EIDGENÖSSISCHE FINANZKONTROLLE
CONTRÔLE FÉDÉRAL DES FINANCES
CONTROLLO FEDERALE DELLE FINANZE
CONTROLLA FEDERALA DA FINANZAS
SWISS FEDERAL AUDIT OFFICE

Artificial Intelligence Audit

Guide

TABLE OF CONTENTS

Version 1.00, as of 20 May 2026

1. INITIAL SITUATION

The integration of artificial intelligence (AI) into administrative processes offers significant opportunities for the Federal Administration. At the same time, it introduces new challenges relating to economic viability, reliability, security and public trust. Investment in AI applications should be supported by demonstrable benefits that can be realised within a reasonable timeframe. In sensitive areas of application, AI systems must also be developed to the highest quality standards in order to maintain public trust.

AI applications based on machine learning rely on complex algorithms with a large number of parameters and a high demand for computing capacity. The associated investments in infrastructure and expertise are considerable. At the same time, the complexity of these algorithms makes it difficult to understand how such systems arrive at their decisions, giving rise to risks such as the reinforcement of societal bias or the lack of explainability of automated decisions. Ensuring the responsible and transparent use of AI therefore requires assessments that address a wide range of risk areas. This helps to ensure that AI applications operate efficiently while also complying with legal and ethical requirements.

The regulatory framework governing the use of AI within the Federal Administration is based on two key documents. The 'Artificial Intelligence' Guidelines for the Federal Government (2020)[1] set out fundamental requirements regarding transparency, traceability and non-discrimination. The Federal Government recognises its responsibility to establish the necessary framework for the safe use of AI by implementing technical and organisational measures that prevent erroneous decisions, minimise discrimination and ensure compliance with data protection requirements. In addition, the Sub-strategy on the Use of AI Systems in the Federal Administration defines three strategic areas of action — Developing skills, earning trust and increasing efficiency – thereby setting out the strategic priorities for the use of AI in the Federal Administration.[2]

Against this background, this guide provides a flexible and pragmatic framework for the risk-based assessment of AI applications within the Federal Administration. It enables relevant aspects to be assessed systematically throughout the entire life cycle of an AI application. The guide is deliberately designed to be modular: it is meant to be adapted to the specific characteristics and risks of the respective audit and combined with existing audit methods, such as project auditing.

---

1 See: Swiss Confederation. (25 November 2020). 'Artificial Intelligence' Guidelines for the Federal Government: A framework for the use of artificial intelligence in the Federal Administration.

2 See Federal Chancellery. 'Digital Transformation and ICT Steering (DTI)' division. (11 December 2025). SB021 – Strategy Use of AI Systems in the Federal Administration. Retrieved on 7 April 2026, https://www.bk.admin.ch/en/sb021-strategy-use-of-ai-systems-in-the-federal-administration.

2. OVERVIEW

2.1 Risk dimensions and risk areas

The Swiss Federal Audit Office's (SFAO) methodology for auditing AI applications within the Federal Administration is based on a structured framework that focuses on the key risk dimensions and the different phases of an AI application's life cycle (see Chapters 3 to 5). Each chapter addresses one of the three principal risk dimensions and explains its relevance across the various phases of the AI life cycle. The three risk dimensions are trustworthiness, cost-effectiveness and competencies.

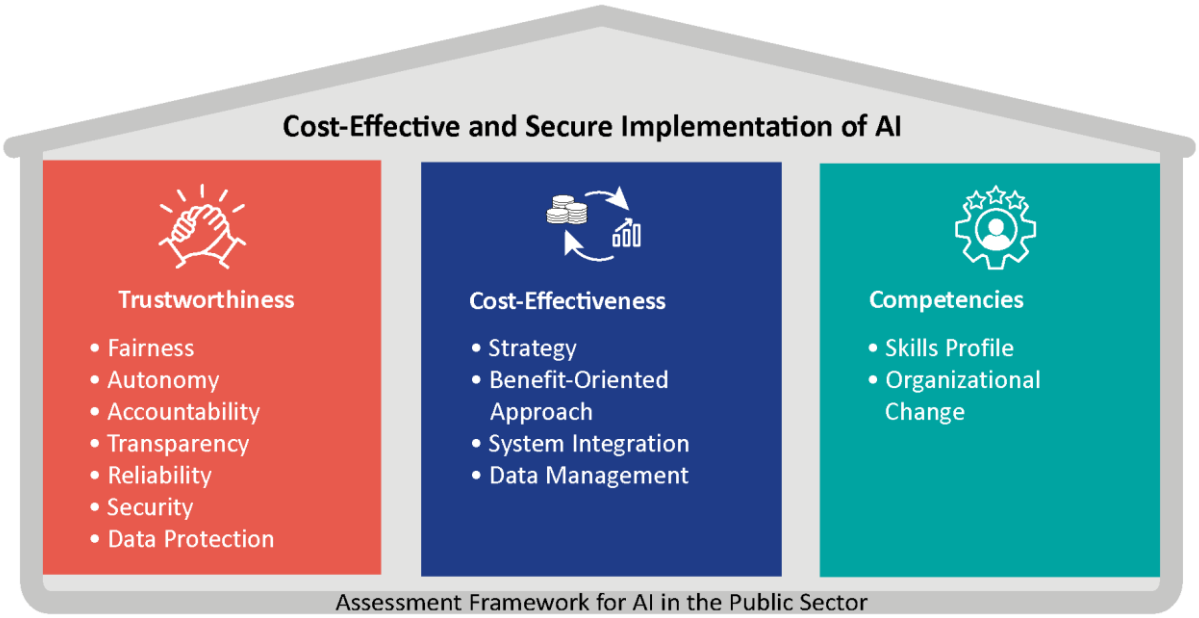

Figure 1 : Overview of the risk dimensions and areas of artificial intelligence.

Each dimension is further divided into individual risk areas, allowing specific aspects of the respective dimensions to be assessed in a structured and differentiated manner. This approach highlights the distinctive characteristics of each risk dimension (see Figure 1).

The selection of these three risk dimensions is based on the **Sub-strategy on the Use of AI Systems in the Federal Administration**, adopted by the Federal Chancellery's Digital Transformation and ICT Steering (DTI) division. The sub-strategy identifies three strategic areas of action: **Developing skills, earning trust and increasing efficiency.** It thus forms the basis for the present audit framework. The guide's three risk dimensions are deliberately aligned with these:

Trustworthiness addresses the inherent risks of AI systems with regard to fairness, transparency, robustness, and data protection – aspects that carry weight in the public sector, since decisions made by the federal administration can affect fundamental rights and public trust. **Cost-effectiveness** ensures that the use of AI corresponds to a demonstrable benefit and that the associated costs and risks are in a reasonable proportion to one another. This benefit can take various forms: direct cost reductions, increased processing volume, shortened turnaround times, quality improvements, relief for scarce specialist staff, or improved compliance and resilience. Sustainability aspects must also be taken into account here, and these are explored further in the risk areas of security, strategy, data management, and organizational

change. **Competencies** capture the organizational and personnel foundation without which even technically mature AI applications can fail.[3]

ONGOING EXAMPLE

To illustrate the concepts presented throughout this guide, a consistent practical example drawn from the Federal Administration is used for each life-cycle phase and risk area.

The example demonstrates how an AI-based system can support recruitment decisions and highlights the principal risks that should be considered throughout the AI application's life cycle and across the different risk areas.

2.2 Risk Phases

In addition to the substantive structure of the risk dimensions, the temporal aspect of AI development is also taken into account. The life cycle of an AI application differs fundamentally from that of conventional IT systems: AI models are developed, trained and adapted dynamically and iteratively, enabling them to continuously adapt to new data and changing circumstances through machine learning. Whilst traditional IT systems perform fixed functions, AI applications can develop new patterns and capabilities over the course of their lifecycle, which are often difficult to predict. In the literature, the lifecycle of an AI application is typically divided into five phases, each of which presents specific challenges and risks.[4] For the purposes of this guide, these five phases are grouped into three audit-relevant **risk phases**: **planning**, **development** and **operation**. This classification enables a structured and comprehensive risk assessment without requiring in-depth technical knowledge, and it clearly maps out the key transitions and responsibilities from an audit perspective. The life cycle is not a linear process, but an iterative cycle: insights gained during operation feed directly back into the planning phase and may trigger a new development or fundamental adaptation. The three phases are therefore closely interlinked and mutually dependent (seeFigure 2).

