities - can be used to verify that post-hoc explanations are correctly implemented and provide meaningful insights.
- Captum - can be used to perform Integrated Gradients or Layer Conductance analysis on different test cases. If feature attributions shift erratically across near-identical inputs, this suggests poor generalization capability.

Reference to EU AI Act

- 11 Technical Documentation
- 13 Transparency and provision of information to deployers
- 18 Documentation keeping

PERF-03

Measures against overfitting

Relevance based on use case parametrization

- Traning of at least one integrated model was performed partially (e.g., with finetuning) or from scratch (Q5)

Evaluation requirement

[Measures] must be implemented to minimize overfitting and underfitting, ensuring that the model is adequately trained and performs effectively for its intended purpose.

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Overfitting:
A situation where a model learns the training data too well, including noise and specific details, resulting in poor performance on new, unseen data due to lack of generalization.
Underfitting:
A situation where a model fails to capture the underlying patterns in the training data, leading to poor performance both on the training data and unseen data.

Exemplary [parameters]

[Measures]: e.g., Cross-validation, regularization.

Additional guidance regarding the requirement

Not applicable.

Evaluation method

Verify, through code review and/or performance testing, that measures to avoid or reduce overfitting, underfitting and to ensure performance are implemented. Also check model's generalization ability.

Evaluation tools

- SHAP (SHapley Additive exPlanations) - can be used to analyze feature importance and detect cases where the model overly depends on specific training set patterns, indicating overfitting.
- Adversarial Robustness Toolbox (ART) - can be used to introduce slight perturbations in input data and observe the model's response, checking if small changes cause drastic misclassifications, which would indicate poor generalization.
- DeepCheck - can be used to evaluate model stability by applying distribution shift tests, ensuring that it maintains expected performance on slightly altered data distributions.

Reference to EU AI Act

- 10 Data and data governance
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)

PERF-04

Periodic retraining and model changes

Relevance based on use case parametrization

- Traning of at least one integrated model was performed partially (e.g., with finetuning) or from scratch (Q5)

Evaluation requirement

It shall be ensured that continuous improvement of the model's performance is achieved through [retraining methods].

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Not applicable.

Exemplary [parameters]

[retraining methods]: e.g., Periodic retraining, adjustment of the conceptual model or on-demand retraining.

Additional guidance regarding the requirement

Retraining should in all cases incorporate new ('unseen') data. Retraining is relevant especially when there are model or concept drifts. If issues persist after retraining, experts shall reassess the model, evaluate risks, and document all changes.
Tools which could be used for implementation:
- DeepChecks to perform side-by-side comparisons of pre- and post-retraining model performance, detecting anomalies, biases, and statistical shifts in feature behavior
Data Drift and Model Drift Monitoring Tools to check if triggered when necessary:
- AI Explainability 360 (AIX360) to assess explainability differences before and after retraining,

verifying that retraining improves model decisions rather than introducing hidden biases or reduced interpretability.
- Review of Azure Event Grid Triggers
- Review of CI/CD Flows

Evaluation method

Verify that a retraining method, see supportive guidance, can be performed. Ensure that triggering events, such as model drifts, are detected to appropriately trigger the retraining method and logged accordingly.
Identify if the retraining method mitigated the risk respectively improved performance, e.g., by using the tools stated in the supportive guidance and comparing the retrained version with the former system. Otherwise a procedure should be in place for a domain expert to reevaluate the risk exposure of the model and the model concept. Confirm that all adjustments and model changes are documented.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)

PERF-05

Testing

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

Performance testing of the AI System shall be performed and documented.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Not applicable.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Testing should include, for example:
- Stress testing;
- Performance testing on performance KPIs with inputs from the operational design domain of the system;
- Performance testing against expected input perturbations/deviations;
- Performance testing on Out-of-Distribution data;
- Test scenarios/test cases for, e.g. interpretable model decisions.

Tools which could be used for implementation:
- tensorboard

Evaluation method

Verify that suitable performance tests were performed and the execution and the corresponding results are documented in a dedictated documentation. The documentation shall comprehensive, so that reproducibility of the performed tests is given.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 9 Risk Management System
- 11 Technical documentation
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)
- 18 Documentation keeping

PERF-06

Assessment of uncertainty

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

Uncertainty assessment of the AI system shall be performed with [measures] and documented.

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Uncertainty Assessment:
The process of evaluating and quantifying the confidence or reliability of an AI system's predictions or outputs, often by analyzing how variations in inputs or conditions impact its performance and decision-making.

Exemplary [parameters]

[measures]: e.g., Sensitivity analysis.

Additional guidance regarding the requirement

E.g., how do changes in individual features affect predictions, particularly for tabular data?
What impact does user-provided data have on future changes to the AI service?
If there are contractually defined performance requirements, deviations from these shall be made transparent to the user.

