6 Criteria and Supplementary Information

Where useful, references to C5 criteria are given within the criteria of the AIC4. The references are indicated by "C5" at the beginning of the reference followed by the criteria ID (i.e. C5-XX-XX).

6.1 Preliminary Criteria

PC-01 General Cloud Computing Compliance

Criterion

The AI service provider demonstrates compliance of the AI service with general cloud computing compliance criteria, as set out in the Cloud Computing Compliance Criteria Catalogue (C5).

The service is compliant according to the C5. This is shown in a report that covers at least the following aspects:

- **Scope**: The subject of the C5 report specifically covers the AI service;
- **Coverage Period:** The C5 report (or multiple reports) covers the full audit period of the attestation according to the AI Cloud Service Compliance Criteria Catalogue. Alternatively, a bridge letter is provided for the gap;
- **Report Type:** The C5 report is of at least the same type (Type 1 or Type 2) as the attestation according to the AIC4;
- **Qualified Service Auditor:** The C5 report is issued by a qualified auditor;
- **Opinion**: The C5 opinion for the AI service in scope of attestation according to the AIC4 is unqualified;
- **Timing**: The C5 report is made available prior to the issue date of the AI Cloud Service Compliance Criteria attestation report.

Supplementary Information

About the Criterion

In case the C5 opinion for the AI service in scope of attestation according to the AI Cloud Service Compliance Criteria Catalogue is qualified, the independent auditor should make this transparent in the report and evaluate the impact on the opinion for his engagement according to the AIC4.

PC-02 Standard for Documentation of the AI Service Provider

Criterion

Policies and instructions covering system robustness, development, deployment, operation and maintenance of the AI service as well as relevant subsystems are documented.

At least following requirements are fulfilled:

- **Structure**: Documentation follows a clear structure in which the information is divided into sections in a coherent manner;
- **Access**: The document is accessible for all relevant parties;
- **Coverage**: The document covers all relevant points of the topic;
- **Roles and Responsibilities:** Authorities and competencies for managing the matter to be documented are defined;
- **Accurate:** Information contained in the documentation is correct;
- **Versioning**: The edit history of the documents is tracked;
- **Components**: Qualitative and quantitative elements are used where applicable to aggregate the relevant information;
- **Review**: The documentation is reviewed and updated on a regular basis (at least annually).

Supplementary Information

About the Criterion

This criterion is closely related to the criterion for Documentation, Communication and Provision of Policies and Instructions (C5-SP-

01). For clarity reasons, this criterion is replicated and slightly modified to meet the AI-specific needs.

Policies and instructions are required for the following criteria in which the content is specified in more detail:

- Results of the risk exposure assessment (SR-02)
- Implemented countermeasures (SR-06, SR-07)
- Model selection process and decisive factors (PF-05)
- Final model specifications and achieved performance (PF-05)
- Test methodology and results of business testing (PF-06)
- Issues identified during performance reviews (PF-10)
- Resource planning procedure (RE-01)
- Policies and instructions for the logging process (RE-02)
- Processes and detected inconsistencies related to AI specific security incidents (RE-05)
- Policies and instructions related to backup and disaster recovery (RE-06)
- Specifications of the data quality requirements for development and operation (DQ-01, DQ-02)
- Assessment requirements and results of the data selection process (DQ-04)
- Concept of data ownership (DM-01)
- Streams of user feedback included for training purposes (DM-03)
- Assessment of the credibility of data sources (DM-04)
- Results of the assessment of the required degree of explainability (EX-01)
- Results of the bias assessment (BI-01)

6.2 Security & Robustness

Objective

1. Risks caused by malicious attacks to the AI system are assessed.

2. Relevant threat scenarios are considered.

3. The effectiveness of defense measures is evaluated.

SR-01 Continuous Assessment of Security Threats and Countermeasures

Criterion

Procedures are implemented by the AI service provider to continuously monitor and assess new threats related to the AI model(s) within the scope of the AI service. In line with PC-01 the principles of the Risk Management Policy (C5-OIS-06 and C5-OIS-07) must apply.

Results are consolidated in threat scenarios. A documented description of a threat scenario contains at least:

- Details of the model architecture or machine learning algorithm that are vulnerable and concrete attack vectors against such threats;
- Characteristics of the data the attack vector operates on or with, such as structure or type;
- If available, references to the implementation of the attack vector or a concrete explanation on how to implement the attack vector and respective countermeasures.

The threat scenarios must incorporate actual security incidents according to RE-05.

Identified threat scenarios are followed up in the risk exposure assessment in SR-02 and SR-03.

Supplementary Information

About the Criterion

The AI service provider should continuously (at least quarterly) investigate state-of-the-art research and methodologies in order to stay up to date to new threat scenarios and attacks. Relevant threats for this criterion are in particular those that can lead to:

- leakage or corruption of data or AI models;
- violation of the integrity, confidentiality or availability of the AI service;
- intentional misuse of functionality or malfunction of the AI service.

Threats related to AI model(s) include for instance adversarial examples, poisoning attacks, model stealing attacks, model backdoors and membership inference attacks.