[3] See Federal Chancellery. 'Digital Transformation and ICT Steering (DTI)' division. (11 December 2025). SB021 – Strategy for the Use of AI Systems in the Federal Administration. Retrieved on 7 April 2026, https://www.fch.admin.ch/fch/de/home/digitale-transformation-ikt-lenkung/vorgaben/sb021-strategie-einsatz-von-ki-systemen-in-der-bundesverwaltung.html.

[4] OECD (2023), "Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI", OECD Digital Economy Papers, No. 349, OECD Publishing, Paris, https://doi.org/10.1787/2448f04b-en.

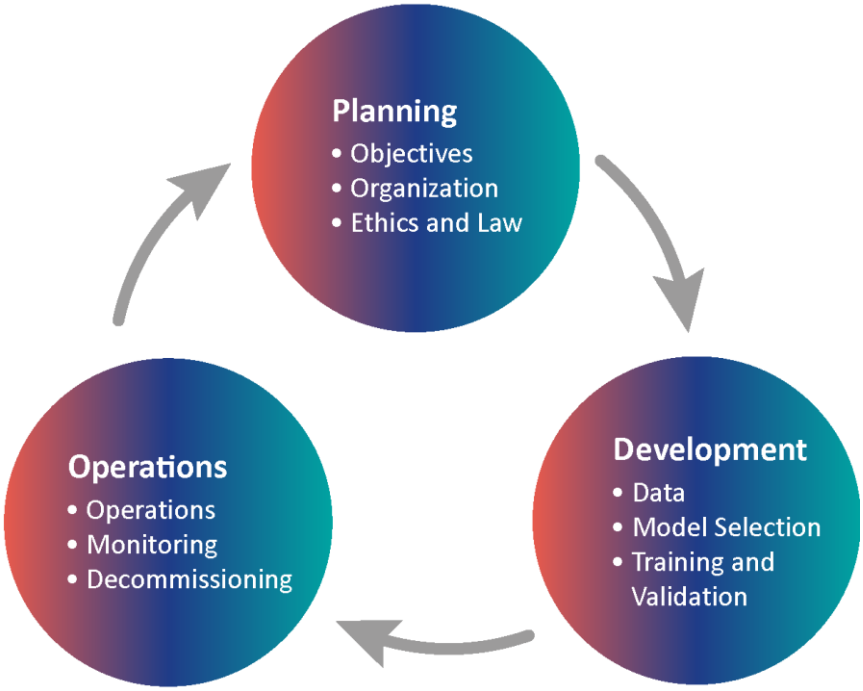

Figure 2 : The life cycle of an AI application as an iterative process.

The **planning phase** forms the strategic foundation of any AI application. It encompasses the definition of the overarching objectives, the needs analysis, and an early examination of legal, ethical and organisational requirements. During this phase, it is determined which specific challenges or problems are to be addressed through the use of AI, and to what extent these align with the organisation's overarching objectives. The responsible bodies clearly delineate the relevant processes and assess whether, and to what extent, the targeted cost-effectiveness appears realistic. Unlike conventional IT systems, strategic planning for AI applications requires particular attention to be paid to ethical issues and to the long-term adaptability of the systems, as AI models are constantly evolving and their future implications are difficult to predict.

ONGOING EXAMPLE

During the planning phase, the responsible body specifies that the application should correctly pre-select at least 90 per cent of incoming applications in order to reduce the manual workload. Qualifications, professional experience and language skills are defined as selection criteria and documented in writing. Characteristics such as gender or ethnic origin are excluded from the assessment. The final recruitment decision remains with the relevant line managers, who assess and take responsibility for the AI application's recommendations on a case-by-case basis.

The **development phase** comprises three closely interlinked sub-steps: algorithm and model selection, data selection and preparation, and model training and validation. The first step is to determine the appropriate algorithm – that is, the method that defines how the model learns from the data. The choice of algorithm depends on the data structure, the desired outcomes and the technical constraints, and has a direct influence on the accuracy, robustness and reliability of the resulting model. In a second step, the relevant data sources are selected and prepared. The data must be checked for quality, completeness and representativeness; preparation involves data cleansing, transformation and enrichment. Unlike conventional IT systems, an AI application requires continuous adaptation and review of the data set, as new data sources and changing data profiles directly influence model accuracy. In the third step, the model is trained on the basis of the prepared data and subsequently validated using test data. Unlike

traditional IT systems, AI applications undergo an iterative learning and adaptation process. After each training cycle, checks should be carried out to ensure that the model continues to operate fairly, transparently and reliably.

ONGOING EXAMPLE

During the development phase, a neural network is selected as the appropriate model based on the defined requirements, as it can recognise and weigh complex relationships between various application characteristics. Historical application documents, consisting of CVs, proof of qualifications and employment references, serve as training data. The data is checked for completeness and representativeness, adjusted for bias and converted into a standardised format to create a fair basis for model training. The trained model is then validated using a separate test dataset, achieving a prediction accuracy of 85 per cent. In addition, checks are carried out to ensure that the model correctly weights the previously defined selection criteria and does not incorporate any impermissible characteristics into its decisions.

The **operational phase** covers the productive use, ongoing monitoring and the orderly decommissioning of the AI application. After the transition into the production environment, a monitoring system is set up to continuously monitor the application's performance and trigger adjustments where necessary; for example, if the prediction accuracy falls below defined thresholds or anomalies occur in the decision-making processes. Unlike conventional IT systems, an AI application must be continuously monitored and adjusted, as its results react to changes in the input data and may lose accuracy and reliability without regular corrections. The end of the lifecycle involves an orderly decommissioning: the model is deactivated, all relevant data and model parameters are archived, and steps are taken to ensure that no security risks arise from unauthorised access or outdated data. This step goes beyond simply switching off a piece of software and requires careful documentation to meet legal requirements and satisfy the demands of future audits.

ONGOING EXAMPLE

During the operational phase, following implementation, the performance of the AI application is monitored by the responsible team using defined key performance indicators. An automated monitoring routine alerts the team if the prediction accuracy falls below 80 per cent or if unusual patterns occur in the assessment decisions. In such a case, the model is reviewed and, if necessary, retrained. After three years of operation, the administrative unit determines that a more up-to-date model achieves significantly better results. The existing AI application is then systematically taken out of service: all application data, model parameters and decision logs are backed up in accordance with the applicable archiving requirements to ensure the traceability of past decisions and to facilitate future audits.

The phases of the AI application lifecycle outlined above must be assessed in a systematic and risk-based manner. The aim is to identify potential vulnerabilities at an early stage in order to ensure the long-term security, reliability and fairness of the systems and to promote their responsible use. For each phase of the lifecycle, a risk area is identified which, based on experience, is of particular significance. However, these phase-specific priorities serve as a guide, not as an exhaustive list: depending on the nature, complexity and context of use of the AI application under review, risk areas from other phases may also be relevant and must be taken into account accordingly. The risk areas are always assessed in the context of the specific circumstances of the application in question.

2.3 Structure of the guide

The following chapters describe the three risk dimensions – 'Trustworthiness', 'Cost-effectiveness' and 'Competencies' – as well as the associated risk areas. Each risk area is structured in a consistent manner: first, the content of the risk area is outlined and its relevance within the context of the Federal Administration is explained. The key risks that may arise if the respective risk area is not adequately addressed are then identified. Finally, the text highlights the life cycle phases in which the risk area requires particular attention and sets out the audit questions that could address these risks. It should be noted that the risk areas under the 'Trustworthiness' dimension are primarily focused on a specific AI application, whilst risk areas under the 'Cost-effectiveness' and 'Competencies' dimensions are deliberately applied at the organisational or departmental level. This specifically broadens the audit perspective beyond the individual application: structural prerequisites such as an organisation's strategic direction, competence profile or willingness to change are decisive in determining whether AI applications can be used economically and responsibly in the long term. The guide thus enables both an application-specific and an organisation-specific perspective, thereby creating a more comprehensive basis for the audit.

