Bundesamt
für Sicherheit in der
Informationstechnik

Deutschland
Digital•Sicher•BSI•

AI Cloud Service Compliance Criteria Catalogue (AIC4)

Federal Office for Information Security
Post Box 20 03 63
53133 Bonn
E-Mail: aicloudsecurity@bsi.bund.de
Internet: https://www.bsi.bund.de/AIC4
© Federal Office for Information Security 2021

1 Preface by the President

Artificial intelligence (AI) is gaining importance in many areas, including information security. It enables novel attack and defense methods that achieve a high degree of automation and scalability. As the national cyber security authority and the center of excellence for cryptography, the BSI has been dealing with this topic for several years. Our approach is holistic: We not only consider mathematical-technical aspects but also the economic, political and socio-political significance of AI. On the one hand, we want to provide a foundation for secure development and usage of AI products and services. On the other hand, we want to leverage artificial intelligence and machine learning technologies to set and further develop national and international standards in cryptography and other domains of cyber security. Thereby, we contribute to improving information security in digitization to increase trust in new technologies and applications.

As a result of advances in virtualization, cloud computing and big-data technologies, methods of artificial intelligence and data-driven methods in particular have found widespread use in the IT world with great success. With the wide range of applications for AI systems at hand, the question of their trustworthiness with regard to their individual application environments becomes increasingly urgent. Since the established standards are not sufficient in this context, the BSI is actively addressing this issue in order to provide solutions in a systematic fashion. The Artificial Intelligence Cloud Services Compliance Criteria Catalogue (AIC4) marks the first step in allowing users to evaluate the trustworthiness of AI-based services developed and operated in a cloud environment.

A standard for testing AI based cloud services that is both open to the market and future-proof has to satisfy two basic needs. Firstly, AI cloud service users must be able to assess the trustworthiness of the provided AI service for their individual applications as detailed and accurately as possible. This calls for a comprehensive catalogue containing specific criteria and suitable audit methods, so that audit results are transparent for the customers. Secondly, the catalogue must only specify requirements and not dictate controls. Giving the AI cloud service providers the freedom to design individual controls ensures that a large number of market participants can apply the criteria catalogue.

The AIC4 criteria catalogue of the BSI fulfils both basic requirements. It is based on the BSI's internationally recognized and well-established C5 criteria catalogue for cloud computing and defines additional, AI-specific requirements for the AI-based cloud service. Furthermore, the AIC4 adopts the evaluation methodology of the C5.

The AIC4 criteria catalogue forms a strong foundation for evaluating the trustworthiness of AI-based cloud services. It serves as a prerequisite for ensuring the information security of future-proof AI applications in concrete development and deployment environments. With this catalogue, the BSI provides renewed proof that Germany still plays an important role in shaping new technologies such as AI.

Arne Schönbohm

Arne Schönbohm

Table of Content

2 Introduction

Methods of Artificial Intelligence (AI) play an increasingly important role for a wide range of businesses and institutions. Systems building on AI methods progressively influence critical processes and decisions. While AI methods allow for new opportunities and applications, the systems based on them are exposed to new security threats, which do not apply to standard IT systems and are therefore not covered by classic IT security standards. To close this gap for AI services running in cloud environments, this AI Cloud Service Compliance Criteria Catalogue (AIC4) was developed. Within this catalogue, an *AI service* refers to a cloud service utilizing AI methods deployed either in a public or private cloud infrastructure.

This AI Cloud Service Compliance Criteria Catalogue provides AI-specific criteria, which enable an evaluation of the security of an AI service across its lifecycle. The criteria set a baseline level of security, which can be reliably assessed through independent auditors. The catalogue has been developed for AI services that are based on standard machine learning methods and iteratively improve their performance by utilizing training data. This includes, for instance, algorithms like gradient boosting algorithms, random forests and deep neural networks. Typical applications of these algorithms include for example image classification or segmentation, time series prediction, natural language processing, voice recognition or scoring. Federated Learning and Reinforcement Learning is currently not covered by the present criteria and will be addressed by future updates.

The AIC4 allows an independent auditor to conduct an attestation engagement on the AI service's compliance with the criteria. The criteria and hence the assurance engagement cover the full AI-lifecycle, i.e. an evaluation of all relevant processes and control measures covering development, testing, validation, deployment and monitoring of such services. The AIC4 is developed in a way that it would allow for a so-called attestation engagement, if requested. In such an engagement, the auditor documents the results of the audit procedures in a detailed and comprehensive report stating compliance with the criteria and possible findings. However, the purpose of such a report from an independent auditor is not to imply that the usage of AI methods is suitable for a specific application or particular cloud customer. It is the responsibility of the cloud customers themselves to, where required, make use of the transparency provided by the attestation report and conduct its own risk evaluation to verify whether the level of information security is sufficient for the specific application at hand or not. It is further important to emphasize that the BSI is not involved in neither the selection of auditors, nor the audits itself and does not check the reports created.

The research regarding AI methods is ongoing and many challenges are not yet solved, e.g. regarding provable robustness and explainability. This AI Cloud Service Compliance Criteria Catalogue does not claim to provide solutions for those unsolved problems. However, in order to comply with the AIC4 and to achieve a successful attestation, the AI Service Provider has to prove that along the lifecycle, processes and controls following state-of-the-art methods are used. A respective audit report would provide a basis for customers to assess the suitability of the AI service and related AI methods for the application at hand.

In order to gain a full picture over the security of an AI service provided by the AI service provider, the classic cloud service risks also need to be considered. For this reason, one of the criteria (PC-01) covers an attestation according to the BSI Cloud Computing Compliance Criteria Catalogue (C5)[1]. This is an indispensable prerequisite for compliance with AIC4 and, where appropriate, an attestation according to the AI Cloud Service Compliance Criteria Catalogue. This Criteria Catalogue is therefore understood

[1] Cloud Computing Compliance Criteria Catalogue (C5), Federal Office for Information Security, 01/2020

as an add-on or extension to the C5 attestation for specific Cloud services based on AI to address AI-specific risks.

Within this catalogue the following concepts are used:

- **AI models within the scope of the AI service:** Refers to the AI methods required for the provisioning of the AI service.
- **AI service:** Refers to a cloud service utilizing AI methods deployed either in a public or private cloud infrastructure.
- **AI service provider:** Legal entity that offers the AI service to individuals, groups or other entities.
- **Target Application:** Refers to the intended use and the performed task of the AI service.

The following definitions shall apply (as also used in ISAE 3000 and ISAE 3402):

- **Attestation engagement**: An audit engagement in which the auditor verifies that the written statement is free from material misstatement.
- **Criteria**: The criteria used to assess the robustness, security, performance, reliability, data quality, explainability or unwanted bias of AI services as defined in the AI Cloud Service Compliance Criteria Catalogue.
- **Control**: Process-integrated or process-independent measure to reduce the probability of occurrence of events or to detect events that have occurred in order to maintain the robustness, security, performance, reliability, data quality, explainability or absence of unwanted bias of the AI service.
- **Material misstatement:** Shortcomings in the statement, for instance information not indicating the insufficient design of controls, false or missing information and/or information including inappropriate generalization.
- **Service organization's system:** The principles, procedures and measures applied by the legal representatives (management) of the AI service provider towards the organizational and technical implementation of management decisions.
- **Written statement:** Assertions on the description of the service organization's system for the provision of the AI service and on the suitability of the design and, if relevant, operating effectiveness of the controls to meet the AIC4 Criteria prepared by the legal representatives of the AI service provider.