Release logs or similar sources of information for software packages implementing adversarial examples, data poisoning attacks and privacy methods should be carefully investigated with regards to the feasibility and applicability.

SR-02 Risk Exposure Assessment

Criterion

A risk exposure assessment is carried out by formulating a threat model that specifies the conditions under which the AI model(s) in scope of the AI service can be attacked. In line with PC-01 the principles of the Risk Management Policy (C5-OIS-06 and C5-OIS-07) and Managing Vulnerabilities, Malfunctions and Errors (C5-OPS-19) must apply.

The threat model includes at least following points:

- Threat scenarios derived from SR-01;
- Adversary's goals;
- Adversary's knowledge about the AI service;
- Adversarial capabilities.

Based on estimated impact and probability of occurrence, threat models are prioritized and assigned to risk owners who formally define

and document which risks have to be mitigated.

The results of the risk exposure assessment are documented in accordance with PC-02.

Supplementary Information

About the Criterion

Based on the mitigation decisions, subject matter experts implement concrete attacks and test the AI service against specific weaknesses as specified in SR-04 and SR-05, if applicable. The prioritization of the risks identified should be conducted according to a risk matrix taking into account the probability of occurrence and the impact of the threat.

Adversary's goals include targeted or untargeted misclassification, confidence reduction, membership inferences or tampering with training data.

Adversary's knowledge about the AI service can be white box, grey box or black box and can contain knowledge about data preprocessing such as filters.

Adversarial capabilities include perturbation domains, bounds of the adversary and computational resources.

SR-03 Regular Risk Exposure Assessment

Criterion

The Risk Exposure Assessment is re-evaluated at regular intervals (at least annually) or in case of events such as:

- Changes to the AI system that affect the operating principles;
- Newly identified threats according to SR-01.

Supplementary Information

About the Criterion

Changes, which affect the operating principles of the AI system include:

- introducing new features;
- extending the applicability of the service for larger user groups;
- retraining according to PF-07.

SR-04 Testing Learning Pipeline Robustness

Criterion

Based on the mitigation decisions for specific threat models for the learning pipeline of the AI model(s) within the scope of the AI service (e.g. based on data poisoning or data tampering through backdoors) derived from the risk exposure assessment in SR-02 and SR-03, the AI model(s) within the scope of the AI service are tested by simulating attacks. These tests take into account the integrity of the relevant data sets and their impact on the AI model(s) within the scope of the AI service. Threat models, attack vectors and identified vulnerabilities are followed up as specified in SR-06.

Subject matter experts perform a sensitivity analysis to estimate the impact of data contributed by users on future changes to the AI service in order to measure the risks associated with the inclusion of user data into the learning pipeline.

Data access management according to DM-02 is taken into consideration.

Supplementary Information

About the Criterion

Known state-of-the-art vulnerabilities of the learning pipeline include following types of data poisoning attacks:

- Logic corruption;
- Data manipulation;
- Data injection.

Note: In contrast to DM-02 this criterion focuses on protection of data integrity against external threats, while DM-02 aims to protect the data used for development and operation.

SR-05 Testing of Model Robustness

Criterion

Based on the mitigation decisions for concrete threat models for the AI model(s) within the scope of the AI service (e.g. based on adversarial attacks or privacy attacks) derived from the risk exposure assessment in SR-02 and SR-03, the AI model(s) are tested by implementing attacks to exploit identified vulnerabilities.

Specifications of the implementation and configuration of the tested attacks are documented, including the results of the tests.

The attacks tested are documented including the observed system behavior of the AI service. Threat models, attack vectors and identified vulnerabilities are followed up as specified in SR-06.

Supplementary Information

About the Criterion

Depending on the threat model, testing of the AI model(s) within the scope of the AI service can include following types of adversarial attacks:

- White box attacks;
- Black box attacks;
- Adaptive attacks;
- Transferability attacks;
- Physical attacks;
- Targeted and untargeted attacks.

Furthermore, basic sanity checks should be performed (e.g. iterative attacks perform better than single-step attacks and use sufficient iterations to converge, considering computational time and respective results after convergence).

SR-06 Implementation of Countermeasures

Criterion

Countermeasures to protect the AI model(s) within the scope of the AI service and its learning pipeline against threats are implemented by the AI service provider based on the susceptibility to attacks investigated in SR-04 and SR-05 as well as in line with PC-01, following the principles of Handling Vulnerabilities and Malfunctions and Errors (C5-OPS-18 and C5-OPS-20). The countermeasures are tested adequately for effectiveness regarding identified threat models as specified in SR-02 and SR-03.

This includes prioritization and implementation of adequate proactive and reactive measures for both learning pipeline and model robustness depending on their feasibility and criticality.

The implemented countermeasures must be tested by subject matter experts not involved in their design and implementation. In order to assess the effectiveness of the countermeasures, adaptive attacks are performed.

The countermeasures are documented according to PC-02.

The suitability of implemented countermeasures as well as residual risks must be formally accepted by the risk owner. In case the risk owner does not accept the remaining level of risk, SR-07 must be considered.