Evaluation method

Verify that an uncertainty assessment was conducted and documented and that it was performed correctly, e.g., by testing whether variations in input features lead to unexpected changes (in the sense of not considered during uncertainty assessment) in the system's predictions.

Evaluation tools

- InterpretML - can be used to conduct sensitivity analysis and test how variations in individual input features affect the model's predictions. Verify if the uncertainty quantification aligns with the documented assessment.
- Deep Ensemble Tester - can be used to verify if ensemble-based uncertainty quantification produces stable confidence intervals and meaningful variance across multiple perturbed runs.

Reference to EU AI Act

- 11 Technical documentation
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)
- 18 Documentation keeping

PERF-07

Error handling

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

A procedure shall be in place to detect failures of the system. In case of a detected failure, the system shall answer with [error fallback functions].

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Procedures (exemplary):
Monitoring and Logging, Self-Checking Mechanisms, Error Detection Algorithms, Threshold-Based Alerts, Redundancy Checks, User Feedback Loops

Exemplary [parameters]

[error fallback functions]: e.g., Error correction, determination of the erroneous state of the system or deactivation of the system and notification of the bodies and parties concerned.

Additional guidance regarding the requirement

Not applicable.

Evaluation method

Verify that the system can successfully detect failures. Ensure that a fallback function, see supportive guidance, is in place in the Event of a failure.

Evaluation tools

- Chaos Monkey - can be used to randomly terminate system components or induce controlled failures to evaluate whether the system detects faults and responds correctly with error fallback mechanisms.
- Metasploit - perform controlled penetration testing to determine if failures lead to unintended system behavior, ensuring fallback mechanisms handle faults securely.
- Selenium - can be used to automate test scripts that simulate erroneous inputs, UI malfunctions, or system crashes, verifying whether failures are detected and appropriate fallback actions are triggered.
- Gremlin - can be use to inject controlled faults into the system, such as CPU spikes, memory leaks, or network failures, to test whether the system identifies these disruptions and executes error fallback functions accordingly.

Reference to EU AI Act

- 9 Risk Management System
- 11 Technical Documentation
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)
- 18 Documentation keeping
- 72 Post-market monitoring by providers for high-risk AI systems

Chapter 15

Transparency

Contents

TR-01

Assessment of the required degree of explainability

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

The need and required degree for explanation of the system shall be assessed and documented.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Explanation of the system: An explanation of a system refers to the information provided to make the system's decisions, outputs, or behavior understandable to different stakeholders. This can include describing how, why, and under what conditions a decision or outcome was made.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

During the assessment, considerations should be given to:
- Purpose of the system
- Affected stakeholders
- Potential damages resulting from the system
- Needs and prerequisites for human decision making
- Adequate handling of outliers
- Organization's obligations to reporting information to interested parties and regulators
- The necessary information to users
- Trade-off between model interpretability and performance

- Applicable legal and regulatory requirements and international standards that require the explainability of actions of the AI service
- Justified interest by users, which requires the implementation of methods to improve explainability

Evaluation method

Verify that the degree of explanation for the system has been assessed. Ensure that the documentation contains the assessed degree.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 11 Technical Documentation
- 13 Transparency and provision of information to deployers
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)
- 18 Documentation keeping
- 86 Right to explanation of individual decision-making

TR-02

Unique identification of AI actions

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

Actions carried out by the AI shall be uniquely identifiable to ensure authenticity.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Action:
Any action or event initiated or completed by the AI system that results in a change of state, data exchange, or processing outcome.
Uniquely Identifiable:
Each transaction has a distinct identifier (e.g., a transaction ID or hash) that makes it distinguishable from all other transactions.
Authenticity:
The assurance that the transaction is genuine, unaltered, and can be verified as having been carried out by the AI system as intended, without unauthorized modifications or tampering.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Unique identifiers can for example be established with transactions ids, timestamps, digital signatures, user or system metadata involved in the transaction.

Evaluation method

Verify that each AI Action is assigned a unique identifier.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 12 Record keeping
- 17 Quality management system (QMS)
- 19 Automatically generated logs

TR-03

Plausibility check

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

The output of the AI system shall be plausible for [inputs].

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Plausible: The AI system's output should be credible, reasonable, and appropriate given the context of the inputs.

Exemplary [parameters]

[inputs]: e.g., corner cases, boundary values, OOD input, expected Inputs, regular input.

Additional guidance regarding the requirement

The model's decision on failed tests should be explained.