A comprehensive list of relevant definition can be found in Section 7.2.

3 Structure and Content of the Criteria

3.1 Structure

The AI Cloud Service Compliance Criteria Catalogue is divided into eight criteria areas representing the security requirements of AI services related to the use of machine learning methods. Each area contains various criteria which provide comfort that the security objectives have been met in the course of an audit (cf. Section 3.2).

Each criterion contains two sections; the criterion itself and supplementary information. The criteria define the minimum scope of an audit according to the AI Cloud Service Compliance Criteria Catalogue. The supplementary information provides further guidance on how to achieve the requirements of the respective criterion.

According to the BSI the criteria make up the minimum requirements for professional AI usage. They ensure that the AI service provider uses state-of-the art processes and controls along the lifecycle. A successful audit does not necessarily imply that the robustness, performance, reliability, data quality, explainability or absence of unwanted bias is suitable for a given application. It is up to the AI service users to assess the extent to which these criteria adequately meet their specific information protection needs based on their planned application of the AI service.

An audit report provides a basis for such a risk analysis including details about the controls and processes in place at the AI service provider. Potential AI service users should in particular consider the information on the general conditions of the AI service (cf. Section 5).

3.2 Content of the AI Cloud Service Compliance Criteria

The AI Cloud Service Compliance Criteria are subdivided into eight areas.

No.	Area (identifier)	Objectives
1	Preliminary Criteria (PC)	Ensure that the AI service provider demonstrates compliance of its AI service with general cloud computing compliance criteria according to C5 as well as consistent documentation of policies and instructions applicable for the AI service.
2	Security & Robustness (SR)	Secure the AI service and improve its robustness against attacks by ensuring confidentiality and integrity of data along the service's training pipeline through testing against malicious input data as well as the implementation of countermeasures.
3	Performance & Functionality (PF)	Ensure accurate processing in line with performance requirements for the AI service by training, evaluating and testing extensively before deployment. Tailored assessment methods are used in accordance with requirements for accurate processing. Training of AI model(s) within the scope of the AI service follows established procedures.
4	Reliability (RE)	Ensure continuous operation of the AI service in the production environment and investigate possible failures through appropriate procedures for resource management, logging, failure management and backups.

No.	Area (identifier)	Objectives
5	Data Quality (DQ)	Ensure that data used by the AI service (e.g. training and test data) comes from trustworthy sources, fulfills quality criteria, is annotated correctly and is protected adequately.
6	Data Management (DM)	Ensure that a framework providing guidelines for data handling, data quality, data access as well as handling of data sources for the AI service is in place.
7	Explainability (EX)	Provide measures so that users can understand and explain decisions made by the service when necessary while depending on the sensitivity of the target application making the lack of explainability transparent to users.
8	Bias (BI)	Provide methods to assess and mitigate possible bias within the service or training data using appropriate techniques.

Table 1: Areas of the criteria catalogue with assigned objectives

4 Providing Conformity through independent Audits

Independent auditors can provide proof that an AI service complies with the AIC4 criteria. However, a simple statement of conformity has usually limited use for (potential) customers of the service. Professional cloud customers typically want to conduct their own risk analysis tailored to the specific AI use case at hand. Moreover, they usually want to implement additional security measures in their organization to complement or strengthen measures implemented at the service provider. Therefore, the customer needs reliable, detailed and transparent information regarding

- controls and processes implemented at the service provider,
- safeguards to be implemented in the customers environment to make the providers measures effective, and
- audit results and actions taken by the auditor.

Moreover, AI cloud services are usually deployed continuously, i.e. in small time intervals. A security audit should therefore not only verify that the service provider implemented suitable processes and controls at a specific point in time. It is also crucial to check that those were effective in the past. Moreover, due to the agility of AI cloud services, it is recommended that audits are repeated at least yearly and therefore they need to be conductible in reasonable small time frames. Another important aspect is modularity of AI cloud services, i.e. that an AI cloud service from one provider itself may be based on a cloud service from another provider. In those cases, the customer needs to know about crucial dependencies, which might affect the security of the service, and wants proof that security risks originating from the integration are managed.

The audit standards ISAE 3000 (Revised) in combination with ISAE 3402 and AT-C, as well as national equivalents, provide mechanisms allowing to realize the different aspects described above. Furthermore, up to now the respective standards provide the only well-established auditing methodology which can be applied straightaway for the below criteria in a timely manner. Therefore, in the following, the terminology of those standards is adapted. In order to present a possible way of applying the presented criteria, this chapter exemplifies how those audit standards are effectively applied to proof conformity with AIC4 criteria. However, it is up to the cloud customers themselves to decide what audit standards and qualifications of the auditors are considered trustworthy and what contractual details should be the content of the audit report.

4.1 Introduction

The BSI has published the Cloud Computing Compliance Criteria Catalogue (C5:2020). Within the C5:2020, the BSI describes “its view of the requirements for proof of conformity and reporting to the Cloud Service Provider and its customers”. These C5 audit-requirements refer to the execution of and the reporting on an independent assurance engagement. They are to be applied when performing an engagement according to this AI Cloud Service Compliance Criteria Catalogue - analogous to an engagement according to the BSI C5.

This chapter therefore follows closely the structure of the corresponding chapter 4 of the C5:2020 and most of the C5 audit-requirements were adopted directly. The terminology used in the C5-Audit requirements has been adjusted as necessary (for example from “Cloud Service” in the C5:2020 to “AI Service”). In addition, partial cuts and linguistic adaptations were made in order to emphasize those aspects, which are essential for use in an AI environment. Particular AI-specific additions are, as relevant, highlighted by footnotes or summarized at the end of each chapter.

For readers who are already familiar with the C5 audit requirements, Section 7.1 provides a brief overview of the particularities to be considered when transferring these requirements to an assurance engagement according to this AIC4.

In the following section, the detailed requirements are outlined regarding the methodology of an independent audit according to this AIC4.

4.2 Applicable Audit Standards

Nationally and internationally established standards shall form the foundation for proving conformity with the AI Cloud Service Compliance Criteria. Specifically the following standards prove to be adequate and suitable for such an audit: The International Standard on Assurance Engagements (ISAE) 3000 (Revised) "Assurance Engagements Other than Audits or Reviews of Historical Financial Information", the German Audit Standard (PS) 860 "IT-Prüfung außerhalb der Abschlussprüfung" of the Institut der Wirtschaftsprüfer (IDW), which is in line with ISAE 3000 (Revised), or other national equivalents to ISAE 3000 (Revised). Up to now there are no other international or national standards available proving to be equally suitable for conducting such an audit. Auditors should therefore carefully choose one of these standards or national equivalents as basis for their planning, execution and reporting.