Depending on the results of the sensitivity analysis performed in SR-04, the AI service provider must implement measures in order to limit the impact of data that users can contribute such that the functionality of the AI service stays intact while attack capabilities are reduced.

Supplementary Information

About the Criterion

The AI service provider should implement state-of-the-art countermeasures in order to be robust against new kinds of attacks. Following examples of countermeasures can be considered:

Adversarial defenses:

- Reactive defenses act on the input before it reaches the AI model(s) within the scope of the AI service:
  - Detection of adversarial examples;
  - Input transformation as a preprocessing step (e.g. filters).
- Proactive defenses aim at building inherently robust models:
  - Adversarial training;
  - Provable defenses;
  - Robust deep architectures (distillation);
  - Defenses based on generative adversarial networks (GAN).

Data poisoning defenses:

- Data sanitization;
- Anomaly detection;
- Golden dataset;
- Bounded Norm Defense.

Privacy measures:

Countermeasures to privacy attacks should be considered. An example could be the use of privacy preserving machine learning techniques (e.g. differential privacy, federated learning).

SR-07 Residual Risk Mitigation

Criterion

In case countermeasures derived from the tests performed in SR-04 and SR-05 do not lead to a residual risk level formally accepted by the risk owner or in case no concrete implementations are available at all, countermeasures not necessarily linked to a specific threat scenario must be implemented and tested.

The implemented countermeasures must be tested adequately by subject matter experts not involved in their design and implementation. The countermeasures are to be documented according to PC-02.

Supplementary Information

About the Criterion

Examples of alternative countermeasures are filters, cropping-rescaling or compression and decompression.

6.3 Performance & Functionality

Objective

1. The performance requirements to evaluate the AI service are appropriate given the characteristics and specifications of the target application.
2. To provide the service as set out in the system description suitable AI model(s) within the scope of the AI service are chosen.
3. Established procedures and recognized methodologies are applied for training and validation of the AI model(s) within the scope of the AI service to ensure correct functioning of the AI service.

PF-01 Definition of Performance Requirements

Criterion

Performance requirements for the AI service are defined and included in the system description according to BC-03. The defined performance requirements include at least the following aspects:

- **Performance metrics**: Performance metrics to measure the quality of the AI service must respect the established rules of technology. Target values for those metrics are set in a way that the AI service fulfills the intended purpose as outlined in the system description. The metrics used to assess the accuracy of the AI service can differ based on the respective target application.
- **Sensitivity analysis**: The stability of the performance metrics is assessed regarding uncertainties in respective input or metadata in order to estimate confidence levels.

Changes to performance requirements are also documented in the system description according to BC-03.

Supplementary Information

About the Criterion

The AI service provider selects adequate performance metrics to measure the quality of the AI service. The following metrics may be used and are open for further extension:

- Scoring: ROC curve, AUC curve, Gini coefficient;
- Classification: confusion matrix, F1-score, recall, precision;
- Regression: Mean square error, mean absolute error, root mean square error, R^2, backtesting;
- Computer Vision: Peak signal-to-noise ratio, structural similarity;
- NLP: Perplexity, BLEU score.

It can be appropriate to use sampling methods (e.g. stratified sampling) to obtain a more meaningful representation of the population and the depiction of performance thereof.

Note that in order to measure the performance of the AI service it is necessary to measure the performance of the AI model(s) within the scope of the service.

PF-02 Monitoring of Performance

Criterion

The AI service provider assigns personnel to continuously compute and monitor the performance metric(s) defined in PF-01. In scheduled intervals (at least quarterly) reports on the performance of the service are communicated to the responsible management of the AI service provider.

Supplementary Information

About the Criterion

To provide an overview of the performance of the service, dashboards should be implemented to aggregate relevant information.

The dashboards should cover the defined performance metrics of the AI service as well as KPIs that measure the underlying infrastructure performance.

PF-03 Fulfillment of Contractual Agreement of Performance Requirements

Criterion

If the target values for the performance requirements defined in PF-01 and the description of the performance measurement procedures are incorporated in contractual agreements, identified material deviations to these contractual obligations are made transparent to users. In case of deviations responsible personnel of the AI service provider request retraining of the AI model(s) within the scope of the AI service in line with PF-07.

Supplementary Information

–

PF-04 Model Selection and Suitability

Criterion

Different algorithms and model approaches are considered taking into account established rules of technology, the amount of available data, the task at hand and the performance requirements in PF-01. The documentation addresses at least the following aspects:

- **Model concept**: The suitability of the conceptual model to perform the intended task is described.
- **Model boundaries**: Limits of the conceptual model and operational boundaries are identified and their impact on the AI service is assessed.

Supplementary Information

About the Criterion

Objectives, impact and purpose of the AI service are defined in the system description according to BC-03.

The AI service provider may define templates that help to formalize the documentation process of objectives, impact and purpose of the AI service.

The templates may include the following points:

- Model concept: the AI model is in theory capable to capture the complexity of the learning task, e.g. for tasks where nonlinearity is a fact, linear models are not used.
- Model boundaries: The AI model has to be able to cover all cases required for the target application, e.g. an AI model trained to recognize German text, cannot be applied to English text without adjustments.