The resulting breadth of the audit approach at the same time requires a consistently risk-oriented application of the guide. Not all risk areas need to be treated with equal weight in every audit. Auditors prioritise those risk areas where the highest risk is suspected on the basis of previous findings, contextual analysis or initial assessments. Risk areas with low risk may be deferred following a reasoned assessment. The selection of the risk areas to be audited and the rationale for this prioritisation must be recorded in the audit plan.

Furthermore, the risk dimensions should not be viewed as independent of one another. In practice, there are numerous interdependencies between them: a lack of expertise can jeopardise the trustworthiness of an AI application just as much as its cost-effectiveness. Auditors are therefore required to assess the three dimensions in a structured manner, but always within the overall context. Finally, the guidance applies to both in-house developed AI applications and the use of third-party solutions; depending on the use case, the audit priorities within the risk areas may shift.

3. TRUSTWORTHINESS

The 'trustworthiness' dimension assesses whether AI applications within the Federal Administration are used in an ethically justifiable, technically robust and socially responsible manner. It thus draws on the principles laid down by the Federal Council and Parliament and translates these into verifiable requirements. Trustworthiness is not an optional feature, but a fundamental prerequisite for the legitimate use of AI in the public sector: decisions taken by the Federal Administration affect fundamental rights and public trust; they must therefore be transparent, fair and legally sound.

AI applications, however, place qualitatively different demands on trustworthiness compared to conventional IT systems. Whilst traditional IT systems perform fixed, predictable functions, the output of AI applications can gradually deteriorate in quality if the underlying reality diverges from the training data. Consequently, the behaviour of AI applications must be continuously monitored and adjusted where necessary. This creates specific risks regarding fairness, human oversight, transparency and data protection, which do not exist in this form in conventional systems and which require a structured audit framework.

To enable a well-founded risk assessment of the trustworthiness of AI applications in practice, the seven risk areas are described below:

- Fairness (3.1),
- Autonomy (3.2),
- Responsibility (3.3),
- Transparency (3.4),
- Reliability (3.5),
- Security (3.6) and
- Data protection (3.7)

This selection is based on internationally recognised frameworks for trustworthy AI.[5] [6] [7] This nuanced approach enables the risks associated with artificial intelligence to be assessed in a targeted manner in practice, ensuring that the use of AI applications within the Federal Administration complies with applicable requirements not only technically and economically, but also ethically and legally.

3.1 Risk area: Fairness[8] [9]

AI applications should ensure fairness by not discriminating against any groups of people or legal and natural persons throughout their entire life cycle. This also means that AI systems must be user-friendly and accessible to all citizens – regardless of age, gender, abilities or other individual characteristics.[10] The

---

5 See Poretschkin, M., Schmitz, A., Akila, M., Adilova, L., Becker, D., Cremers, A. B., ... & Wrobel, S. (2021). Guidelines for the Design of Trustworthy Artificial Intelligence (AI Assessment Checklist).

6 See High-Level Expert Group on Artificial Intelligence. (2020). Assessment list for trustworthy artificial intelligence (ALTAI). European Commission. https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence.

7 See National Institute of Standards and Technology (NIST). (n.d.). Trustworthy and Responsible AI. Retrieved 1 October 2024, from https://www.nist.gov/trustworthy-and-responsible-ai.

8 See 'Guideline 1: Putting people first, keyword: discrimination', Swiss Confederation. (25 November 2020). 'Artificial Intelligence' Guidelines for the Federal Government: Framework for the use of artificial intelligence in the Federal Administration, p. 3.

9 See 'REQUIREMENT #5 Diversity, Non-discrimination and Fairness' of AI systems, High-Level Expert Group on Artificial Intelligence. (2020). Assessment List for Trustworthy Artificial Intelligence (ALTAI). European Commission (EU). p. 16.

10 Definition of discrimination according to the Federal Commission against Racism: 'unequal treatment of persons on the basis of essential and unchangeable characteristics of identity'. EKR: Discrimination (admin.ch).

'Fairness' risk area assesses AI applications for potential biases and discriminatory effects to ensure they comply with ethical standards and the principle of equal treatment.

Risk: Lack of fairness encompasses the risk that AI applications may make discriminatory decisions or systematically discriminate against certain groups. Biases in the underlying data can lead to AI models disadvantaging certain groups, either directly or indirectly, or reinforcing any existing prejudices. This is of particular importance to the Federal Administration, as public institutions are bound by the principle of equal treatment and must ensure that public services are provided without discrimination. The use of an unfair system would pose a major reputational risk for the Federal Government.

Phases: The focus of this risk area lies in the planning phase. This is where the foundations are laid that determine the fairness of an AI application: it is established which groups of people are affected by the application, which characteristics may be included in the assessment, and which fairness requirements the application must meet. If these issues are not clarified during the planning phase, any resulting biases can only be rectified in later phases at considerable expense. During the development phase, it must be ensured that the training data used does not contain any systematic biases and that the trained model complies with the defined fairness requirements. During operation, the application must be continuously monitored to check whether new biases creep in over time, for example due to changes in data profiles or societal developments.

Phase	Possible key questions for the review
Planning	Have the key demographic and socio-economic variables been identified and taken into account during the planning stage? Have potentially affected groups been identified and taken into account when defining the fairness requirements?
Development	Are the data sources representative of the scope of the AI application? Has the training data been systematically checked for bias, and have appropriate corrective measures been taken and documented? Is the trained model explicitly tested for discriminatory effects on relevant groups of people?
Operation	Is the AI application continuously monitored to detect whether new biases or discriminatory patterns emerge over time? Can erroneous or discriminatory decisions be reported within the administrative unit, and are these reports systematically investigated?

ONGOING EXAMPLE

If the AI recruitment system uses historical data that contains systematic disadvantages affecting certain socio-economic or demographic groups, women, for example, might be invited to interviews significantly less often because they have historically been under-represented in certain roles or management levels.

3.2 Risk area: Autonomy[11]

AI applications should support human agency and decision-making, as required by the principle of respect for human autonomy. This requires that AI applications act as catalysts for a democratic and just society by supporting the user's agency, whilst also safeguarding fundamental rights, which should be underpinned by human oversight. The 'Autonomy' risk area assesses AI applications in terms of their respect for human agency and human oversight.

Risk: A lack of autonomy carries the risk that AI applications will make decisions that restrict the freedom of choice and control of the individuals involved. An excessive degree of autonomy in an AI application could lead to human control mechanisms being weakened or even completely bypassed. This is particularly relevant in the Federal Administration, as decisions must always be transparent and consciously steered by human actors in order to uphold fundamental democratic principles. Furthermore, over time, an increasing dependence on the suggestions of an AI system may develop, leading to a loss of critical distance.

Phases: The focus of this risk area lies in the operational phase. It is here that it becomes apparent whether the intended human control mechanisms are actually effective in day-to-day practice, or whether a creeping dependence on the system's recommendations develops, which undermines a critical examination of the results. It is essential to monitor on an ongoing basis whether decision-makers are still critically evaluating the AI application's recommendations or are, in effect, adopting them unquestioningly. During the development phase, the technical foundations for appropriate human oversight are laid, for example through the implementation of override mechanisms or transparent explanations of the system's recommendations. Finally, the planning phase establishes the normative framework: it defines which decisions must be subject to human assessment and what degree of system autonomy is justifiable for the specific use case.