Auditors should consider further audit standards for individual questions of execution and reporting. These include ISAE 3402 "Assurance Reports on Controls at a Service Organization", the German IDW PS 951 n.F. “Die Prüfung des internen Kontrollsystems bei Dienstleistungsunternehmen”, which is in line with ISAE 3402, or other national equivalents to ISAE 3402 or generally equivalents to such standards. Requirements for the contents of the system description provided by the AI service provider as part of the auditor’s report were derived from these standards (cf. Section 4.4.4.1).

In addition, the audit standard AT-C section 105, 205 and 320 of the AICPA, the American Institute of Certified Public Accountants, have been taken into account in this AIC4 to supplement ISAE 3402 and IDW PS 951 with requirements for the consideration of subservice organizations.

4.3 Connection to Other Audits

If the AI service provider is carrying out audits according to other international standards, the provider has already considered corresponding principles, procedures, measures and controls in its operations. It is possible that these principles, procedures, measures and controls are also to some extent relevant to an engagement according to this standard. In those cases, it makes sense to combine already existing audits with an audit according to the AIC4 in terms of organization and time. This enables auditors and AI service providers to use records in parallel for reporting according to other standards, as well as for reporting according to the AIC4.

When results obtained from other audits are used to assess the coverage of the AIC4 Criteria, the auditor shall give particular consideration to the nature of the audit and compare it with the “reasonable assurance” required for an attestation engagement or a direct engagement (cf. Section 4.4.1). In addition, it must always be assessed individually and specifically to what extent the controls and measures set up by an AI service provider cover the AI Cloud Service Compliance Criteria.

4.4 Supplementary Requirements of the BSI

The following sections outline the application of the above-mentioned audit standards.

4.4.1 Audit Engagement

As mentioned above, up to now the international standard ISAE 3000 (Revised) is the only well-established auditing methodology which can be applied straightaway for the below criteria in a timely manner. Therefore, proof of conformity shall be based on the international standard ISAE 3000 (Revised).

The ISAE 3000 (Revised) distinguishes between engagements with “reasonable assurance” and engagements with “limited assurance”. According to the BSI, reasonable assurance engagements are needed to provide conformity with the AIC4.

A distinction is also made between "attestation engagements" and “direct engagements”. Both variants are suitable for proving conformity with this AI Cloud Service Compliance Criteria Catalogue.

Further, the engagements may refer to the suitability of the design only or, in addition, the operating effectiveness of the controls. According to the BSI, the operating effectiveness is to be included in order to provide sufficient information about the AI service. Engagements solely covering the suitability of the design should only be carried out in the case of an initial engagement according to the AI Cloud Service Compliance Criteria Catalogue. As such, engagements only covering the suitability of the design are not to be recurring.

4.4.2 Criteria to be applied

4.4.2.1 Criteria for Information Security of the AI Service

The criteria[2] define the minimum scope of an assurance engagement according to the AI Cloud Service Compliance Criteria Catalogue.

The AI service provider must explain in the system description if individual criteria are not applicable due to the nature and design of the AI service. Consequently, the remaining criteria are considered to be applicable. Based on the information provided by the AI service provider, the auditor must assess to what extent the AI Cloud Service Compliance Criteria are not applicable, and if applicable, whether they are fully or partially covered by the service-related internal controls.

The applicable AI Cloud Service Compliance Criteria are to be presented in the audit report’s section containing the AIC4 Criteria, controls, test procedures and results.

4.4.2.2 Further Criteria for Transparency and Reporting

Further criteria define the information on the general conditions of the AI service (cf. Section 5) as well as the requirements concerning the system description and written statement (cf. Section 4.4.4.1 and Section 4.4.4.2; these sections also provides guidance for handling the general conditions in a direct engagement). These criteria shall enable for appropriate information about the information security of the AI service for users in order to support them with assessing the suitability of the AI service for their individual use case. The criteria also ensure comparability of the reporting in order to make it easier for users to compare several AI service providers or AI services for which an AIC4 report has been issued.

[2] In contrast to the C5 the AI Cloud Service Compliance Criteria Catalogue does not consider additional criteria.

4.4.3 Subject Matter and Objective of the Audit

4.4.3.1 Attestation Engagement

The subject of an attestation engagement is the description of the AI service provider's service-related internal controls to meet the AI Cloud Service Compliance Criteria ("system description"). The system description is to be prepared by the AI service provider. In addition, the management of the AI service provider should provide a written statement about the suitability of the design of controls to meet the applicable criteria at a specified date (type 1 report) or throughout a specified period (type 2 report), respectively. In case the engagement covers a specified period (type 2 report) the statement also covers the operational effectiveness throughout this specified period.

The objective of the engagement is to enable the auditor to provide conclusion with reasonable assurance as to whether:

- the system description fairly presents the AI service provider's service-related internal controls to meet the AI Cloud Service Compliance Criteria on a specified date (type 1 report) or throughout a specified period (type 2 report) as defined in the AIC4;
- the controls specified in the system description have been appropriately designed and implemented to meet the applicable AI Cloud Service Compliance Criteria on a specified date (type 1 report) or throughout a specified period (type 2 report); and
- where mandated (type 2 report), the controls specified in the system description operated effectively throughout a specified period.

According to the BSI, Cloud Service Providers who already have a system description can reuse it in audits according to this criteria catalogue. However, an existing system description that meets the requirements of another standard must be adapted to this criteria catalogue, as necessary.

4.4.3.2 Direct Engagement

In a direct engagement, the auditor takes stock of controls established by the AI service provider. In contrast to an attestation engagement, the AI service provider does not provide a system description. Identifying the relevant parts of the service-related internal controls takes place during the execution of the engagement.

The objective of the engagement is to enable the auditor to provide a conclusion with reasonable assurance as to whether:

- the controls established by the AI service provider were suitably designed and implemented to meet the applicable AI Cloud Service Compliance Criteria at a specified date; and,
- where mandated, the controls established by the AI service provider operated effectively throughout a specified period.

According to the BSI, the direct engagement is particularly suited for AI service providers who have not yet documented their service-related internal controls completely or in enough detail in a system description.

4.4.4 Requirements for the System Description and the Written Statement

4.4.4.1 System Description

The system description provided by the AI service provider shall include at least the following aspects:

- Name, type and scope of the AI service(s) provided;

- Description of the system components for providing the AI service including the infrastructure used for operation;
- Information on the general conditions of the AI service in accordance with the criteria in Section 5 and in Section 6 of the AIC4 that enable potential users of the AI service provider to assess its suitability for their use case;
- Applicable AI Cloud Service Compliance Criteria;
- Policies, procedures and measures as well as the controls implemented to provide (develop and operate) the AI service with respect to the applicable AIC4 Criteria;
- Dealing with significant events and conditions that represent exceptions to normal operation, such as security incidents or the failure of system components;
- Complementary customer controls assumed in the design of the AI service provider's controls and
- Functions and services with respect to the applicable AI Cloud Service Compliance Criteria provided by subservice organization including the type and scope of such functions and services, the location of processing and storage of data and requests, the complexity and uniqueness of the functions and services as well as the resulting dependency of the AI service provider and the availability of audit reports according to the criteria in the AIC4.