PF-05 Model Training and Validation

Criterion

The model(s) selected under consideration of their suitability according to PF-04 are trained, tested and validated with designated data according to DQ-06 taking into account feature selection and feature engineering.

Model performance is assessed using performance metrics specified in PF-01 and uses an independent test set (i.e. data not seen by the model during training or validation). Based on the results obtained, models may need to be adjusted and retrained with different configurations (e.g. with different architectures, parameter settings or feature engineering).

The trained models are validated on independent validation data (c.f. DQ-06- Preparation of Training, Validation and Test Data) which is used to benchmark different models and to adjust hyperparameters, if necessary.

Inaccuracies of the models such as overfitting and underfitting are evaluated and addressed. In addition, the training process includes safeguards to ensure the absence of bias with regards to BI-01.

Especially, trade-offs between performance, bias mitigation according to BI-03 and hardening according to SR-02 are considered when se-

lecting a model. The selection process and decisive factors are documented according to PC-02.

The final model specifications and achieved performance are documented according to PC-02.

Supplementary Information

About the Criterion

Depending on the model and the intended purpose, feature engineering and data cleansing/transformation (e.g. one-hot-encoding or stratified sampling) are carried out to transform the data to a form usable by the model.

For example: Solving a classification problem, a subject matter expert should start by training a linear regression model, a random forest and a neural network. For tasks, where neural networks evidently outperform other methods, three networks with different weights should be trained at the beginning. The AI service provider should use cross validation or grid search to tune the hyperparameters. Backtesting should be applied in case of timeseries data.

Evaluating and addressing overfitting:

- One indicator for overfitting can be a significantly better performance on training than on test data. Measuring feature importance can also provide insights. This can be done by applying saliency maps or tree interpreters.
- To overcome overfitting one can potentially use regularization, simpler models or fewer features. For deep learning the options of adding dropout and early stopping can be used. In addition, the number of free parameters in the model (i.e. the weights in a neural network) should be at least 5 times smaller than the number of training examples.

Evaluating and addressing underfitting:

- A bad performance on both training and test set can be an indication for underfitting or for not including appropriate features.
- To overcome underfitting, one can add more complexity to the model e.g. increase the number of free parameters or chose a different model concept.

PF-06 Business Testing

Criterion

Tests are implemented by the AI service provider and performed by subject matter experts to ensure that the AI model(s) within the scope of the AI service meet the requirements of the business process or respective target application scenario in accordance with PC-01, following the principles of Testing Changes (C5-DEV-06).

The tests are performed on a regular basis in accordance with training frequency, before go-live of the AI service and after major changes (e.g. retraining).

Test methodology and results are documented according to PC-02. The go-live is approved based on test results by authorized personnel.

Supplementary Information

About the Criterion

To test the model, subject matter experts can work with a carefully chosen “golden” dataset which should cover (all) the possible cases the system might encounter in production extensively. This dataset can be derived from real data or be sampled to meet a special composition of features reassembling cases.

When multiple AI models are chained together, correlation between errors of the respective models may affect the performance of the AI service itself.

In addition, it can be useful to compare the AI services output/ decision with the decision made by subject matter experts.

PF-07 Continuous Improvement of Model Performance

Criterion

If necessary, continuous improvement of the model performance is achieved through retraining the AI model(s) within the scope of the AI service and adjusting the conceptual model.

Model retraining is either carried out at regular intervals (defined by the AI service provider), when the AI service is subject to model/concept drift or upon demand of responsible personnel assigned in PF-02 (Monitoring of Performance).

Retraining a model follows the same principles as outlined in PF-05 and must always incorporate new ("unseen") data.

If retraining a model does not lead to a mitigation of the issue that triggered the retraining process, subject matter experts reconsider the model concept according to PF-04 and the risk exposure assessment according to SR-02 and SR-03. The adjustments and model changes are documented.

Supplementary Information

About the Criterion

Concept drift: conceptual changes such as changes in products, exposures, activities, clients, user groups, frequency of requests or quality of input data can lead to a diminished performance of the service. Subject matter experts should verify that any extension of the model beyond its original scope is valid and retrain the model if necessary.

PF-08 Additional Considerations when using Automated Machine Learning

Criterion

If parts of the development process are subject to Automated Machine Learning (Automated ML), the following aspects are considered:

- Evaluation of the degree to which automated ML is applicable and how it provides suitable and adequate functionality to satisfy the services as set out in the system description;
- Documentation of the development process undergone as well as of the model chosen in the end considering potential recombination of features, feature transformation and combination of different models (if applicable);
- Documentation of the integration of automated ML components.

The monitoring of the automated Machine Learning functionalities must provide all required information to measure the performance of the AI service as specified in PF-01 and information required for the model selection according to PF-04 and PF-05.