Evaluation method

Verify, using appropriate explanation methods, that the output is reasonable across a variety of input cases. Refer to the provided supportive guidance for further clarification.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 13 Transparency and provision of information to deployers
- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 17 Quality Management System (QMS)

TR-04

Protect user from overreliance

Relevance based on use case parametrization

- The model interacts directly with external consumers (Q7) OR
- The AI system operates autonomously with no human oversight (Q9)

Evaluation requirement

Measures shall be in place to protect the users from overreliance of the AI system.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Overreliance:
The tendency of users to depend excessively on the AI system's outputs or recommendations, potentially leading to reduced human judgment, critical thinking, or oversight, even in cases where the AI might produce incorrect or incomplete results.

Exemplary [parameters]

[measures]: e.g., informing the user about the performance and robustness of the system, integration of checks and questions to ensure that the user questions the AI systems output before proceeding.

Additional guidance regarding the requirement

Not applicable.

Evaluation method

Verify that measures are in place to protect the users from overly reliying on the AI models output. The needed measures depend on the degree of autononmy of the system, human oversight measure put in place and the overall risk of the system.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 15 Accuracy, Robustness, Resilience, Cybersecurity
- 14 Human oversight
- 17 Quality Management System (QMS)

TR-05

User awareness of AI interaction

Relevance based on use case parametrization

- The model interacts directly with external consumers (Q7) OR
- The AI system operates autonomously with no human oversight (Q9)

Evaluation requirement

Users shall be made fully aware when they are interacting with the AI system and not a human.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

User Awareness:
Ensuring users clearly understand that they are interacting with an AI system rather than a human.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

To ensure user awareness, it would be suitable:
- to explicitly inform the user at the beginning of an interaction,
- use explicit labels for the system,
- maintain visible indicators throughout the interaction.

Evaluation method

Confirm that users are adequately informed that they are interacting with an AI system.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 50 Transparency obligations for providers and deployers of certain AI systems

TR-06

Marking of AI content

Relevance based on use case parametrization

- The system integrates at least one Generative AI Foundation Model (Q1) OR
- The model interacts directly with external consumers (Q7) OR
- The AI system operates autonomously with no human oversight (Q9)

Evaluation requirement

AI-generated or AI-modified content shall be appropriately marked as AI content using [techniques].

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

AI-Generated Content:
Content entirely created by an AI system without direct human creation or authorship.
AI-Modified Content:
Content that originates from human input but has been altered, enhanced, or adapted by an AI system.

Exemplary [parameters]

[techniques]: e.g., DNN (Deep Neural Network) Watermarking, Spread-Spectrum Watermarking, Perturbation-Based Watermarking, Adversarial Watermarking.

Additional guidance regarding the requirement

Not applicable.

Evaluation method

Review the implementation of techniques as referenced in the supportive guidance, e.g. using huggingface, and assess whether they are appropriate. For example, verify if AI-generated or AI-modified content is appropriately marked as AI content by attempting to remove, alter, or evade to test the robustness of the watermarking method with the proposed evaluation tool.

Evaluation tools

- Watermark-Robustness-Toolbox - can be used to evaluate the resilience of watermarking techniques by simulating various attacks to determine if AI-generated labels remain intact.

Reference to EU AI Act

- 50 Transparency obligations for providers and deployers of certain AI systems

TR-07

Provision of tailored explanations and transparency

Relevance based on use case parametrization

- The model interacts directly with external consumers (Q7) OR
- The model must explain reasoning for specific inference results (Q8) OR
- The AI system operates autonomously with no human oversight (Q9)

Evaluation requirement

A tailored explanation of why a particular output has been produced by the AI system shall be provided to the users with [explanation method].

Evaluation principle

Test-based

Supportive guidance

Supplementary Definitions

Tailored Explanation:
A specific, context-aware description provided to users, detailing the reasoning, factors, or data that led the AI system to produce a particular output or decision. The explanation is adapted to the user's needs, expertise, and the complexity of the output

Exemplary [parameters]

[explanation method]: e.g., Transparency of the data, the AI algorithm, the inference (if applicable); explainability of the AI system; the model's decision at the boundary values; the model's decision on corner cases; feature importance; consider local and global model behaviour if necessary.

Additional guidance regarding the requirement

Not applicable.

Evaluation method

Verify that the explanation methods, see supportive guidance, have been properly implemented, e.g. using Shap, and that these explanations are readily accessible to users.
This can be done by systematically stress-testing the systmens response consistency while checking if explanations remain logically consistent when inputs are slightly altered

Evaluation tools

- Counterfactual Testing - can be used to assess if explanations logically shift when similar but distinct cases are presented, ensuring the system maintains a valid rationale for different scenarios.