Phase	Possible guiding questions for the assessment
Planning	Are there provisions for intervention to control or correct the system where necessary? Is the level of system autonomy explicitly defined and aligned with legal and ethical requirements? Has it been specified which decisions must be subject to human assessment and which may be supported or prepared by the AI application?
Development	Have technical mechanisms been implemented that allow a human to override the system's recommendations at any time? Is the effectiveness of the implemented control mechanisms in practice verified as part of the validation process?
Operation	Are there mechanisms in place to intervene in the event of erroneous decisions? Is there ongoing monitoring to check whether decision-makers independently assess the AI application's recommendations or, in practice, accept them unquestioningly?

[11] See 'REQUIREMENT #1 Human Agency and Oversight' of AI systems, High-Level Expert Group on Artificial Intelligence. (2020). Assessment List for Trustworthy Artificial Intelligence (ALTAI). European Commission (EU). p. 7.

ONGOING EXAMPLE

If the AI system decides independently – without human review – which applicants are invited to an interview, erroneous decisions could go unnoticed. As a result, the relevant administrative unit effectively no longer exercises control over the recruitment process.

3.3 Risk area: Responsibility[12] [13]

AI applications should ensure that clear responsibilities are defined and verifiable accountability mechanisms are in place at every stage of their lifecycle. This includes mechanisms that guarantee that the AI system's decisions and actions are traceable and that responsible parties can be identified in the event of adverse effects. However, the formal allocation of responsibilities alone is not sufficient. The crucial factor is whether these responsibilities can actually be fulfilled in practice, which is closely linked to the competence profile of the bodies involved (see 5.1Risk area: Competence profile). The 'Responsibility' risk area ensures the traceability of an AI application's decisions. In the event of undesirable outcomes, this provides appropriate mechanisms for clarification and accountability.

Risk: Lack of accountability encompasses the risk that, in the event of erroneous or unfair decisions by the AI application, there are no clear structures in place to assign responsibility or to scrutinise decisions. This can lead to a loss of trust in the application. Within the federal administration, accountability is crucial for safeguarding the integrity of the civil service and ensuring that incorrect decisions are rectified by the responsible authorities.

Phases: The focus of this risk area lies in the operational phase. It is here that it becomes apparent whether the defined responsibilities are actually effective in day-to-day operations and whether the system's decisions are documented in a transparent manner. In the event of incorrect decisions, the relevant bodies should be notified and corrective measures initiated. An orderly decommissioning is equally critical: unclear responsibilities during this phase can lead to model parameters, logs and data not being properly archived, thereby losing the traceability of past decisions. During the planning phase, the formal foundations are laid by bindingly defining roles and responsibilities for the entire life cycle. In the development phase, technical mechanisms (logging, audit logs, etc.) must be implemented, as these are what enable the seamless traceability of system decisions during operation.

Phase	Possible key questions for the audit
Planning	Are the responsibilities for the entire life cycle of the AI application (from development to decommissioning) clearly defined and documented? Has it been established which body is accountable in the event of erroneous or unfair decisions made by the AI application? Are mechanisms in place enabling affected individuals to report erroneous decisions and request a review?
Development	Have clear responsibilities for model monitoring been defined? Have technical mechanisms (e.g. logging data) been implemented to ensure full traceability of the system's decisions?

[12] See 'Guideline 4: Accountability, keyword: accountability', Swiss Confederation. (25 November 2020). Guidelines on 'Artificial Intelligence' for the Federal Government: Framework for the use of artificial intelligence in the Federal Administration, p. 3.

[13] See 'REQUIREMENT #7 Accountability' of AI systems, High-Level Expert Group on Artificial Intelligence. (2020). Assessment List for Trustworthy Artificial Intelligence (ALTAI). European Commission (EU). p. 21.

	Is it ensured that the documentation of the model architecture and training decisions is complete, so that decisions can be understood retrospectively?
Operation	Have clear responsibilities been established for the period during and after decommissioning? Are erroneous decisions made by the AI application systematically recorded, reported to the relevant authorities and corrective measures initiated? Is it ensured that decision documentation is kept up to date throughout the entire operational period and is properly archived upon decommissioning?

ONGOING EXAMPLE

If the AI application were to make erroneous decisions during the recruitment of staff, without it being possible to clearly trace who is responsible for the application's decision, this could lead to a lack of auditability and insufficient scope for correction.

3.4 Risk area: Transparency[14] [15]

AI applications should ensure transparency so that their development and decision-making processes are traceable. This is important for building trust in the technology and enabling decisions to be scrutinised. Furthermore, the limitations of an AI application should be clearly communicated to users so that they can make informed decisions about its use. The 'Transparency' risk area assesses AI applications on the basis of how well they ensure traceability, explainability and communication in order to guarantee the necessary transparency.

Risk: Lack of transparency encompasses the risk that the functioning and decision-making processes of AI applications are not comprehensible to users and auditors. Particularly in the case of complex AI models, such as neural networks, there is a risk that even experts may not be able to fully grasp the decision-making logic. For public authorities, however, transparency is of particular importance, as they are legally obliged to provide clear justifications for their rulings and decisions in order to guarantee the right of appeal. There are therefore high standards of transparency required within the federal administration to strengthen citizens' trust in the decisions taken and to comply with legal requirements.[16]

Phases: The focus of this risk area lies in the development phase. The choice of model plays a decisive role in determining the extent to which decisions can later be explained in a comprehensible manner. Complex models such as neural networks generally offer greater predictive accuracy, but make it considerably more difficult to explain individual decisions. The key risk here is that models are inadequately documented: if there is a lack of documentation – appropriate to the intended use – of the model architecture, the training data and the design decisions made, technical traceability will be permanently compromised, with direct consequences for the verifiability of the system. Transparency requirements must be defined in a binding manner during the planning phase, as these directly influence the choice of model and the scope of documentation. During operation, checks must be carried out to ensure that the documentation produced is kept up to date.

[14] See 'Guideline 3: Transparency, Traceability and Explainability, Keyword: Transparency', Swiss Confederation. (25 November 2020). Guidelines on 'Artificial Intelligence' for the Federal Government: Framework for the Use of Artificial Intelligence in the Federal Administration, p. 3.

[15] See 'REQUIREMENT #4 Transparency' of AI systems, High-Level Expert Group on Artificial Intelligence. (2020). Assessment List for Trustworthy Artificial Intelligence (ALTAI). European Commission (EU). p. 14.

[16] Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., ... & Liang, P. (2023). The Foundation Model Transparency Index. arXiv preprint arXiv:2310.12941.

Phase	Possible key questions for the audit
Planning	Have the transparency requirements for the AI application been defined at an early stage, particularly with regard to the explainability of decisions to affected persons and auditors? Is it defined to what extent the functioning of the AI application must be disclosed to the public, affected persons, and internal bodies? Have the legal requirements regarding the duty to provide reasons been taken into account when selecting the model type?
Development	Have the criteria for selecting the algorithm and the model been defined and documented? Has documentation of the model architecture, the training data and the design decisions made been produced in a manner appropriate to the intended use? When selecting the model, was explicit consideration given to whether the accuracy achieved justifies or outweighs the associated loss of transparency (as well as any potential loss of autonomy)? Have mechanisms been implemented that provide comprehensible explanations for individual system decisions?
Operation	Are the model decisions documented and comprehensible in a manner appropriate to the intended use? Are protocols in place for the orderly and secure decommissioning of the AI application? Is the documentation for the AI application continuously updated during operation, so that changes to the model or the data remain traceable? Can affected individuals, upon request, receive a clear explanation of a system decision that concerns them?

ONGOING EXAMPLE

If the reason for the AI system's rejection of a job application cannot be clearly explained, the relevant administrative unit will find it difficult to understand the AI's decision and to justify it to those affected or to the public. This poses reputational risks for the authority, particularly if the impression arises that personnel decisions are made in a non-transparent manner or are algorithmically discriminatory.