Supplementary Information

About the Criterion

When leveraging automated machine learning in addition to the final model, a report should be provided which covers the following areas:

- Recombination of features performed through the process;
- Feature transformation such as scaling or one-hot encoding;
- Models and feature combinations evaluated;
- Parameter grid evaluated and corresponding results;
- For ensembles: combination of different models.

The results in the report should be made plausible by applying domain knowledge of subject matter experts.

PF-09 Impact of Automated Decision-making

Criterion

In case of automated decision making, procedures and measures are in place that allow us-

ers of the AI service to update or modify the decisions made by the AI service as specified in BC-03.

Supplementary Information

About the Criterion

If there are no specifications on the extent to which users are able to correct or object to the results or decisions made by the AI service, this criterion might not be applicable.

PF-10 Regular Service Review

Criterion

Mechanisms for the review of the AI service are set up in accordance with the principles of Managing and Handling Vulnerabilities, Errors and Logs (C5-OPS-20 and C5-PSS-04). These mechanisms are executed by subject matter experts at regular intervals (at least quarterly). The review includes at least the following aspects:

- **User feedback:** Review of user/customer feedback about service output, impact and complaints;
- **Failure reports:** Evaluation of failures and problem management records that occurred during operation.

All issues identified during performance reviews are documented in accordance with PC-02 and reported in an aggregated form to the management of the AI service provider, following the principles of Managing Vulnerabilities (C5-OPS-18, C5-OPS-20 and C5-OPS-21). Identified issues with an impact on the users are made transparent to them according to the procedures outlined in the system description (according to BC-03). Appropriate measures are defined and followed up. Following points are considered:

- **Prioritization**: Measures for the remediation of identified failures and malfunctions are prioritized (e.g. in terms of criticality, impact and effort).
- **Remediation**: An action plan with defined measures to remediate identified issues is documented and includes scheduled objectives for implementation.
- **Implementation**: Realization of defined measures based on the defined action plan. Necessary retraining is carried out in accordance with PF-07.
- **Change management**: The process is subject to the change management procedures and is reevaluated at regular intervals (at least annually) on its effectiveness.

Supplementary Information

About the Criterion

User feedback provides additional information on the performance and functionality of the AI model(s) within the scope of the AI service, which can lead to new measures to improve the quality of the AI service. In the context of this criterion, failure reports shall address the operation of AI model(s) within the scope of the AI service.

6.4 Reliability

Objective

1. Defined performance thresholds are achieved by providing sufficient resources for the operation of the AI service.
2. Interactions with the AI service are monitored and assessed.
3. Safe functioning of the AI service is ensured by appropriately handling system security incidents wherever they occur.
4. Service components are recovered in reasonable time, by establishing backup plans, if needed.

RE-01 Resource Planning for Development

Criterion

The planning of capacities and resources (technical and human) for the development and further improvement of the AI service is in line with PC-01 and follows the principles from Capacity Management - Planning (C5-OPS-01).

The procedure must be documented according to PC-02.

Supplementary Information

About the Criterion

This criterion extends and builds on C5-OPS-01 as follows:

In addition to resource planning for the operation of the AI service required by C5-OPS-01, this criterion covers resources for development, validation, testing and further improvement according to PF-07.

RE-02 Logging of Model Requests

Criterion

The logging of requests should allow the backtracking of incidents and failures of the AI service to specific AI model(s).

The AI service allows logging of requests to the AI service to investigate failures or incidents. In line with PC-01 the principles for Logging of Relevant Information (C5-OPS-11, C5-OPS-12 and C5-OPS-13) must apply. The log files contain at least type of request, processing times including time stamps and metadata on the user requesting the AI service.

Log files are kept for intervals that are appropriate for the application (for at least three months) taking into account the sensitivity of the application and requirements of users.

Policies and instructions with technical and organizational safeguards for the logging process are documented and provided to authorized personnel if required. The policies and instructions are documented according to PC-02. In addition, the AI service provider outlines the information contained in the logs and their storage periods in the system description according to BC-04.

Supplementary Information

–

RE-03 Monitoring of Model Requests

Criterion

The AI service provider performs continuous checks (at least monthly) for irregularities within user requests in order to detect malicious requests against the AI model(s) in scope of the AI Service according to RE-05.

Supplementary Information

About the Criterion

In addition to security monitoring issues addressed in C5-OPS-13, irregularities can arise from different sources, e.g. an unusual large number of requests or similar requests in terms of content which should be limited.

The monitoring of AI models should also consider model theft and data poisoning scenarios according to SR-01.

RE-04 Corrective Measures to the Output

Criterion

If the AI service allows for human intervention or correction of the AI service output, only authorized subjects are allowed to correct the output based on their rights and responsibilities. A corresponding role and rights concept is in place in accordance with the Policy for User Accounts and Access Rights (C5-IDM-01).

Supplementary Information

About the Criterion

For the purpose of retraining a model, suggestions made by the users of the AI service are collected and assessed through established procedures.

RE-05 Handling of AI specific Security Incidents

Criterion

Identified security incidents related to the AI model(s) within the scope of the AI service are addressed by the AI service provider in accordance with the Policy for Security Incident Management (C5-SIM-01)

The processes and detected inconsistencies are documented according to PC-02.