When auditing operating effectiveness (type 2 reporting), the following minimum items shall be appended to the system description:

- Details on significant changes to the policies, procedures and measures as well as the controls to govern the provisioning (development and operation) of the AI service with respect to the applicable AI Cloud Service Compliance Criteria that have been implemented during the period under review;
- Details on significant events and conditions that are exceptions to normal operation, that have occurred throughout the specified period and have resulted in:
    - Contractual agreements regarding the availability/performance of the AI service not being fulfilled, or major security breaches.
    - Violation of the integrity of the output of the AI service with regards to robustness against adversarial attacks, fair treatment or required explainability and of the integrity, confidentiality or availability of the AI service.[3]
    - Intentional misuse of functionality or malfunction of the AI service
    - Leakage or corruption of data or AI models.

  as well as the measures initiated by the AI service provider to prevent such events and conditions in the future.

  An incident is typically significant when it affects AI service users and the AI service provider informs the affected parties or the public. The information about the incidents and the protection measures put in place should be as transparent as possible, without revealing vulnerability or potential points of attack. Furthermore, the reporting must not jeopardize the confidentiality of information concerning individual AI service users and therefore should not contain a detailed description of individual incidents.

The system description shall not omit or distort any information relevant to the fulfillment of the applicable AI Cloud Service Compliance Criteria. This does not mean that all aspects of the service-related internal controls that can be considered important from the point of view of individual users of the AI service provider should be presented. It should be noted that the system description is intended to achieve an appropriate level of transparency for a broad range of users and that some of the processes can be customized.

[3] This bullet is an AI-specific addition that cannot be found in the C5. Note that the last two bullets found in the C5 are not included in the AI Cloud Service Compliance Criteria Catalogue.

In case of a direct engagement, the auditor shall present the above-mentioned minimum content in all material aspects as part of the audit report so that the intended users can obtain an appropriate understanding of the information security of the AI service, including the principles, procedures, measures and controls applied. This includes sufficient information on the general conditions of the AI service (cf. Section 5).

4.4.4.2 Written Statement

In the written statement, the management of the AI service provider confirms that:

- The system description fairly presents the AI service provider's service-related internal controls to meet the AI Cloud Service Compliance Criteria on a specified date (type 1 report) or throughout a specified period (type 2 report) and comprises at least the items set forth in Section 4.4.4.1 of the AIC4;
- the controls stated in the system description were suitably designed and implemented to meet the applicable AI Cloud Service Compliance Criteria as at a specified date (type 1 report) or throughout a specified period (type 2 report); and,
- where mandated (type 2 report), the controls stated in the system description operated effectively throughout a specified period.

4.4.5 Consideration of Subservice Organizations

If necessary, the AI service provider will outsource parts of its business processes for the provision of the AI service to other service providers (use of subservice organizations). The AI service provider describes this procedure in the system description and the auditor takes this into consideration as specified in the audit standard ISAE 3402. The standard offers a suitable proceeding by distinguishing between the "inclusive method" and the "carve-out method".

- Inclusive method: In case of the inclusive method, the controls and measures of the subservice organizations are to be included in the scope of the assurance engagement as well as in the reporting. In case of the inclusive method, the AI service-related internal controls of the subservice organization are also included in the system description and are in scope of the audit.
- Carve-out method: This method merely describes the services provided by the subservice organization in accordance with the minimum contents of the system description (cf. Section 4.4.4.1). The service provider's system description covers the controls that are designed and implemented to monitor the operating effectiveness of the controls at the subservice organization. The controls of the subcontractor itself are not included.

The AI service provider shall select the method to be used at its own discretion and state it accordingly in the system description (cf. Section 4.4.4.1 on Minimum Contents of the System Description).

For the purposes of the AI Cloud Service Compliance Criteria Catalogue, a service organization is a subservice organization if the following two characteristics apply:

- The services provided by the subservice organization are likely to be relevant to the users' understanding of the applicable AI Cloud Service Compliance Criteria.
- Complementary controls at the subservice organization are required in combination with the controls of the AI service provider to meet the applicable AI Cloud Service Compliance Criteria with reasonable assurance.

In case of a direct engagement, the above remarks shall be applied mutatis mutandis.

4.4.6 Assessing the Fulfillment of Criteria within an Attestation Engagement

The C5:2020 covers the case that the AI service provider already performs audits according to other standards and publications. Therefore, it is possible that the controls presented in the system description may be optimally aligned with the criteria of these standards and publications, but that their description does not fully meet all elements of the AI Cloud Service Compliance Criteria. The AI service provider shall include these controls in the system description or adjust the existing control descriptions and present these changes in an appropriate form.

An adjustment of the system description may be waived if the descriptions of the auditor's test procedures clearly state how the elements of the AI Cloud Service Compliance Criteria not covered by the control description were audited. Such test procedures shall be marked in an appropriate form (e.g. "Further test procedure for assessing full coverage of the AI Cloud Service Compliance Criterion").

This applies mutatis mutandis to a direct engagement.

4.4.7 Deviation Handling

In assessing whether applicable AI Cloud Service Compliance Criteria are not met due to identified deviations and whether the conclusion needs to be qualified, the auditor must consider the following procedures:

- Inquiry of management of the AI service provider regarding their assessment of the cause of the identified deviation;
- Assessment of the AI service provider’s handling of the identified deviation;
- Assessment whether comparable deviations have been identified by the AI service provider's monitoring processes and what measures have been taken as a result; and,
- Verification whether compensating controls and measures are in place and effective to address the risks arising from the deviation in such a way that the AI Cloud Service Compliance Criterion is met with reasonable assurance. This concerns, for example, the assessment of alternative organizational and technical approaches of the AI service provider to meet the applicable AI Cloud Service Compliance Criteria, which have not been considered in the design of the criteria set out in the AIC4.

Irrespective of the assessment as to whether a deviation leads to a qualified opinion, further information should be presented in the audit report. This information is intended to enable report recipients to assess whether the AI service provider is taking appropriate actions to handle errors and optimize its policies, procedures and actions. The following additional information shall be added by the AI Service provider to the information provided by the system description:

- If the deviation was detected by the AI service provider itself, when and in the course of which measures the deviation was detected.
- If the deviation was already stated in a report of a previous audit, an indication should be given of when and by what means the deviation was detected, together with a separate indication that the detection occurred in a previous audit period. This requires that the auditor has access to prior reports from the AI service provider. In case of doubt, the auditor shall have the inspection of these reports separately assured in his engagement letter.
- The measures to be taken to remedy the deviation in the future and when these measures are likely to be completed or effectively implemented.
- When issuing a type 2 report, the time period for which the deviation was in place should be specified.[4]

[4] This bullet point is an AI-specific requirement.

This additional information is not subject of the audit, and, accordingly, the auditor does not express an opinion thereon.

4.4.8 Reporting

The report on an attestation engagement should contain the following elements:

1. Independent auditor's report
    a. Scope and version of the AIC4
    b. The AI service provider’s responsibility
    c. Independence and quality control of the auditor/auditing firm (including information on compliance with qualification requirements, cf. Section 4.4.9)
    d. Auditor ‘s responsibility
    e. Inherent limitations
    f. Audit Opinion
    g. Intended users and purpose
    h. General terms of the engagement;
2. Written statement by the AI service provider's management responsible for the AI service(s);
3. Description of the service-related internal controls established by the AI service provider to meet the AI Cloud Service Compliance Criteria;
4. Presentation of the applicable AIC4 Criteria, the associated controls (part of the system description), test procedures performed and the individual test results of the auditor;
5. Optional: Other information provided by the AI service provider (this information is not subject of the audit, and, accordingly, the auditor does not express an opinion thereon).