Reference to EU AI Act

- 13 Transparency and provision of information to deployers
- 86 Right to explanation of individual decision-making

TR-08

Provision of suitable AI system documentation for relevant parties

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

The necessary technical documentation of the AI system shall be determined for each relevant interested party and provide them with the appropriate documentation in a suitable form.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Interested Parties:
Individuals or entities who have a vested interest in the AI system and require documentation to fulfill their roles or responsibilities. Examples include:
- Developers and engineers
- End users
- Compliance officers and regulators
- Business stakeholders
- Supervisory authorities
- Partners

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Aspects to consider in the documentation/manual:
- Product version = Specific version number or identifier
- Quality measures = Metrics such as accuracy, precision, recall
- User understanding = Comprehension level, ease of use

Evaluation method

Verify that the documentation includes relevant and necessary information in accordance with the supportive guidance, and ensure that it is accessible for the user.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 11 Technical documentation
- 13 Transparency and provision of information to deployers
- 18 Documentation keeping

TR-09

Claims submission and appeal process

Relevance based on use case parametrization

- The model interacts directly with external consumers (Q7) OR
- The AI system operates autonomously with no human oversight (Q9) OR
- The model has a societal impact (Q10)

Evaluation requirement

The system shall allow individuals impacted by it to submit claims and complaints. The user feedback should be regularly reviewed.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

User Feedback:
Any input, complaint, suggestion, or observation provided by users or stakeholders regarding their experience with the system.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Details on how to appeal should also be provided to the users.
Tools which could be used during implementation:
- Form and Workflow Testing: Selenium
- Usability Testing Tools: Hotjar
- Accessibility Testing: Wave

Evaluation method

Verify that there are instructions on how to submit claims and complaints. Verify that claims and complaints can be submitted, e.g. check corresponding tools (see supportive guidance) are implemented, e.g., to evaluate user experience.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 17 Quality Management System (QMS)
- 26 Obligations of deployers of high-risk AI systems

TR-10

Notification of system use in terms and conditions and EULA

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

Existing terms and conditions and EULAs shall be reviewed and updated for any AI considerations.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Terms and Conditions (T&Cs):
The legal agreement that defines the rules, guidelines, and expectations for users interacting with a service or product. It specifies user rights, responsibilities, and liabilities.
End-User License Agreement (EULA):
A legal contract between the provider of software or services and the user, granting the user rights to use the software while outlining restrictions and obligations.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Considerations should be given to how the AI handles
- User prompts;
- Output rights and ownership;
- Data privacy;
- Compliance;
- Liability;

- Privacy;
- Limits on how output can be used;
- Privacy and protection of sensitive data.
Tools which could be used during implementation:
- Appropriateness of Language: Grammarly
- Legal and Compliance Check: Compliance.ai

Evaluation method

Verify that AI considerations are addressed and represented in terms, conditions and EULA.

Evaluation tools

Not applicable.

Reference to EU AI Act

Not applicable.

TR-11

Establishing regular error review and incident reporting

Relevance based on use case parametrization

Always relevant.

Evaluation requirement

An error report evaluation shall be established and executed.
Discovered incidents should be communicated to users of the AI system.

Evaluation principle

Document-based

Supportive guidance

Supplementary Definitions

Error Report Evaluation:
A systematic process for collecting, analyzing, categorizing, and addressing errors or malfunctions within the AI system to ensure continuous improvement, reliability, and safety.
Incident:
Any unexpected or incorrect behavior, malfunction, bias, or failure in the AI system that impacts performance, accuracy, safety, or user experience.

Exemplary [parameters]

Not applicable.

Additional guidance regarding the requirement

Errors should be categorized according to their severity, type and impact.

Evaluation method

Verify that a process is established for receiving and reviewing claims and complaints reports, and verify that a mechanism for reporting incidents back to users is effectively implemented.

Evaluation tools

Not applicable.

Reference to EU AI Act

- 13 Transparency and provision of information to deployers
- 17 Quality Management System (QMS)

Glossary

Access control concept: A framework that defines how access to resources, data, or systems is managed and restricted based on user roles, permissions, and security policies.

Accessibility: The system should be operable by people with a wide range of impairments, including visual impairments, hearing impairments and physical impairments; it should support multiple languages.

Active informed consent: A type of consent where individuals actively and explicitly agree to the processing and storage of their data after being fully informed about the nature, purpose, and implications of that processing.

AI policy: A set of guidelines and principles designed to govern the development, deployment, and use of artificial intelligence systems.

AI-Generated Content: Content entirely created by an AI system without direct human creation or authorship.

AI-Modified Content: Content that originates from human input but has been altered, enhanced, or adapted by an AI system.

AI-Specific Security Incidents: Events where AI systems are compromised, misused, or malfunction due to vulnerabilities, attacks, or breaches, potentially impacting model integrity, data privacy, or system behaviour.