Supplementary Information

About the Criterion

The identified incidents are consolidated into new threat scenarios according to SR-01. The effectiveness of the countermeasures implemented according to SR-05 should be assessed taking into account the security incidents and further improved.

RE-06 Backup and Disaster Recovery

Criterion

Policies and instructions with technical and organizational safeguards are documented and provided according to PC-02 by the AI service provider to avoid loss of relevant data and AI model(s). In line with PC-01 the principles for Data Protection and Recovery (C5-OPS-06) must apply.

They provide reliable procedures for backup management (e.g. snapshots) and recovery of models (e.g. roll-back mechanisms). Access to the backups is limited to authorized subjects.

The recovery procedures are tested at least annually. Actions required by the user must be outlined in the system description according to BC-04.

Supplementary Information

About the Criterion

Versioning, tracking and storing of datasets and AI models for development and in production should be done according to a predefined structure (type, manner and frequency) along the learning pipeline.

6.5 Data Quality

Objective

1. Data used for the training and operation of the AI service fulfills quality requirements.
2. Establish transparency, which regulations and laws the service provider meets regarding the use of data for the AI service.

DQ-01 Data Quality Requirements for Development

Criterion

Data quality requirements for development are defined to ensure a proper functioning of the AI service according to PF-01. The following aspects apply to data exploration as well as training, validation and testing data:

- Accessibility
- Amount
- Completeness
- Relevance
- Correctness
- Structural integrity

The specifications of the data quality requirements are documented according to PC-02.

For external data sources, reports on the suitability and quality of the data must be provided and compliance with applicable legal and regulatory requirements and international standards according to BC-01 must be ensured.

Supplementary Information

About the Criterion

When it comes to data exploration and during training, validation and testing of the data, the following aspects should be considered:

- Accessibility: The data sets should be easy to locate, access, obtain and view.
- Amount: Depending on the volume of model parameters, the data sets used for training, testing and validation should be sufficiently large to avoid underfitting and to reflect all relevant real-world scenarios;
- Completeness: Missing values should be replaced in an appropriate manner. This depends highly on the feature itself. Special care should be taken when dropping missing values since this can lead to a serious imbalance in the training data;
- Relevance: Extensive data exploration should help to derive underlying relationships and to determine relevant features to predict another feature;
- Correctness: The extent to which real world phenomena are incorporated in the data should be evaluated.
- Structural integrity: Data should be consistent in terms of schema and design.

External data sources include data acquired from third parties as well as openly available data.

DQ-02 Data Quality Requirements for Operation

Criterion

Data quality requirements for operation are defined to ensure a proper functioning of the AI service according to PF-01. The following aspects apply to data required for productive use of the AI service:

- Origin;
- Completeness;
- Structural integrity.

The specifications of the data quality requirements must be documented according to PC-02. In case that users of the AI service provide data required for productive use (i.e. for inference), quality requirements are made transparent according to BC-05.

For data sources acquired by the AI service provider, reports on the suitability and quality of the data must be provided and must be mapped to the data quality requirements defined above.

Supplementary Information

About the Criterion

Data quality requirements for development and operation may differ significantly depending on the type (e.g. streaming data vs. static data), number and origin (e.g. internal vs. external).

DQ-03 Data Quality Assessment

Criterion

The quality of gathered data is continuously assessed according to DQ-01 or DQ-02 respectively. Corrective measures are in place to ensure stable data quality. The steps undertaken during data assessment are documented and outlined in the system description according to BC-05.

These systematic data checks are carried out at regular intervals (at least quarterly) and detected inconsistencies are documented and followed up in a timely manner which is defined by the AI service provider.

Supplementary Information

About the Criterion

Handling of inconsistencies should be addressed immediately at best but not later than 14 days after detecting the issue.

DQ-04 Data Selection

Criterion

The AI service provider assesses data selected for training purposes as well as for the operation of the AI service based on defined assessment requirements. The assessment requirements are designed according to the criticality of the target application as well as the frequency of the learning process and include at least the following aspects:

- **Correctness:** Information contained in the data is true (does not refer to faulty data in the sense of poor data quality);
- **Bias**: The selection and aggregation of data used is statistically representative and free of unwanted bias;
- **Dimensionality**: The number of features is determined under consideration of sparseness of data, feature correlation and the curse of dimensionality;
- **Data provenance:** During the data lineage process a log file is kept, that documents changes made to the data.

The assessment requirements and results of the selection process are documented according to PC-02.

Supplementary Information

-

DQ-05 Data Annotation

Criterion

Requirements to ensure annotation accuracy and quality are defined in line with DQ-01 and documented. At least following points are considered:

- Domain knowledge of the personnel assigned;
- Quality assurance of annotation by independent personnel (e.g. four eyes principle).

Supplementary Information

-

DQ-06 Preparation of Training, Validation and Test Data

Criterion

Training, validation and testing of the AI model(s) within the scope of the AI service need

to be carried out with datasets that fulfill at least the following aspects:

- The unsplit data set is separated into training-, validation- and test data in a reasonable proportion;
- Test datasets are separated from training and validation data and therefore must not be used for training or validation. The sample size of the test data is selected depending on the variability of the input;
- Training, validation and test data shall have a similar distribution.