In the case of a direct engagement, these are applied mutatis mutandis. The reporting on an attestation engagement is based on the requirements of ISAE 3402.
In case of a direct engagement, the components 2 ‘Written statement' and 3 ‘System Description' are omitted. Nevertheless, the minimum content of the system description mentioned in Section 4.4.4.1 shall be presented in all material respects in the audit report so that the intended users can obtain an appropriate understanding of the information security of the AI service, including the principles, procedures, measures and controls applied as well as sufficient information on the general conditions of the AI service. Such information shall be provided in a separate section, e.g. "Description of the AI service and the policies, procedures and measures applied by the AI service provider".

The test procedures performed shall be described for both the suitability of design (type 1 and type 2 report) and the operating effectiveness (type 2 report only) engagements.

4.4.9 Qualification of the Auditor

According to ISAE 3000 (Revised), the auditor must determine before accepting an engagement that the professional duties (for auditors in Germany § 43 WPO, German Law regulating the Profession of Wirtschaftsprüfer: Wirtschaftsprüferordnung), including the duty of independence, are complied with. Based on the auditor's knowledge of the subject matter, the auditor shall assess whether the members of the audit team entrusted with the engagement have the necessary competency and understanding of the industry as well as capabilities to perform the audit and whether sufficient experience with the relevant formal requirements is available or can be obtained.

According to the BSI, audits based on the AI Cloud Service Compliance Criteria Catalogue place special requirements on the qualification of the auditor and the members of the audit team. From the BSI's point of view, the following aspects on professional qualifications and professional experience are suitable indications that these special requirements are met.

They have to be fulfilled by those members of the audit team who, according to the International Standard on Quality Control (ISQC) 1 "Quality Control for Firms that Perform Audits and Reviews of Financial Statements, and Other Assurance and Related Services Engagements" or the German IDW quality assurance standard "Anforderungen an die Qualitätssicherung in der Wirtschaftsprüferpraxis" 20 Federal Office for Information Security Providing Conformity through Independent Audits 4 (IDW QS 1) or other national equivalents of ISQC 1, supervise the execution and review the results of the engagement (including evaluation of the work performed, review of the documentation and the planned reporting):

- 3 years relevant professional experience with IT audits in a public audit firm

or one of the following professional examinations/ certifications;

- Information Systems Audit and Control Association (ISACA) – Certified Information Systems Auditor (CISA), Certified Information Security Manager (CISM) or Certified in Risk and Information Systems Control (CRISC);
- ISO/ IEC 27001 Lead Auditor or BSI certified ISO 27001 Auditor for audits based on BSI IT Grundschutz;
- Cloud Security Alliance (CSA) – Certificate of Cloud Security Knowledge (CCSK);
- (ISC)^2 – Certified Cloud Security Professional (CCSP).

For the members of the engagement team who conduct the audit on a technical/operational level the BSI recommends that at least one member, who is involved in the actual testing of the AI service, has at least 3 years relevant professional experience as data scientist or as developer of machine learning models.[5]

It is recommended, at the client's request, that the auditor provides appropriate evidence that the audit team meets the qualification requirements.

Compliance with the qualification requirements shall be confirmed in the section "Independence and quality control of the auditor/auditing firm" of the independent auditor's report.

4.5 Handling of Updates for the AI Cloud Service Compliance Criteria

The BSI intends to update the AIC4 regularly in line with general technical developments and the ongoing development of the underlying standards.

In this context, AI service providers and auditors shall have sufficient time to make the necessary adjustments to the processes, systems and controls and to the execution of the audit by the auditors associated with the updates of the AIC4.

According to the BSI, the adjustments to processes, systems and controls and the audit must be considered within audits with a specified date (in case of a type 1 engagement) or period end date (in case of a type 2 engagement) that are set 12 months after the new version has been published. Any deviations from this must be justified within the report.

Hence, since updates of the AIC4 must be considered within 12 months, it may happen that the assessment of the effectiveness of the processes, systems and controls applied by the AI service provider relates both to the status before and after the implementation of such adjustments. The system description should include the adjustments made (cf. Section 4.4.4.1). In the case of a direct engagement, the auditor must obtain and disclose this information.

[5] This is an additional requirement on top of C5.

If the audit period ends between six and twelve months after the publication of the updated AIC4, the AI service provider shall provide additional information in the system description regarding the necessary changes to its service-related internal controls which have not been completed. The details should include what measures are to be completed or effectively implemented. In the case of a direct engagement, the auditor shall obtain and disclose this information.

5 Information on the General Conditions of the AI Service

The purpose of this section is to specify the minimum contents that an AI service provider must include in the system description by defining the general condition of the AI service.

BC-01 System Description

Information on the General Conditions of the AI service

In the system description, the AI service provider sets out precise and comprehensible specifications regarding the AI service. Goals, design and application of the AI service are documented. Policies and guidelines for the provision of the AI service are outlined.

The system description covers at least the following general aspects:

- System specifications for the compatibility of AI model(s) within the scope of the AI service and how it is integrated into general IT systems;
- Regulatory and legal requirements as well as international standards applied for the AI service itself or related data;
- Description of the infrastructure, network and system components used for development and operation of the AI service as well as measures taken to ensure the integrity of the latter;
- Complementary responsibilities of the user and subservices.

Further information is provided in the criteria areas in Section 6.

BC-02 Security & Robustness

Information on the General Conditions of the AI service

As part of the system description (BC-01), the AI service provider states comprehensive information allowing a (potential) AI service user to understand the suitability of robustness and security measures for the specific AI service.

The information provided covers at least the following aspects:

- Procedure for the measurement and quantification of robustness;
- Level of robustness the AI service provider guarantees and why it is sufficient for the service at hand;
- Limits of the robustness of the AI model(s) within the scope of the AI service.

Further information is provided in the Security & Robustness criteria in Section 6.2.

BC-03 Performance & Functionality

Information on the General Conditions of the AI service

As part of the system description (BC-01), the AI Service provider presents comprehensive information allowing a (potential) AI service user to understand and evaluate the suitability of the performance and functionality for intended use.

The information provided includes at least the following aspects:

- Definition of objectives, impact and purpose of the AI service;
- Procedures and measures implemented to develop and operate the AI model(s) within the scope of the AI service;
- Performance measures used to evaluate the AI model(s) within the scope of the AI service;
- Selection of the implemented AI method (algorithms and data processing mechanisms) and an explanation why it is suited for the target application. Outline of the limitations and assumptions of the model;

- Training frequency of the AI model(s) within the scope of the AI service. Information whether it is a continuous learning system or whether there are defined learning cycles;
- Functionalities of the AI model(s) within the scope of the AI service. This includes a description of the task to be solved and inputs and outputs;
- Degree and potential impacts of automated decision making;
- The extent to which users are able to correct or object to the results or decisions made by the AI service;
- Details on significant changes made during the audit period to procedures, controls and measures concerning the AI service. This includes changes made to the AI model(s) within the scope of the AI service itself (e.g. retraining, model change).