3.5 Risk area: Reliability[17] [18]

AI applications should operate reliably over the long term and under varying conditions. They must be capable of performing their tasks correctly and consistently, even when their operating environment changes. In this context, reliability means that the AI application delivers precise and reproducible results, regardless of external influences over time or internal variations. Reliability also refers to the system's ability to handle varying or incomplete data, which requires that minimum requirements for input data

[17] See 'Guideline 5: Security, Keywords: Robust and Resilient Design', Swiss Confederation. (25 November 2020). 'Artificial Intelligence' Guidelines for the Federal Government: Framework for the Use of Artificial Intelligence in the Federal Administration, p. 5.

[18] See 'REQUIREMENT #2 Technical Robustness and Safety' of AI systems, High-Level Expert Group on Artificial Intelligence. (2020). Assessment List for Trustworthy Artificial Intelligence (ALTAI). European Commission (EU), p. 9.

be defined and that relevant edge cases be taken into account during training and validation. Furthermore, the ability to respond appropriately to unexpected inputs plays a central role in assessing reliability.

Risk: A lack of reliability entails the risk that AI applications may produce incorrect or inconsistent results under unforeseen conditions or when faced with erroneous inputs. A lack of reliability can lead to erroneous or unpredictable decisions, which can have serious consequences in safety-critical contexts. This is of great importance within the Federal Administration, as many AI applications are used in areas with high reliability requirements.

Phases: The focus of this risk area lies in the development phase. This is where the technical foundation for the reliability of the AI application is laid: the model is trained on representative and robust data and then systematically validated for its performance under various conditions, including erroneous, incomplete or unexpected inputs. Inadequate testing at this stage can result in vulnerabilities only becoming apparent during live operation, where the consequences of incorrect decisions are considerably more severe. During operation, reliability must be continuously monitored, as AI models can lose accuracy over time due to changing data patterns. In the planning phase, the minimum reliability requirements – in the form of accuracy thresholds or permissible error rates – must be defined in a binding manner, as these directly determine the requirements for training and validation.

Phase	Possible key questions for the assessment
Planning	Are mechanisms in place for continuous monitoring and adjustment? Have the minimum reliability requirements for the AI application been bindingly defined and documented? Has it been specified what measures will be taken if the defined reliability requirements are not met?
Development	Have the performance and robustness of the algorithm been demonstrated? Have suitable quality assurance measures been implemented during data preparation? Has the model been systematically tested for its behaviour in the event of incorrect, incomplete or unexpected inputs?
Operation	Have procedures been put in place to validate the model's performance under real-world conditions? Is model performance continuously monitored against defined key performance indicators, and are deviations systematically recorded and reported?

ONGOING EXAMPLE

If the AI application were to fail when evaluating applications involving unusual CVs or incomplete data, this could lead to incorrect decisions – and consequently to the rejection of suitable candidates.

3.6 Risk area: Security[19] [20]

AI applications must be protected against potential threats, such as malicious attacks or erroneous and fraudulent use. They should continue to function reliably and securely even in the event of malicious interactions – whether by human or artificial agents. Mechanisms must be implemented to proactively identify and minimise risks at the 'system' (safety) and 'algorithms/data' (security) levels, so that the safe operation of the AI application is guaranteed at all times. The 'security' risk area assesses AI applications in terms of their ability to detect and prevent potential hazards and threats.

Risk: Lack of security in AI systems encompasses the risk that AI applications may malfunction, causing unintended hazards to the outside world. In particular, the aim is to minimise potential risks of harm arising from faulty design or unsafe operating conditions. This is of particular importance to the Federal Administration, as AI applications could be used in safety-critical areas where malfunctions could result in harm to people or property.

Phases: The focus of this risk area is on the operational phase. The security of an AI application must be actively ensured throughout its entire operational lifespan: as attack vectors are constantly evolving, regular security reviews and updates are essential to ward off cyberattacks or manipulation by third parties. At the end of the lifecycle, it must also be ensured that all IT components are completely and securely deactivated upon decommissioning, as system components that have not been fully deactivated may continue to present vulnerabilities even after being taken out of service. The development phase plays a significant role in this regard: if sensitive data sources and training data are not adequately protected against tampering, or if the algorithm contains unaddressed vulnerabilities, this poses a structural and long-term threat to the system's security during operation. It is therefore important to establish robust security protocols throughout the entire training and implementation process. During the planning phase, security requirements at both the system and algorithm levels must be defined in a binding manner and coordinated with the relevant information security bodies.

Phase	Possible key questions for the review
Planning	Have the security requirements for the AI application been defined and coordinated with the responsible information security units? Has a risk analysis been carried out that identifies potential threat scenarios (e.g. data manipulation, unauthorised access or misuse)?
Development	Have the criteria for selecting the algorithm and the model been defined and documented? Have security aspects been taken into account when selecting the algorithm? Are IT security measures in place to protect the model against manipulation? Have the training data and data sources been protected against unauthorised access and manipulation? Has the algorithm been systematically tested for known vulnerabilities and unintended security flaws?
Operation	Are IT security checks in place to protect the AI application from attacks? Are protocols in place for the orderly and secure decommissioning of the AI application?

[19] See 'Guideline 5: Security, keyword: Secure Design', Swiss Confederation. (25 November 2020). 'Artificial Intelligence' Guidelines for the Federal Government: Framework for the Use of Artificial Intelligence in the Federal Administration, p. 5.

[20] See 'REQUIREMENT #2 Technical Robustness and Safety' of AI systems, High-Level Expert Group on Artificial Intelligence. (2020). Assessment List for Trustworthy Artificial Intelligence (ALTAI). European Commission (EU), p. 9.

	Is it ensured that, upon decommissioning, all IT components are fully and securely deactivated, and that no security vulnerabilities arise from unauthorised access to obsolete system components?

ONGOING EXAMPLE

If the AI recruitment application were vulnerable to manipulation or attacks, it could be deliberately influenced to unlawfully favour certain applicants who lack any objective qualifications for a position.

3.7 Risk area: Data protection[21] [22]

AI applications should ensure responsible data processing. Data protection involves minimising data access to what is strictly necessary, ensuring that processing procedures comply with data protection regulations, and maintaining clear access logs. AI applications must be designed in such a way that the processing of personal data is carried out securely and in accordance with data protection regulations. The 'Data protection' risk area assesses AI applications in terms of their ability to protect privacy and safeguard the integrity of the data being processed.

Risk: Inadequate data protection encompasses the risk that personal data – in particular, sensitive personal data[23] – may be processed or disclosed unlawfully. These risks are particularly high when data is processed or linked without adequate security measures, which may, for example, lead to the re-identification of individuals. For the Federal Administration, compliance with data protection regulations and respect for the right to informational self-determination are of paramount importance.

Phases: The focus of this risk area lies in the development phase. Personal data contained in the training data must be protected by means of appropriate anonymisation or pseudonymisation procedures, as inadequate anonymisation could constitute a direct breach of data protection regulations and jeopardise the right to informational self-determination.[24] The planning phase lays the normative foundation: in accordance with the principle of 'data protection by design' (Art. 7 FADP)[25] , data protection requirements must be incorporated into the design at an early stage. It must be specified which personal data may be processed, for what purpose and on what legal basis. During operation, it must be continuously checked whether the data being processed still corresponds to the original purposes and whether new risks have arisen as a result of data combinations that enable re-identification. Upon decommissioning, it must also be ensured that all personal data is duly deleted or archived in accordance with the applicable data protection requirements.

Phase	Possible key questions for the audit
Planning	Have the data protection requirements for the AI application been identified at an early stage and agreed with the relevant data protection officer?

[21] See 'Guideline 1: Putting people first, key terms: privacy and data protection regulations', Swiss Confederation. (25 November 2020). 'Artificial Intelligence' Guidelines for the Federal Government: Framework for the Use of Artificial Intelligence in the Federal Administration, p. 3.

[22] See 'REQUIREMENT #3 Privacy and Data Governance' of AI systems, High-Level Expert Group on Artificial Intelligence. (2020). Assessment List for Trustworthy Artificial Intelligence (ALTAI). European Commission (EU). p. 12.