Additionally, it is ensured that training, validation and test data have the same shape as the data used for operation and fulfill the data quality requirements described in DQ-01.

Supplementary Information

About the Criterion

In case that only insufficient validation data can be used (e.g. unsplit data set is too small to train the desired model), techniques such as cross validation are applied to validate the model.

6.6 Data Management

Objective

1. Data acquisition for the training and operation of the AI service is done in a structured manner.
2. A viable data management framework for the data sources relevant for development and operation of the AI service is in place.

DM-01 Data Management Framework

Criterion

A framework is in place to provide guidance for acquisition, distribution, storage and processing of data required for development, operation and further improvement of the AI model(s) in scope of the AI service. This includes the assignment of tasks, responsibilities as well as rights and roles for data handling along the learning pipeline. The following aspects are addressed:

- Granting and changing (provisioning) of access authorizations based on the least-privilege principle and need-to-know principle;
- Separation of duties;
- Regular review (at least quarterly) of granted authorizations;
- Withdrawal of authorizations in case of changes in the employment relationship or role of the employee in a timely manner which is defined by the AI service provider.

Data access applies to all relevant data (including data stored on premise) used for development and further improvements.

The concept of data ownership is documented according to PC-02.

Supplementary Information

About the Criterion

Access to data should be withdrawn immediately at best but not later than 14 days after performing the required task.

DM-02 Data Access Management

Criterion

The AI service provider protects the data used for development, operation and further improvement. In line with PC-01 the principles for identity and access management (C5-IDM-01, C5-IDM-02, C5-IDM-04 and C5-IDM-05) must apply and regulatory and legal requirements specified in with BC-01 must be considered.

The implemented safeguards are outlined in the system description according to BC-05.

This includes at least the following aspects:

- Access to data for unauthorized subjects is denied;
- Training and validation data sets are secured to prevent unauthorized subjects from compromising the datasets (for instance by frequent data quality checks).

Supplementary Information

-

DM-03 Traceability of the Data Source

Criterion

Data sources used by the AI service are documented to ensure traceability of data. The documentation includes all internal and external data sources used and specifies the purpose of their use. Data sources that contain user data and that are used by the AI service are outlined in the system description according to BC-05.

An AI service that includes user feedback for training purposes highlights feedback streams as an additional data source in the documentation in line with PC-02.

In case synthetic methods are used for artificial data creation, the process is documented and made transparent to relevant users.

Supplementary Information

About the Criterion

Data factsheets and templates for datasets should provide a structured way for the required documentation.

DM-04 Credibility of Data Sources

Criterion

The data sources selected for the development of the AI service are assessed in terms of their credibility and usability by the AI service provider in accordance with the principles of the Risk Assessment for Service Providers and Suppliers (C5-SSO-02) for external data sources. The data origin, gathering process (e.g. survey, streaming) and the level of protection of the latter are taken into account.

The assessment is documented according to PC-02 and describes the type of data source (e.g. internal vs. external data collection) as well as requirements for credibility and usability. The following points are considered additionally:

- Data needed is available in reasonable time (defined by the AI service provider) and with the required quality (see also DQ-01 or DQ-02 respectively);
- The data collection process avoids unfavorable tendencies of data according to BI-01;
- Data is retrieved in compliance with applicable legal and regulatory requirements and international standards according to BC-01, whereby these requirements shall be identified in line with the principles of the identification of applicable legal, regulatory, self-imposed or contractual requirements (C5-COM-01).

Identified issues are followed up in a timely manner which is defined by the AI service provider. The assessment is carried out by subject matter experts of the AI service provider before model training/validation takes place.

External data sources are described according to BC-05.

Supplementary Information

About the Criterion

To protect the credibility of data adequately, data should be stored in encrypted form whenever possible.

For additional information about compliance checks see C5-COM-01 (Identification of applicable legal, regulatory, self-imposed or contractual requirements).

6.7 Explainability

Objective

1. Decisions of the AI service are made explainable, if necessary.
2. Appropriate techniques are used to provide explainability for decisions made by the AI service, if necessary.

EX-01 Assessment of the required Degree of Explainability

Criterion

Based on the criticality of the AI service, an assessment for the need for explainability is carried out by persons with relevant domain knowledge, taking into account:

- Purpose;
- Potential damages;
- Needs and prerequisites for human decision making;
- Adequate handling of outliers.

The results are documented in line with PC-02 and must consider the following aspects:

- Applicable legal and regulatory requirements and international standards according to BC-01 that require the explainability of actions of the AI service;
- Justified interest by users, which requires the implementation of methods to improve explainability.

The identified need of explainability to be provided by the AI service is outlined in the system description according to BC-06.

Supplementary Information

About the Criterion

A need for explainability may for example arise during the debugging process of an AI model within the AI service.