Further information is provided in the Performance & Functionality criteria in Section 6.3.

BC-04 Reliability

Information on the General Conditions of the AI service

As part of the system description (BC-01), the AI service provider presents comprehensive information allowing a (potential) AI service user to understand and evaluate the reliability of the AI service, taking the provisioning of resources and incident handling into account.

The information provided contains at least the following aspects:

- Logging carried out during operation. This includes an overview of the content kept in logs as well as storage periods and usage;
- Handling of significant incidents and conditions that lead to exceptions to regular operations. This includes a definition of such incidents and conditions as well as implemented safeguards and disaster recovery management;
- Location, duration and responsibilities for storing and processing involved data and models.

Further information is provided in the Reliability criteria in Section 6.4.

BC-05 Data Quality & Data Management

Information on the General Conditions of the AI service

As part of the system description (BC-01), the AI service provider states comprehensive information allowing a (potential) AI service user to understand and evaluate the suitability of the data quality and data management for the AI service.

This information provided covers the following aspects:

- Description of data sources used for training and operation of the AI service;
- Roles and responsibilities assigned to business functions of the AI service provider regarding access and use of data;
- Description of the data selection;
- Description of the performed data pre- and post-processing steps.

Further information is provided in the Data Quality and Data Management criteria in Sections 6.5 and 6.6.

BC-06 Explainability and Bias

Information on the General Conditions of the AI service

As part of the system description (BC-01), the AI Service provider states comprehensive information allowing a (potential) AI service user to understand the degree of explainability and potential sources of bias for the AI service.

This information covers the following aspects:

- Explainability of the AI model(s) within the scope of the AI service. This includes a description of the level of explainability and, if present, the parts of the AI model(s) that are not explainable;
- Technical limitations of used methods and shortcomings regarding the identified needs for explainability;
- Possible effects of bias that may impact the functionality of the service in a critical way are outlined;
- Metrics and tolerance intervals for assessing bias are outlined;
- Critical bias currently not mitigated are outlined.

Further information is provided in the Explainability as well as the Bias criteria in Sections 6.7and 6.8.

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 selecting 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 users 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

-

7 Key Concepts & Glossary

This section includes an overview that outlines key differences between chapter 4 in the AIC4 and chapter 4 of the C5:2020. A detailed glossary explaining all the relevant definitions mentioned in the catalogue to enable a broad understanding of the reader is also provided.

7.1 Adaptation of the C5-Audit Methodology for AI services

For the sake of completeness, the comprehensive explanation of the methodology and amendments made to the C5 can be found subsequently.

- “Cloud Service” is replaced with “AI service”
- “C5 Criteria” is replaced with “AI Cloud Service Compliance Criteria”
- Section 4.3 is applicable but explicit references to other audit standards have been removed as well as references to the respective mapping to other standards.
- Section 4.4.2 is applicable, however refers to the general requirements an AI service must fulfill with respect to security, reliability, data quality and bias assessment and mitigation.
- Section 4.4.2.1: the AI Cloud Service Compliance Criteria Catalogue does not include the concept of additional criteria.
- Section 4.4.4.1 outlines additional information on type 2 reporting in the second paragraph. Bullet point two outlines details of “significant events and conditions that are exceptions to normal operation, that have occurred throughout the specified period and have resulted in [...]”. The following two sub bullet points in the C5 are removed and replaced with: “violation of the integrity of the outputs of the AI service with regards to robustness against adversarial attacks, fair treatment or required explainability”.
- Section 4.4.7: in the second paragraph an additional bullet point should be added: “When issuing a type 2 report, the time period for which the deviation was in place should be specified”.
- Section 4.4.9. outlines details of the qualification for the auditor. The following should be added: “For the members of the engagement team who conduct the audit on a technical/operational level the BSI requires that at least one member, who is involved in the actual testing of the AI service, has at least 3 years relevant professional experience as data scientist or as developer of machine learning models“.

7.2 Glossary

The following definitions aim to provide clarity to the terminology used for the specification of both a process and control audit as well as a technical/functional audit.

The general definitions (see Global section) mainly refer to the Stanford Encyclopedia of Philosophy (SEP). The SEP is a scientifically acknowledged base for comparable questions in other fields. Several publications of the European Commission, including the “Ethics Guidelines for trustworthy AI” and the whitepaper on Artificial Intelligence (see section 4.1 for detailed sources) provide further detail as well.

Four different sources are mainly used as a reference for audit-related terminology, including the ISAE 3000, C5, the Common Criteria (CC) catalogue and the “Glossary of Terms” provided by the International Auditing and Assurance Standards Board (IAASB). More specific definitions reflect the technical terms of current scientific research.

Global

Term	Description
AI Lifecycle	defines the steps an organization follows to leverage AI.
AI service provider	Individual(s) or organization that develops/deploys/operates/uses AI services excluding end-users or consumers.
AI service	cloud service utilizing AI methods deployed either in a public or private cloud infrastructure.
Artificial Intelligence (AI)	a method that enables a computer to solve problems that, if done by a human, would require intelligence – such as performing certain complex tasks (visual perception or speech recognition) that involve acquisition, processing and rational analysis of data.
Automated	ability to act without direct human control.
Bias	systematic tendency or error in the process of data processing, which results in misleading results.
Black Box Model	inputs and outputs are known, however there is little to no knowledge about the inner workings of the model.
Data Quality	concept to ensure that data points are in good shape for the intended use.
Deep Neural Network	an artificial neural network built of multiple layers between the input and output layers.
Disaster Recovery Management	stands for implementing sufficient protocols and systems that ensure a company recovers fast after a disastrous incident, for instance loss of access due to malware. This is often done by third parties.
Documentation	is the process of writing down the details of the characteristics of the datasets, models and processes that an AI service goes through, from design, to development, to the deployment. It separates between: AI Service Design and Setup Stage: service framing and high-level objective design. This includes pondering the motivation developing the AI service and defining the goals of the AI service, as well as determining team priorities and objectives throughout the AI service design process. AI Development: building a thorough and well-documented overview of the elements of the AI development pipeline - from the data used to train the model to the internals of the service architecture and output characteristics. AI Deployment: testing and examining different ways to assess a service's effectiveness in achieving the desired goals, while keeping undesirable side effects in the output minimal. AI Operation: integrate AI projects with existing applications and processes successfully within the entity and manage the complete end-to-end lifecycle of AI. AI Service Maintenance and Monitoring: documenting the continued functionality of the service and maintaining quality in service performance. Service Feedback: details of the elements that most impacted the service's deployment should be documented along with the key decisions that were made to ensure the well continued functioning of the service.
Explainability	is the extent to which the algorithms of a machine or deep learning model can be explained in human terms. Intrinsic explainability means