All user groups: All people including marginalized groups, people with disabilities (e.g., accessible to screen readers, including alt text for images, colour-blind friendly palettes, etc.).

Anonymization: The process of irreversibly transforming data so that an individual can no longer be identified, directly or indirectly.

Architecture: The architecture refers to the structural design of the AI model, including the number, size and properties of network layers. It defines how data flows through the model.

Asset inventory: A detailed list of all resources, components, or items owned or used within a system. It includes physical assets, software, data, and third-party solutions.

Associated Risks: The likelihood and potential impact of adverse effects, such as security vulnerabilities or functional failures, caused by handling non-conforming inputs.

Attack Vectors: Paths or methods leveraged by adversaries to exploit vulnerabilities in systems, such as data poisoning, model manipulation, or injection of malicious triggers or backdoors.

Authenticity: The assurance that the transaction is genuine, unaltered, and can be verified as having been carried out by the AI system as intended, without unauthorized modifications or tampering.

Authenticity of data: The quality of being genuine, unaltered/altered with traces, and verifiably from a trusted source.

Automation bias: Occurs when a human decision-maker prefers recommendations made by an automated decision-making system, even when the system makes errors, over non-automated information.

Availability: Ability of a system, to remain accessible and operational for authorized users whenever needed.

Black Box: A scenario where attackers have no internal knowledge of the system or model and rely on observable input-output behaviour.

Boundary Values: Inputs at the edge of an allowed range or threshold, used to evaluate a system's performance and stability under extreme conditions.

Consistent logging: Systematic recording of events, actions, and decision made by the system.

Conventional IT security: Traditional methods and practices for protecting systems, data, and infrastructure from unauthorized access, misuse, or harm.

Coverage bias: Happens when the population represented in a dataset does not match the population that the machine learning model is making predictions about.

Critical level of bias: A threshold or degree of bias that, if exceeded, poses a significant risk to fairness, functionality, or compliance with ethical and legal standards.

Data acquisition: The process of collecting, measuring or obtaining raw data from various sources to be used for training and validating machine learning models.

Data characteristics: Attributes or properties that describe the nature, structure, quality, and behavior of data. Documenting these characteristics helps ensure that data is well-understood, appropriately used, and suitable for its intended purpose.

Data decommissioning: The process of securely and systematically retiring, archiving or disposing of data that is no longer needed.

Data Management: The practice of organizing, storing, protecting and maintaining data to ensure its quality, accessibility, and usability throughout its lifecycle.

Data planning: The process of strategically defining how data will be collected, processed, managed, stored, analyzed, and maintained to meet organizational objectives and compliance requirements.

Data Preparation: The process of cleaning, transforming, and organizing raw data into a format suitable for analysis, modeling, or machine learning. This step ensures that data is of high quality and ready for use.

Data provisioning: The process of making data available to users, applications, or systems in a secure, efficient, and controlled manner. This can involve extracting, preparing, and delivering data from various sources to its intended destination.

Data quality: The degree to which data meets the standards for (among others) consistency, informativeness, representativeness, timeliness, validity, reliability, completeness.

Data split ratios: Proportions of data allocated to training, validation, and test sets.

Deep Learning: Deep learning is a subset of machine learning involving neural networks with multiple layers, enabling the model to learn complex patterns and make predictions from large amounts of data.

Disruptions: For example server crashes, network outages, high traffic or load, hardware failures, power failures etc.

End-User License Agreement (EULA): A legal contract between the provider of software or services and the user, granting the user rights to use the software while outlining restrictions and obligations.

Error Report Evaluation: A systematic process for collecting, analyzing, categorizing, and addressing errors or malfunctions within the AI system to ensure continuous improvement, reliability, and safety.

Evasion Attacks: Methods where attackers modify inputs to trick a system or model into making wrong or unexpected decisions without changing the system itself.

Event of a failure: An incident where a system, component, or process does not perform as intended, leading to disrupted functionality, reduced performance, or complete inoperability. Common types of failures include hardware failures, software failures, security failures, network failures or data failures.

Explanation of the system: An explanation of a system refers to the information provided to make the system's decisions, outputs, or behavior understandable to different stakeholders. This can include describing how, why, and under what conditions a decision or outcome was made.

Extraction Attacks: Methods where attackers try to obtain sensitive information from a model, such as its training data, parameters, or underlying logic, by analyzing its responses or internal structure.

Fairness: Fairness in AI systems refers to the principle that decisions and predictions should be made without prejudice or discrimination. A fair AI system ensures that all individuals or groups are treated equally, regardless of gender, ethnicity, age, or other protected characteristics. It involves ongoing measures to review, assess, and correct any unfair outcomes to maintain equitable treatment.