[23] See Article 5(1)(c) of the Federal Act of 25 September 2020 on Data Protection (Data Protection Act, FADP; SR 235.1).

[24] See Article 13(2) of the Federal Constitution of the Swiss Confederation of 18 April 1999 (FC; SR 101)

[25] See Article 7(1) of the Federal Act of 25 September 2020 on Data Protection (Data Protection Act, FADP; SR 235.1).

	Has it been determined which personal data may be processed, for what purpose and on what legal basis?
Development	Is personal data in the training data protected by appropriate anonymisation or pseudonymisation procedures? Has it been ensured that the training data used complies with the principles of purpose limitation and will not be reused for other purposes?
Operation	Are the measures for protecting personal data sufficient? Are measures in place to delete data in accordance with data protection requirements? Is there an ongoing review to check whether the data being processed still complies with the original purpose limitations and whether any new risks have arisen?

ONGOING EXAMPLE

If the AI recruitment application fails to adequately protect applicants' personal data, sensitive information such as previous employment history, addresses or other sensitive data could be disclosed to unauthorised third parties.

4. COST-EFFECTIVENESS

The cost-effectiveness dimension assesses whether the use of an AI application in the Federal Administration is objectively justified, financially viable and organisationally sustainable. It thus addresses a core requirement of the SFAO: public funds must be used sparingly and in a results-oriented manner. This applies particularly to AI applications, as the associated investments in infrastructure, data and expertise are substantial and the benefits are often difficult to quantify.

According to ISSAI 300, the performance audit is divided into three traditional dimensions: **Economy** assesses whether the resources used are procured at the lowest possible cost. **Efficiency** examines the relationship between the resources used and the results achieved. **Effectiveness** assesses whether the intended objectives are actually achieved.[26] [27]

These three dimensions form the conceptual framework for the cost-effectiveness assessment. However, taken in isolation, they are too abstract to specifically capture the specific risks of AI applications in practice. AI applications entail specific economic risks that require a nuanced analysis: Their benefits may be indirect, delayed or only indirectly quantifiable in monetary terms, for example when it comes to building up skills. This is precisely why a structured and transparent cost-benefit assessment is essential, even if it involves uncertainties. Without explicit strategic integration, there is a risk of misallocating investment in systems that do not create genuine organisational added value. Furthermore, AI applications are heavily dependent on the surrounding system landscape, as they must continuously draw on data from various sources and integrate with existing processes. Ultimately, they are fundamentally dependent on the quality, completeness and timeliness of their data – a factor that places specific and often underestimated demands on data management.

To enable a well-founded risk assessment of the cost-effectiveness of AI applications in practice, the four risk areas

- **Strategy** (4.1),
- **Benefit orientation** (4.2),
- **System integration** (4.3) and
- **Data management** (4.4)

are described. The selection of these four risk areas is based on the SFAO's practical audit experience in the field of AI, as well as on an assessment of which financial risks are particularly relevant to AI projects within the Federal Administration. The four risk areas translate the conceptual audit dimensions of ISSAI 300 into concrete, auditable requirements tailored to the specific circumstances of AI applications within the Federal Administration. These risk areas are essential to ensure that the use of AI applications in the Federal Administration is not only technically feasible but can also be carried out in a way that is economically justified and sustainable.

In this context, account must be taken of the fact that AI applications within the Federal Administration are often still at an early exploratory stage. At this stage, the focus is not on providing immediate, quantifiable evidence of benefits, but rather on the targeted accumulation of practical knowledge, the testing of use cases and the development of institutional capacity to act. The requirements regarding the level of detail in cost-benefit analyses and performance indicators must be calibrated appropriately to the maturity of the project.

---

[26] See INTOSAI. (2019). ISSAI 300. Principles of Performance Auditing. International Standards of Supreme Audit Institutions. https://www.issai.org/pronouncements/issai-300-performance-audit-principles/

[27] See Swiss Federal Audit Office (SFAO). (n.d.). Performance audits. https://www.sfao.admin.ch/wp-content/uploads/publikationen/vorgehen-und-methoden/flyer-wirtschaftlichkeitspruefungen_d.pdf

4.1 Risk area: Strategy

AI applications should be embedded within a coherent overall organisational strategy. This requires that the use of AI is not viewed in isolation, but as an integral part of the digital transformation of the respective administrative unit. A clear strategic direction ensures that AI projects are derived from the organisation's overarching objectives, that available resources are deployed in a targeted manner, and that duplication of effort is avoided. The 'Strategy' risk area assesses AI applications to determine whether their deployment is strategically justified, organisationally embedded and aligned with the relevant framework conditions of the Federal Administration.

Risk: Lack of strategic anchoring encompasses the risk that AI projects are launched without a clear link to the organisation's objectives and thus tie up resources without creating sustainable added value. A lack of governance structures can lead to unclear responsibilities, delays in decision-making processes and the project failing in the face of resistance or changing framework conditions. This is of particular importance for the Federal Administration, as AI projects must be implemented within a complex institutional environment characterised by high demands on transparency, accountability and resource efficiency.

Phases: The focus of this risk area lies in the planning phase. This is where the strategic foundations are laid that determine the long-term success of an AI project: it is defined which organisational objectives are to be pursued through the use of AI, how the project is embedded within the existing IT and corporate strategy, and which governance structures ensure management and monitoring. An inadequate strategic foundation laid at this stage can hardly be rectified in later phases without incurring significant additional costs. During the development phase, it must be verified whether the strategic guidelines have been translated into concrete functional requirements and whether the developed system meets the defined objectives. Finally, during operation, it is necessary to continuously assess whether the AI system continues to contribute to the achievement of the strategic objectives or whether changes in the organisational framework require a reorientation.

Phase	Possible key questions for the review
Planning	Are the AI-related objectives and strategies of the administrative unit derived from the overarching guidelines of the department and the federal administration, and are they formulated in an outcome-oriented manner? Is there a clear prioritisation of AI projects, as well as structured strategic management that ensures resources are deployed in a targeted manner? Is the selection of areas for AI deployment based on a systematic analysis of potential benefits, rather than, for example, on availability or external influences? Are there analyses of potential benefits and cost components, including indirect costs such as operation, maintenance, training and change management? Have the relevant steering committees considered the AI project and taken a formal decision on its implementation or continuation? Is the use of the AI application embedded in a formal strategy or concept of the administrative unit, and have governance structures been clearly defined? Is the exploratory nature of the project – where applicable – explicitly acknowledged, justified and accompanied by appropriate success criteria?
Development	Have the strategic objectives of the AI project been translated into measurable functional requirements and documented? Is the progress of development regularly assessed against the strategic objectives and corrected in the event of deviations?

Operation	Does the AI application demonstrably contribute to the achievement of the defined strategic objectives during ongoing operation? Is there a process in place that triggers a review of the AI application's strategic relevance in the event of changes to the organisational framework? Is the current level of goal achievement being monitored?

ONGOING EXAMPLE

If an administrative unit introduces an AI application for the pre-selection of job applications without strategically embedding it or obtaining approval from the relevant steering committee, there is a danger that the project will be launched without clear objectives. This creates the risk of misallocation of resources and a lack of accountability.

4.2 Risk area: Benefit orientation

New AI applications should deliver demonstrable and measurable added value compared with existing solutions. This requires that the expected benefits be clearly defined and quantified before development begins, and weighed against the total costs involved. A consistent focus on benefits ensures that AI applications are not deployed for their own sake, but rather targeted specifically where they can achieve realistic gains in efficiency or effectiveness. It should be borne in mind that the benefits of AI applications do not always materialise immediately: building up expertise, gaining experience and the gradual optimisation of processes are legitimate and valuable long-term objectives that should be given appropriate weight in the assessment of benefits. A forward-looking approach, which also takes into account medium-term potential benefits, is therefore just as much a part of a sound economic assessment as the short-term cost-benefit analysis. Particular care is required when assessing projects in the exploratory phase. The 'benefit orientation' risk area is used to evaluate AI applications in terms of whether the expected benefits justify the costs incurred, whether they have been defined in measurable terms, and whether the full potential for benefits is being realised through consistent process optimisation.