Term	Description
	that the reasoning of a model is understandable for humans, whereas post hoc explainability methods are applied to trained models to achieve explainability.
Explanation	reasoning for an action/prediction that is human understandable.
Functionality	describes principles that ensure usability and comprehensiveness of an AI service throughout its lifecycle.
Gradient Boosting Algorithm	machine learning technique for classification and regression problems that produces a prediction model in the form of an ensemble of other prediction models
Grey Box Model	combines both the Black Box and White Box Model approach in order to unite the respective advantages.
Intelligence	the capacity to recognize patterns.
Interpretability	describes the extent to which a cause and effect can be observed within a model.
Learning	group of techniques that refers to, among others, machine learning, neural networks and deep learning. Enabler for AI services to learn how to solve problems that cannot be precisely specified or whose solution method cannot be described by symbolic reasoning rules.
Model	A model is a defined input-output function that takes a set of inputs (i.e. “features”) and provides a prediction for the expected output (i.e. “label”) for that input, based on the learned past relationship of previous input-output pairs it has been exposed to directly or indirectly.
Prediction	refers to the output of an AI-algorithm after it has been trained on historic data and applied to new data to forecast a target.
Privacy	is concerned with the interest of individuals in exercising control over access to information about themselves and is most often referred to as “informational privacy”.
Random Forests	ensemble learning method for classification, regression and other tasks that consists of a large number of individual decision trees that operates as an ensemble.
Reasoning & Decision Making	group of techniques that refers to knowledge representation, reasoning, planning, scheduling, search as well as optimization and allows to perform the reasoning on the data coming from the sensors.
Reliability	aims at ensuring that an AI service functions within its expected behavior.
Responsibility	assumption of consequences that arise with an action. This assumption can result in blame or praise of the action which can eventually (depending on the case) lead into punishment.
Robustness	the ability to perform well under a certain level of uncertainty.
Safety	defines the ability of a system to protect its users from harmful or non-desirable outcomes.
Security	is a state of a system achieved by being protected against malicious planned activities.
Transparency	aims to provide information that allow users, practitioners or other stakeholders to understand the goals, origins and form of an AI service.

Term	Description
Trust	a set of specific beliefs dealing with benevolence, competence, integrity and predictability; the willingness of one party to depend on another in a risky situation; or the combination of these elements.
White Box Model	refers to analytical and physical descriptions, the modeling of which is usually very complex.

Table 2: Terminology Global

Security & Robustness

Term	Description
Adaptive Attacks	the adversary has access to or knowledge about the defense in place and can adapt his attack accordingly.
Adversarial Defense	a countermeasure put in place by subject matter experts to defend against adversarial attacks.
Adversarial Examples	intentional inputs which aim at diminishing the performance of machine learning model while being in close similarity to training data difficult to detect.
Adversarial Perturbation	a small malicious change carefully crafted in an input to fool the model.
Data Corruption	relates to errors in computer data that occur, among others, during writing, reading or processing steps and cause unintended changes to the original data.
Data Poisoning	attacks that include the manipulation of the training dataset used by the machine-learning model.
Integrity of Data	ensures the accuracy and consistency of data over its entire lifecycle and its maintenance.
(Machine) Learning Pipeline	integral component for the improvement of the performance of AI model(s) involving acquiring and processing of data as well as making and validating predictions.
Membership Attack	is an attack with which an attacker can establish whether a given individual's data were in a training set or not.
Model Stealing/theft	is an attack where the adversary finds a way to access where the model is stored and secretly duplicates it.
Privacy Attacks	are actions designed to breach the service and extract private data. Examples of such attacks can be model inversion attacks and membership inference attacks.
Privacy-Preserving Machine Learning Techniques	are techniques that prevent the extraction of privacy-relevant information from the AI service and for instance, allow multiple input parties to collaboratively train ML models without releasing their private data in its original form.
Proactive Defense	consider protective measures during the development process and incorporate these directly into the model.
Reactive Defense	catch adversarial examples after model development.

Term	Description
Risk Exposure Assessment	systematic estimation on how vulnerable the service is to possible weaknesses of the AI model regarding leakage or corruption of data and compromising the integrity and confidentiality of the model and to determine its degree of robustness with respect to the model and data.
Security Violation	is an incident that can happen when an attacker aims to get malicious input misclassified as legitimate which is an integrity violation or augment the wrong classification rate if he aims for an availability violation which can render the model as unusable. Examples of such attacks can be adversarial attacks.
(Data) Tampering	intentionally altering data (editing, manipulating, destroying) through unauthorized channels.
Threat Scenario	describes the extent by which potential hazards are identified from a hypothetical attacker's point of view. It consists of the attacker's goals (targeted or untargeted misclassification or confidence reduction), knowledge (white-box or black-box) and perturbation space (Distance norm for image, character, word or phrase level for NLP, ...).

Table 3: Terminology Security & Robustness

Performance and Functionality

Term	Description
Automated Machine Learning	the automated training and testing of various models to find a good solution.
Benchmarks	can be understood as standardized tasks with predefined datasets.
Concept Drift	A model built on old data becomes inconsistent with new input data and requests and requires updates since the concept of the target variable has changed over time.
Hyperparameters	parameters of the model set by the subject matter expert.
Maintenance	process of monitoring and adjusting the model after the developing phase in order to ensure usability/applicability of the service in its environment.
Model Change/Evolution	changes in the AI pipeline or model parameters after adjusting the AI model(s) in scope of the AI service for example after retraining.
Model Deployment	putting an approved model into production.
Model Development	the process of developing an AI model. This involves data preparation, model selection, training and validation.
Model Drift	Changes in the environment over time can cause a model's performance to degrade due to violated model assumptions. This makes retraining necessary.
Model Parameters	parameters of the model learned from data.
Model Retraining	the process of adjusting parts of the developed model after deployment/approval.

Term	Description
Model Training	a process, in which model parameters are constantly updated in order to improve the model's capability to approximate a target function by only observing input/output-pairs of that function.
Model Validation	the process of testing the model on unseen data to assess and compare its performance.
Overfitting	a training state in which the model encodes input/label-pairs directly into model parameters, which drastically restricts the model's ability to generalize.
Performance Metric/Accuracy	a measure to assess the degree a given AI service contributes to the solution of the underlying problem. Depending on the application this measure can differ.
Production Environment	environment where the ML model can provide predictions to other systems.
Reproducibility	ability of an AI model(s) in scope of the AI service to generate the same output using a given input at different points in time.
Scalability	Exists in several different manners. These include: Horizontal Scalability: Adding more machines will lead to a reasonable increase in processing time of the AI service. Vertical Scalability: Adding more power (CPU/RAM) to existing machines will lead to a reasonable increase in processing time of the AI service. Algorithmic Scalability: describes whether the training or prediction of a ML algorithm is the more computation complex task. (Eager vs Lazy Learners).
Suitability	the capability of an AI model to achieve acceptable performance (especially regarding the risk of failure) with respect to defined requirements for a given problem.
Timeliness	The machine learning pipeline is able to process/predict new data in a predefined/required time window.
Underfitting	occurs when a model or algorithm is unable to represent the relationships between a dataset's features and a target variable due to a lack of complexity.
Unit Testing	testing of different well separated parts of the AI service independently from each other by comparing predefined (expected) outputs and generated outputs.

Table 4: Terminology Performance and Functionality

Reliability

Term	Description
Action/Decision	event triggered by a prediction made by the model.
Failure	results/actions of the model that are considered to be incorrect.
Logging	Information necessary for a forensic analysis are saved for a later reference.