Finetuning: Finetuning is the process of tweaking a pre-trained machine learning model on a specific, smaller dataset to optimize its performance and accuracy for a particular task or domain.

Formal incident response procedures: Structured guidelines and processes for identifying, managing, and mitigating security incidents or breaches.

Formal Verification: The use of mathematical techniques to ensure that the system or its components operate as intended by checking against predefined rules and specifications.

General Purpose AI (GPAI): Artificial intelligence systems designed to perform a wide range of tasks across various domains. These systems are adaptable and can solve diverse problems using the same core model, making them versatile for multiple applications.

Generalize: The ability of an AI system to perform well on new, unseen data. This demonstrates adaptability beyond examples used in training datasets.

Generative AI (GenAI): A type of artificial intelligence that creates new content, such as text, images, or audio, by learning patterns from existing data.

Generative AI Foundation Model: A Generative AI Foundation Model is a type of AI technology that uses large amounts of data to generate new content, mimicking various forms of human-like output such as text, images, or speech based on learned patterns and examples.

Graduated set of controls: Security measures that were implemented in layers or tiers.

Gray Box: A scenario where attackers have partial knowledge of the system or model, such as access to some parameters, architecture details, or limited input-output behaviour.

Group attribution bias: Occurs when a human assumes that what is true for one individual or object is also true for all individuals or objects in the group.

Group bias: The tendency of an AI system to favor or disadvantage certain groups based on patterns learned from biased data associated with group characteristics such as gender, ethnicity, or age.

Human Control: This refers to a pure assistance system that cannot trigger any reactions without human confirmation and is therefore completely dependent on human operation.

Human Oversight: The process in which humans closely monitor AI systems and intervene when necessary to ensure safe, ethical, and compliant operations.

Human in-the loop: This refers to a semi-autonomous AI application that is able to process tasks but cannot complete them independently and is dependent on the operation or confirmation of a human.

Human on-the Loop: This means that humans are rarely or not at all involved in decisions under normal conditions, but instead mainly monitor the AI application. It is possible to correct decisions retrospectively, even if immediate intervention is not possible at all times and in all places.

Human out-of-the Loop: This refers to the state in which an AI application acts fully autonomously under all conditions, including errors or unforeseen events.

Impact analysis: An assessment of how changes or risks affect a system's functionality, security, and performance, helping identify risks and plan mitigations.

Incident: Any unexpected or incorrect behavior, malfunction, bias, or failure in the AI system that impacts performance, accuracy, safety, or user experience.

Individual bias: The tendency of an AI system to make skewed decisions based on patterns learned from biased data related to individual preferences or characteristics, e.g. when a person with a university degree from a less known university applies to a job that is not recognized by an AI system.

Infrastructure: IT-Infrastructure that houses the key components of the AI System such as the database, computational resources used for training and inference and the model itself.

Integrity: In terms of data, integrity ensures that data remains unchanged, uncorrupted, and true to its original form through collection, storage, transmission and processing.

Intended purpose: The specific goal or function that the system is designed to achieve. It defines why something was created or implemented and what outcomes are expected from its use.

Interested Parties: Individuals or entities who have a vested interest in the AI system and require documentation to fulfill their roles or responsibilities. Examples include /- Developers and engineers/- End users/- Compliance officers and regulators/- Business stakeholders/- Supervisory authorities/- Partners.

IP (Intellectual property): Concerns content that is subject to copyright, trademark, or patent protection.

KPIs (Key Performance Indicators): Metrics to evaluate the performance.

Large Language Model: A Large Language Model (LLM) is an AI-based model trained on vast datasets to understand and generate human-like text, enabling sophisticated tasks such as translation, summarization, and conversation.

Life cycle: The process from design and development through deployment, operation, and eventual decommissioning.

Machine Learning (ML): A type of AI where systems learn from data to make predictions or decisions without explicit programming. It includes methods like supervised, unsupervised, and reinforcement learning.

Model Theft Attacks: Attempts by attackers to replicate or steal a machine learning model by exploiting access to its outputs, structure, or functionality.

Multiple Models: The practice of training and evaluating different machine learning algorithms or architectures on the same task to compare their performance, robustness, and suitability before selecting the final model.

Non-Conforming Inputs: Inputs that deviate from defined standards, such as malformed, corrupted, or improperly formatted data, potentially affecting system integrity or performance.

Non normality bias: Arises when statistical methods are applied to non-normally distributed data.

Non-Specified Users: Individuals who are either untrained or lack proper authorization to access the system.

On Site / Cloud / Hybrid Infrastructure: A piece of IT-Infrastructure is considered to be On Site, if it is physically present in location owned and controlled by the owner of the AI system. It is considered to be in the Cloud if it is located in a datacenter or compute cluster of a cloud provider such as Azure or AWS. An AI System can have parts of its infrastructure in the Cloud and parts on site in this case it is referred to as hybrid.