Risk: A lack of benefit orientation encompasses the risk that AI applications are developed and operated without any significant and realistically achievable added value being expected. This can arise if no precise key performance indicators (KPIs) are defined or no baseline values are recorded to enable subsequent impact measurement. A further risk is that efficiency potential is only exploited in isolated instances: if existing processes are merely supplemented with AI rather than fundamentally rethought, the achievable benefits fall far short of what is possible. This is of particular importance for the Federal Administration, as public funds may only be used if a reasonable cost-benefit ratio can be demonstrated.[28]

Phases: The focus of this risk area lies in the planning phase. This is where it is determined what specific benefits are to be achieved with the AI application, how these are to be measured, and whether the total costs (including indirect costs) are in a reasonable proportion to the expected added value. If these foundations are not laid during the planning phase, there will be no binding benchmark against which to assess the project's success as it progresses. During the development phase, it must be ensured that the defined KPIs and baselines are incorporated into the technical implementation and that the system is designed in such a way that its impact can be measured at a later stage. During operation, regular checks must be carried out to verify whether the benefits realised meet expectations and whether the full potential for benefits is being realised through consistent process optimisation.

[28] See Art. 12(4) of the Federal Act of 7 October 2005 on the Federal Budget (Finanzhaushaltgesetz, FHG; SR 611.0).

Phase	Possible key questions for the audit
Planning	Have the expected benefits of the AI application been defined in concrete and measurable terms, with clear KPIs and a baseline established for comparison? Have the total costs of the project been fully accounted for, including indirect costs such as operation, maintenance, training and process adjustments? Did the cost-benefit analysis take into account not only short-term efficiency gains but also medium-term potential benefits – such as the development of AI skills and institutional know-how? Are the expected benefits proportionate to the total costs, and has this been substantiated by a formal cost-benefit analysis? Does the use of AI fundamentally optimise existing processes, or does it merely supplement them in specific areas?
Development	Have the defined KPIs and baselines been incorporated into the technical requirements, and is the system designed in such a way that its effectiveness can be measured at a later stage? Is project progress regularly measured against the expected potential benefits, and are any deviations reported to the steering committee?
Operation	Do the benefits realised during operation meet the expectations defined during the planning phase? Is efficiency potential being consistently exploited, or are there still untapped opportunities for optimisation in the relevant processes? If there is a persistent discrepancy between expected and realised benefits, is a review of the decision to continue operations initiated?

ONGOING EXAMPLE

If measurable objectives – such as a specific reduction in manual processing effort – and cost-benefit analyses are lacking, it is not possible to assess retrospectively whether the AI application has delivered the expected added value. Furthermore, if the existing recruitment process is not fundamentally redesigned, but the AI application is merely integrated in specific areas, significant efficiency potential may remain untapped.

4.3 Risk area: System integration

AI applications should be seamlessly integrated into the administrative unit's existing IT landscape and relevant business processes. This requires that interfaces with existing systems be identified at an early stage, technical dependencies clarified and organisational adjustments initiated in good time. Effective system integration ensures that the AI application is not operated in isolation, but can fully realise its benefits in conjunction with existing processes and systems. The 'System Integration' risk area assesses AI applications to determine whether the technical and organisational prerequisites for successful integration into the existing environment are in place.

Risk: Inadequate system integration encompasses the risk that AI applications cannot be integrated into existing IT systems and business processes, or can only be integrated inadequately. Incompatible interfaces, inadequate data flows or a lack of organisational adjustments can result in the AI application failing to realise its full potential or disrupting the operation of existing systems; in the case of AI systems that carry out actions autonomously within conventional systems, errors or malfunctions also have a direct and operational impact. This is of particular significance for the federal administration, as the IT landscape

is complex and has evolved over time, and integration projects often entail underestimated costs and risks.

Phases: The focus of this risk area lies in the development phase. This is where the technical interfaces are defined, compatibility with existing systems is ensured, and the necessary organisational adjustments are implemented. It must also be clarified whether, and to what extent, the AI application can be actively integrated into conventional systems beyond mere data exchange. It is also necessary to determine which control and safeguards are required for this. Inadequate integration work at this stage often leads to costly rectifications during operation. In the planning phase, the relevant interfaces and dependencies must be identified at an early stage and the integration effort realistically estimated. During operation, it is necessary to continuously monitor whether the integration is functioning stably and whether changes to existing systems are affecting the AI application.

Phase	Possible key questions for the audit
Planning	Were the relevant interfaces to existing systems and data sources identified at an early stage, and was the integration effort realistically estimated? Have the technical and organisational dependencies of the AI project been fully documented and taken into account in resource planning? Has it been checked whether the existing IT infrastructure meets the requirements of the AI application, for example in terms of computing capacity and data availability? Are similar solutions from other federal agencies available that can be built upon?
Development	Have the technical interfaces to existing systems been correctly implemented and systematically tested? Are synergies with existing models, platforms, data platforms and infrastructure being utilised? Has an analysis of dependencies (*vendor lock-in*) been carried out? Is the chosen IT architecture and landscape scalable? Is the scope of action for AI components that actively intervene in conventional systems clearly defined and documented?
Operation	Is the integration into the existing IT landscape functioning reliably, and are interface issues systematically identified and resolved? When changes are made to existing systems, is it checked whether these affect the AI application, and are appropriate adjustments made? Are AI components that interact directly with the system subject to structured lifecycle management – including commissioning, change management and decommissioning?

ONGOING EXAMPLE

If the administrative unit's existing HR system has an outdated database structure that prevents direct integration with the AI application, this results in unplanned additional costs for developing an intermediate layer to harmonise the data formats. If such integration risks are not identified at an early stage, they can lead to significant cost overruns and delays, jeopardising the cost-effectiveness of the entire project.

4.4 Risk area: Data management

AI applications are fundamentally dependent on a high-quality, complete and up-to-date database. The acquisition, processing and maintenance of this database incur significant costs, tie up human resources and thus represent a key cost driver in the life cycle of an AI application. Effective data management ensures that these investments are deployed in a targeted manner, that the data set meets the model's requirements, and that ongoing data operations remain economically viable. The 'Data Management' risk area assesses AI applications to determine whether the data set meets the necessary quality requirements and whether the organisational conditions for sustainable and cost-effective data operations are in place.

Risk: Inadequate data management entails the risk that incomplete, outdated or erroneous data will lead to unreliable model results. The resulting manual corrections are often time-consuming and consequently negate the expected efficiency gains. Administrative units with inadequate data infrastructure or a lack of data governance are faced with underestimated data preparation costs, which place a significant strain on project costs and timelines. Furthermore, there is a risk that a poorly maintained database will lead to a gradual deterioration in model performance during operation, resulting in additional costs for retraining and adjustments. This is of particular significance for the federal administration, as many administrative units possess heterogeneous data sets that have accumulated over time, the preparation of which often involves efforts that are underestimated.

Phases: The focus of this risk area lies in the development phase. This is where the relevant data sources are selected, checked for quality and representativeness, and prepared for training and testing the model. Underestimated data preparation efforts in this phase are a common cause of cost overruns and delays. During the planning phase, data requirements must be defined at an early stage, and the availability and preparation effort for the relevant data sources must be realistically estimated. Inadequate planning of the data infrastructure jeopardises the financial viability of the entire project. During operation, the data infrastructure must be continuously monitored, as a gradual deterioration in data quality can lead to additional costs for model adjustments.