EX-02 Testing the Explainability of the Service

Criterion

Based on the assessment carried out in EX-01, the provided explanations must be tailored for the recipient of this information (e.g. subject matter experts, business experts of the AI service provider or users) taking the recipients know-how into account. The applied methods (e.g. saliency maps, feature importance) must consider specific characteristics of the specified input and allow for a plausible indication on why the specified output was produced by the AI service.

In case the required degree of explainability derived in EX-01 cannot be provided, subject matter experts must consider the selection of a less complex model approach (e.g. random forest instead of neural network) and the corresponding trade-off between performance and explainability.

The technical limitations of used methods and shortcomings regarding the identified needs for explainability are outlined in the system description according to BC-06.

Supplementary Information

About the Criterion

Examples for explainability techniques can be divided into three categories that should be considered:

- Pre-Training: PCA, SOM (self-organizing maps), Clustering;
- Inherently explainable architectures: Linearity, monotonicity;
- Post-Training:
  - Gradient-based visualizations (Saliency maps);
  - Statistical Insights into features (Feature importance, PDP, ICE);
  - Surrogates.

6.8 Bias

The topic of bias in AI applications is often linked to moral or ethical questions like the fair treatment of individuals or groups. The BSI does not make any statements regarding ethical questions. From a security perspective, it is crucial that the AI service provider itself and the cloud customers understand the functionality and possible limitations of the AI service to a sufficiently high degree, which depends on the application. However, in order to understand the functionality of the system it is important to analyze which features determine the outcome of the system and whether there are features which have an unwanted strong effect on the outcome (i.e. bias). This objective demands that the provider thoroughly assesses the impact of bias on the functionality and security of the AI service and that corresponding threats or limitations are communicated transparently to cloud customers. Moreover, critical risks need to be mitigated. It is up to the customers to read the audit report and draw their own conclusions whether possible limitations of the functionality are acceptable for their application and whether the AI service provider considers all forms of bias relevant for the customers intended use of the cloud service. However, the outcome of an audit does not make any statements on the moral or ethical suitability of the service towards individuals for a certain application.

Objective

1. Unwanted bias within the AI service is identified.
2. Critical risks regarding existing bias are identified and mitigated.

BI-01 Conceptual Assessment of Bias

Criterion

Based on the specific characteristics of the AI service and required functionalities, a conceptual assessment is carried out by subject matter experts to evaluate the possibility of bias within the AI service and possible implications regarding the functionality, e.g. threats or limitations. Different types of bias and their origins are considered. Implications are rated and prioritized according to their criticality.

Depending on the criticality, identified implications are followed up according to BI-02 in a timely manner.

The results of the assessment are documented in line with PC-02. Identified possibilities for bias and implications affecting the functionality of the system in a critical way are outlined in the system description according to BC-06.

Supplementary Information

About the Criterion

The following types of bias should be considered:

- Direct bias
- Indirect bias
- Systemic bias
- Statistical bias
- Explainable bias
- Unexplainable bias

BI-02 Assessing the Level of Bias

Criterion

Based on the implications identified through the conceptual assessment of bias (BI-01) and the rated criticality, the data and AI model(s) within the scope of the AI service are evaluated through appropriate measures to investigate the level of bias existent in the AI service. Depending on the targeted application, potential bias is evaluated against different metrics to quantify possible effects.

The applied metrics are chosen with respect to the task at hand and expected tolerance intervals are defined by the AI service provider. If applicable, this is supplemented by measuring feature importance.

The selection of bias metrics, tolerance intervals and respective reasons are included in the system description according to BC-06.

The results of the assessment are documented in line with PC-02.

Supplementary Information

About the Criterion

Several metrics exist that can be used to quantify the level of bias. In the scientific literature, they are often called "fairness metrics". Here, fairness is understood as the non-existence of bias and is therefore not necessarily linked to ethical or moral considerations with regard to individuals.

The following fairness metrics may be included in the assessment:

- Equalized Odds;
- Equalized Opportunity;
- Demographic Parity;
- Fairness through awareness/unawareness.

BI-03 Mitigation of detected Bias

Criterion

If the applied metrics express a critical level of bias, i.e. if the defined tolerance levels from BI-02 are exceeded, measures are taken to mitigate the bias. Several mitigation methods are tested on their benefit, depending on the machine learning task and their applicability to the specific domain.

Achieved results by using mitigation methods are compared on both bias measures and standard performance requirements as defined in PF-01.

If a bias occurs that is considered critical for functionality but cannot be mitigated at the time, this limitation is included in the system description according to BC-06.

Supplementary Information

The following mitigation methods may be used and are open for further extension:

- Pre-processing:
  - Disparate impact remover;
  - Reweighting;
  - Optimized pre-processing.
- In-processing:
  - Adversarial debiasing;
  - Prejudice remover.
- Post-processing:
  - Calibrated equalized odds post-processing;
  - Reject option classification.

BI-04 Continuous Bias Assessment

Criterion

As new data is collected and the AI model(s) within the scope of the AI service are adjusted, the bias assessment and measurement are repeated regularly according to BI-01 and BI-02. If necessary, findings are followed up with respect to BI-03.

Supplementary Information

-