Output Integrity Attack: An attack that occurs when an adversary manipulates or alters the final results produced by the AI model without interfering with the input data or performing adversarial attacks on the model itself. In this type of attack, the input data remains untouched, but the generated output is modified or tampered with before it reaches the end user or decision-making system.

Overfitting: A situation where a model learns the training data too well, including noise and specific details, resulting in poor performance on new, unseen data due to lack of generalization.

Overreliance: The tendency of users to depend excessively on the AI system's outputs or recommendations, potentially leading to reduced human judgment, critical thinking, or oversight, even in cases where the AI might produce incorrect or incomplete results.

Override the system: Update or modify decisions or completely stop the operation of the system.

Performance: Indicator that shows how effectively and efficiently an AI system accomplishes its tasks.

Performance Requirements: Criteria that define the expected level of functionality and efficiency of the system, often measured through specific metrics like accuracy, speed, reliability, or resource usage, to ensure it meets its intended goals.

Personally Identifiable Data: Any information relating to an identified or identifiable natural person. This can include anything from a name, a photo, an email address, bank details, posts on social networking websites, medical information, or even a computer IP address. The key aspect is that if the information can directly or indirectly identify a person, it qualifies as personal data.

Plausible: The AI system's output should be credible, reasonable, and appropriate given the context of the inputs.

Poisoning Attacks: A type of adversarial attack where malicious data is injected into the training dataset to corrupt the model's learning process, leading to degraded performance, biased outputs, or hidden malicious behaviours.

Pre-Trained: A model is considered to be pretrained when it has already been trained on a large, general dataset before being customized or finetuned with specific data for targeted tasks. This allows the pretrained model to have a broad understanding which can then be adapted for particular applications.

Privacy: Privacy in AI systems refers to the protection of personal data and ensuring its confidential treatment throughout the system. A privacy-conscious AI system collects, processes and stores personal data only when necessary and in accordance with relevant data protection laws (such as GDPR). It also incorporates measures such as anonymization or pseudonymization, restricts access by unauthorized users, and ensures that individuals are informed about the use of their data and have given their consent.

Privacy Impact Analysis (PIA): A systematic process to identify, assess, and mitigate risks to individuals' privacy posed by a system. It evaluates how personal data is collected, stored, processed, and shared, ensuring compliance with data protection laws and safeguarding individuals' rights.

Procedures (exemplary): Monitoring and Logging, Self-Checking Mechanisms, Error Detection Algorithms, Threshold-Based Alerts, Redundancy Checks, User Feedback Loops.

Pseudonymization: The process of replacing identifiable data with pseudonyms or identifiers so that individuals cannot be identified without additional information.

Quality Management System (QMS): A structured framework that is concerned with the implementation, monitoring, and governance of AI systems.

Reliability: Reliability in AI Systems refers to the ability of the system to consistently perform its intended functions over time and across varying conditions. A reliable AI system operates without significant downtime or failure, ensuring continuous performance even when subjected to diverse workloads or environmental changes. The goal is to ensure long-term operational stability and dependable outcomes.

Resilience: Resilience in AI systems refers to the ability of the system to maintain its functionality and stability in the face of adverse conditions, such as technical failures or unforeseen disruptions. A resilient AI system can self-monitor, detect failures, and take necessary recovery actions to ensure that it remains operational. This includes both the prevention of external threats and the system's capacity to recover quickly and fully after an incident.

Revocation of consent: The ability of individuals to withdraw their previously granted consent at any time, making further processing or storage of their data unlawful unless another legal basis exists.

Right to be forgotten: The right of individuals to have their personal data erased or deleted when it is no longer necessary for the purposes for which it was collected, or when consent has been withdrawn (GDPR Art. 17).

Right to deletion (Right to be forgotten): The right of individuals to have their personal data erased or deleted when it is no longer necessary for the purposes for which it was collected, or when consent has been withdrawn (GDPR Art. 17).

Robustness Reviews: Complete evaluations of a model's ability to maintain functionality and performance under adversarial attacks, unexpected inputs, or challenging conditions.

Robustness: Robustness in AI systems refers to the ability to function reliably and consistently even under uncertain conditions, such as unexpected inputs or disturbances. A robust AI system delivers consistent and correct results on previously unseen data, even when faced with noisy or manipulated data in adversarial attacks. The goal is to ensure that the system is not easily disrupted by external influences.

Runtime Environment: The combination of hardware and software configurations required to execute a system or application, including operating systems, frameworks, libraries, and physical devices essential for proper functionality and performance.

Sampling bias: Occurs when data records are not collected randomly from the intended population.