Term	Description
Monitoring	process of keeping track of the AI service over time.
Noise	disruptive component in the data hiding the underlying signal.
Safeguard	functionality that alerts when unusual behavior is monitored.

Table 5: Terminology Reliability

Data Quality

Term	Description
Accessibility	extent to which relevant data is available or easily and quickly collectable.
Completeness	data points in a set are exhaustive and uncorrupted. In addition, the set itself is a valid representation of the ground truth.
Cross Validation	evaluation technique for ML models which trains several models on different subsets of the available input data and evaluates them on another subset of the data.
Data Accuracy	degree to which data correctly captures the “real-life” objects/phenomena they are intended to represent.
Data Annotation	refers to labeling data including different data types like text, images or videos.
Data Cleansing	preparing data for analysis by removing or modifying incorrect, incomplete, irrelevant, duplicated or improperly formatted data.
Data Creation (Sampling)	Artificial data creation can be necessary to generate data sets meeting specific needs or conditions which are unavailable in real-world data sets.
Data Hierarchy	organizing data systematically, often in a hierarchical approach.
Data Lineage	tracks the data flow through the company to its source system.
Data Ontology	enables to explain the properties of a subject area and their relation to each other by outlining a set of concepts and categories that describe the subject.
Data Ownership	defining responsibility levels for the data.
Data Provenance	log file that tracks the data from recording to its present state.
Data Quality Assessment	applies data quality metrics and objective judgment to verify/ensure data quality over the data and time.
Data Quality Metric	measure that describes/assesses problems within the collected data.
Data Validation System	set of implemented actions/rules that test data on their plausibility.
Dimensionality	number of attributes included in a dataset.
Feature Engineering	refers to the process of transforming raw data into features using domain knowledge.
Least-privilege Principle	refers to the concept that any user/subject matter expert/program/process should only have the minimum required access rights to perform the task.

Term	Description
Metadata	refers to a data set describing and informing about other data.
Need-to-know Principle	a subject matter expert only gets access to certain data if it is essential to conduct the task.
Relevance	The combination of variables can be used to predict another variable and reveals underlying relationships between them.
Systematic Data Checks	assessing the plausibility of the data taking data quality rules and subjective judgment into account.
Test Data	data set used to assess the working accuracy of the model and its ability to generalize. This data was not presented to the model before.
Traceability	extend to which data source(s) can be tracked through an organization's IT landscape/ infrastructure.
Training Data	data set the model is trained on by adjusting the parameters.
Validation Data	data set used during training to assess the training model and to avoid overfitting.
Verifiability	the ability to cross check obtained data points on plausibility/consistency with other sources.

Table 6: Terminology Data Quality

Explainability

Term	Description
Feature Importance	indicates the value of each feature in the construction of the model.
Saliency Map	picture that displays every pixel's specific quality, aiming to alter the representation of the image into something easier to analyze.
Sensitivity	indication how minor changes in input influence the output of the model.

Table 7: Terminology Explainability

Bias

Term	Description
Bias Mitigation Method	technique that reduces the measured bias of an AI service.
Fairness Metric	a quantification of bias in training data or models.
Type of Bias	refers to different sources of bias which can be categorized and have an impact on the functionality of the AI service.

Table 8: Terminology Bias

Audit-related Terminology

Term	Description
Assess	analyze identified risks of to conclude on their significance.
Assurance Engagement	an engagement in which a practitioner aims to obtain sufficient appropriate evidence in order to express a conclusion about the subject matter information.
Attestation Engagement	an assurance engagement in which a party other than the practitioner measures or evaluates the underlying subject matter against the criteria.
Comprehensive	complete and including everything that is necessary.
Confirm	declare that something has been reviewed in detail with an independent determination of sufficiency.
Consider	to give attention to a particular subject or fact when judging something else.
Criteria	the benchmarks used to measure or evaluate the underlying subject matter. The “applicable criteria” are the criteria used for the particular engagement.
Describe	provide specific details of an entity.
Define	explain and describe the meaning and exact limits of something.
Determine	affirm a particular conclusion based on independent analysis with the objective of reaching a particular conclusion.
Direct Engagement	an assurance engagement in which the practitioner measures or evaluates the underlying subject matter against the applicable criteria and the practitioner presents the resulting subject matter information as part of, or accompanying, the assurance report.
Document	the record of work performed, results obtained, and conclusions the practitioner reached.
Ensure	guarantee a strong causal relationship between an action and its consequences.
Evaluate	Identify and analyze the relevant issues, including performing further procedures as necessary, to come to a specific conclusion on a matter.
Evidence	information used by the practitioner in arriving at the practitioner’s conclusion. Evidence includes both information contained in relevant information systems, if any, and other information.
Explain	give argument accounting for the reason for taking a course of action.
Exploit	a procedure designed to take advantage of a flaw in an AI service, typically for malicious purposes.
Identify	to recognize a problem, need or fact and to show that it exists.
Implement	to put a plan or system into operation.
Limited Assurance Engagement	an assurance engagement in which the practitioner reduces engagement risk to a level that is acceptable in the circumstances of the engagement but where that risk is greater than for a reasonable assurance engagement.
Measure	verb, to judge the quality, effect, importance or value of something.

Term	Description
Monitor	watch and check a situation carefully for a period of time in order to discover something about it.
Outline	to give the main facts about something.
Package	named set of either security functional or security assurance requirements.
Preparation	activity in the lifecycle phase of a product, comprising the user's acceptance of the delivered application and its installation which may include such things as booting, initialization, start-up and progressing the application to a state ready for operation.
Prioritize	to arrange things, tasks etc. in order of importance in order to deal with the most important things before the others.
Production	production lifecycle phase follows the development phase and consists of transforming the implementation into a state acceptable for delivery to the user.
Prove	show correspondence by formal analysis in its mathematical sense.
Provide	to give something that is needed or wanted to someone.
Reasonable Assurance Engagement	an assurance engagement in which the practitioner reduces engagement risk to an acceptably low level in the circumstances of the engagement as the basis for the practitioner's conclusion.
Reflect	to think carefully, especially about possibilities and opinions.
Satisfy	to have or provide something that is needed or wanted.
Service Level Agreements	are signed between a service provider and a customer agreeing on a certain level of service, for instance quality, availability or responsibility.
Specify	provide specific details about an entity in a rigorous and precise manner.
Subject Matter Experts	person who has special knowledge and/or skills to work on a particular task, topic or job and fulfills all requirements to work in the environment of the AI service (legal, technical, etc.). For instance, during model development, subject matter experts consist of data scientists and software engineers or persons with similar skills, while during regular operation of AI services subject matter experts consists of system administrators or specialists for application operations.
Subject Matter Information	the outcome of the measurement or evaluation of the underlying subject matter against the criteria, i.e., the information that results from applying the criteria to the underlying subject matter.
Test	do something in order to discover if something is safe, works correctly or if something is present.
Validate	to make something officially acceptable or approved, especially after examining it.
Verify	rigorously review in detail with an independent determination of sufficiency.
Vulnerability	weakness in the AI service that can be used to violate it in some environments.

Table 9: Audit-related Terminology