Secure Integrations: The process of safely connecting systems or components to ensure data integrity, confidentiality, and functionality.

Security: Security in AI systems refers to the protection of the underlying infrastructure, data, and communications from cybersecurity threats. A secure AI system employs various security mechanisms, such as encryption, authentication, and access controls, and includes regular security audits to safeguard against unauthorized access and misuse. The aim is to ensure that both data and system functionality are protected from internal and external threats, including potential cyber-attacks and data poisoning.

Security Controls: Measures and mechanisms implemented to safeguard systems, data, and processes against threats, ensuring confidentiality, integrity, and availability as per applicable regulations and standards.

Security Reviews: Systematic assessments to detect and address vulnerabilities in productive models, ensuring they are safeguarded against potential threats, breaches, and unauthorized access.

Self error detection: Criteria and mechanisms for automatic error detection.

Sensitive Data: Special category of personal data which require more protection due to their sensitivity. This includes information about an individual's racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data for the purpose of uniquely identifying a person, health data, and information concerning a person's sex life or sexual orientation. Such data can only be processed under strict conditions.

Sensitive features: Features that were identified during the assessment in FAIR-01, e.g., gender, ethnicity, religion etc.

Shallow Learning: Shallow learning refers to machine learning methods that involve simpler, often linear models with fewer layers and complexity compared to deep learning, focusing on tasks that require less hierarchical feature extraction.

Simpson's paradox: Manifests when a trend indicated in individual groups of data reverses when the groups of data are combined.

Societal Impact: Societal Impact in AI systems refers to the broader ethical and social consequences of deploying AI technologies. A system designed with societal impact in mind ensures it contributes positively to society and avoids causing harm to communities, individuals, or the environment on all timescales. The goal is to align the AI system with societal values, ensuring its use promotes the common good and is both socially and environmentally sustainable.

Stability: In the context of data, it refers to the degree to which data remains consistent, reliable and unchanges over time or under varying conditions. It is critical for maintaining trust in data-driven processes, ensuring consistency, and supporting robust decision-making.

Stratification method: Criteria and approach for data stratification.

Supply chain: The network of entities and processes involved in creating, delivering, and maintaining a system.

System description: A detailed document ensuring transparency and ease of understanding.

System interfaces: Points of interaction where a system communicates with other systems, applications, or users.

Tailored Explanation: A specific, context-aware description provided to users, detailing the reasoning, factors, or data that led the AI system to produce a particular output or decision. The explanation is adapted to the user's needs, expertise, and the complexity of the output.

Tampering: The unauthorized alteration, manipulation, or interference with data, inputs, or outputs to compromise the integrity, functionality, or security of the system.

Terms/Conditions: The legal agreement that defines the rules, guidelines, and expectations for users interacting with a service or product. It specifies user rights, responsibilities, and liabilities.

The Cloud Computing Compliance Criteria Catalogue: The Cloud Computing Compliance Criteria Catalogue (C5) by BSI outlines minimum security requirements for cloud applications.

Third Party: An external entity (not directly affiliated the system's owner), responsible for providing training data, models, or other services.

Third-Party Model: A model that was developed and trained by a third party, i.e. not trained and developed by the user of the model.

Traceability: Refers to the ability to track, document, and verify the origin, history and transformation of data throughout its lifecycle.

Training Data: Training data is a dataset used to teach machine learning models how to recognize patterns and make decisions, forming the foundational knowledge that guides their predictions and actions.

Transaction: Any action or event initiated or completed by the AI system that results in a change of state, data exchange, or processing outcome.

Transparency: Transparency in AI systems refers to the clarity and openness with which the system's operations, decision-making processes, and data handling practices are communicated. A transparent AI system ensures that stakeholders, including users and regulators, can understand how decisions are made and how data is used. This involves clear documentation of algorithms, data sources, and decision criteria, as well as providing explanations for the system's outputs.

Uncertainty Assessment: The process of evaluating and quantifying the confidence or reliability of an AI system's predictions or outputs, often by analyzing how variations in inputs or conditions impact its performance and decision-making.

Underfitting: A situation where a model fails to capture the underlying patterns in the training data, leading to poor performance both on the training data and unseen data.

Uniquely Identifiable: Each transaction has a distinct identifier (e.g., a transaction ID or hash) that makes it distinguishable from all other transactions.

Unspecified Usage Environments: Physical or operational settings that are not explicitly defined.

User Awareness: Ensuring users clearly understand that they are interacting with an AI system rather than a human.

User Feedback: Any input, complaint, suggestion, or observation provided by users or stakeholders regarding their experience with the system.

White-Box: Scenarios where attackers have full access to the internal structure, parameters, and algorithms of a system or model, allowing detailed analysis and exploitation.