Phase	Possible key questions for the audit
Planning	Have the necessary data sources been identified, and has the effort involved in obtaining, processing and maintaining them on an ongoing basis been realistically estimated? Have the project-specific costs for the one-off data preparation and indexing of the required data sources been fully taken into account in the project's overall cost estimate? Are there clear governance structures in place for data management, including responsibilities for data maintenance and quality assurance? Have the regulatory requirements for the use of the relevant data sources been fully identified and taken into account in the planning?
Development	Have the training and test data been systematically checked for completeness and quality, and has any data preparation work been factored into the project budget? Are the steps involved in data cleansing and preparation documented in a transparent manner and are they reproducible (for both one-off processing tasks and ongoing data pipelines)? Have incompatible data formats or data silos been identified at an early stage, and have measures been put in place to resolve them? Is data risk management integrated into the development process, and are data-related risks systematically identified and addressed?
Operations	Is the quality of the processed data continuously monitored, and are deviations from the expected data profile recorded and rectified?

	Are the costs for ongoing data maintenance and any model adjustments included in the operational budget? Are there clear responsibilities for ongoing data maintenance, and are these fulfilled in practice?

ONGOING EXAMPLE

If the required application data is spread across several incompatible systems and this circumstance was not taken into account during the planning stage, this will result in significant additional costs and delays in data preparation. Furthermore, if no clear responsibilities for ongoing data maintenance are defined, data quality may gradually deteriorate during operation. This results in additional effort required for corrections and model adjustments and jeopardises the project's financial viability – or reduces its benefits.

5. COMPETENCIES

The 'Competencies' dimension assesses whether the individuals and organisations involved in an AI application possess the necessary skills, knowledge and structures to develop, deploy and manage AI systems responsibly. It thus addresses a key insight from practical experience: even technically mature AI applications often fail not because of the technology itself, but due to a lack of competencies, unclear responsibilities or the organisation's unwillingness to embrace change. This applies not only to the departments involved in development and management, but also explicitly to the users who work with the AI application on a day-to-day basis and must critically assess its results.

In this context, two aspects are to be regarded as particularly relevant to the audit: on the one hand, the competence profile of the departments involved — whether the necessary technical, legal and specialist know-how is available or is being specifically developed. On the other hand, organisational change – whether the organisational units concerned are being actively prepared for the changes brought about by the use of AI. Both aspects are closely interlinked: a lack of skills hinders change, and change that is inadequately supported prevents existing skills from being utilised effectively. The following section outlines the two risk areas

- Competence profile (5.1) and
- Organisational Change (5.2)

are described in order to enable a targeted assessment of the competence-related risks of AI applications in the Federal Administration. These risk areas are essential to ensure that the use of AI applications in the Federal Administration is sustainably embedded not only technically and economically, but also organisationally and in terms of human resources.

5.1 Risk area: Competence profile

AI applications require a broad and interdisciplinary range of expertise, encompassing technical, legal, specialist and ethical know-how. This means that the individuals and organisations involved in an AI project must possess the necessary skills to develop, deploy and monitor AI systems appropriately. Users play a particularly important role in this regard: they must have a sufficient understanding of how the AI application works and its limitations in order to critically assess its results, recognise incorrect decisions and consciously assume responsibility for the final decision. An adequate competence profile ensures that risks are identified at an early stage, decisions are made on a sound basis, and the quality of the AI application can be maintained throughout its entire life cycle. The 'Competence Profile' risk area assesses AI projects to determine whether the organisations involved possess the necessary skills and whether appropriate measures to build competence have been put in place.

Risk: Inadequate competence profile encompasses the risk that AI applications are developed, deployed or monitored by individuals who lack the necessary knowledge. A lack of technical know-how can lead to flawed model decisions, inadequate testing or poor monitoring during operation. A lack of legal and ethical expertise increases the risk that legal requirements will not be met or that ethical risks will not be identified. This is of particular significance for the Federal Administration, as AI expertise is not yet widespread and building up the necessary know-how requires time and resources.

Phases: The focus of this risk area lies in the planning phase. This is where it is determined which skills are required for the AI project, whether these are available in-house or need to be sourced externally, and what measures should be taken to build up these skills. An inadequately defined skills profile at this stage jeopardises the quality of development and of operations. During the development phase, it must be ensured that those involved possess the necessary skills to develop, test and validate the model appropriately. During operation, it must be continuously assessed whether the existing skills still meet the changing requirements of the AI application and whether further training is necessary.

Phase	Possible key questions for the assessment
Planning	Have the competencies required for the AI project (technical, legal, subject-matter and ethical) been systematically identified and compared with the skills available in-house? Have measures for building up expertise or for externally procuring missing competencies been defined and taken into account in resource planning? Have roles and responsibilities within the AI project been clearly assigned and aligned with the necessary competencies? Is the development of AI expertise viewed as a strategic investment and prioritised accordingly?
Development	Do the individuals involved in development possess the necessary knowledge to develop, test and validate the model appropriately? Is legal and ethical expertise integrated into the development process through the involvement of data protection officers or legal experts? Are knowledge gaps that emerge during development systematically identified and addressed through targeted measures?
Operation	Do the individuals responsible for operations have the necessary knowledge to monitor the AI application appropriately and adapt it where necessary? Are training programmes continuously adapted to the changing requirements of the AI application? Is it ensured that, should key personnel leave the organisation, the necessary expertise is retained within the organisation?

ONGOING EXAMPLE

If the administrative unit lacks the necessary knowledge in the areas of machine learning and data protection law, there is a risk that the AI application will be used without proper scrutiny. The relevant HR staff will be unable to assess the system's recommendations properly and will therefore fail to recognise incorrect decisions. This undermines human oversight and jeopardises the quality of the entire recruitment process.

5.2 Risk area: Organisational change

AI applications fundamentally change work processes, role definitions and decision-making structures. This requires the organisational units concerned to be actively prepared for these changes and for the transformation to be managed in a targeted manner. Successful organisational change ensures that AI applications are accepted by the staff concerned, used appropriately and integrated into existing workflows. The 'Organisational Change' risk area assesses AI projects to determine whether the organisational prerequisites for the successful introduction and sustainable embedding of the AI application are in place.

Risk: Inadequate management of organisational change encompasses the risk that AI applications will meet with resistance from the staff concerned, be underutilised or destabilise existing processes. A lack of communication, unclear implications for existing roles and tasks, and insufficient involvement of those affected can result in the AI application failing to realise its full potential. This is of particular significance for the Federal Administration, as AI applications are being introduced into a complex institutional environment with established structures and processes, which makes the transition even more challenging.

Phases: The focus of this risk area lies in the operational phase. It is here that it becomes apparent whether the organisational changes are supported by the staff concerned and whether the AI application is sustainably integrated into day-to-day work. Resistance that arises during this phase is often difficult to overcome and can jeopardise the long-term success of the project. The planning phase plays a significant role in this regard: here, the organisational implications of AI deployment must be analysed at an early stage, the relevant departments must be involved, and suitable measures to support the change must be defined. During the development phase, it must be ensured that the AI application is designed in such a way that it is accepted by users and can be used practically in day-to-day work.

Phase	Possible key questions for the review
Planning	Have the organisational implications of AI deployment on roles, tasks and decision-making processes been systematically analysed and documented? Have the affected staff and managers been involved in the project at an early stage and informed about the planned changes? Have measures to support organisational change been defined and taken into account in resource planning?
Development	Is the AI application designed in such a way that it is understandable and user-friendly for users, and can be integrated into their existing daily work routines? Are the affected staff members involved in the development (e.g. through pilot trials or feedback sessions) to promote acceptance at an early stage?
Operation	Is the acceptance of the AI application among the affected staff being continuously monitored, and is any resistance being actively addressed? Have the implications of the AI application for existing roles and tasks been clearly communicated, and have the necessary adjustments been made to job descriptions or processes? When significant changes are made to the AI application in the workplace, is it assessed whether further measures are needed to support the change?

ONGOING EXAMPLE

If the roll-out of the AI application is not actively supported and the HR staff concerned are not involved at an early stage, the administrative unit risks the application meeting with resistance or being used without due consideration. Without targeted communication and training measures, it remains unclear how roles within the recruitment process will change, which poses a long-term threat to the acceptance and proper use of the AI application.

6. AUDIT PROCEDURE

Figure 3 summarises the audit framework of this guide. It maps the three risk dimensions and their respective risk areas against the three life cycle phases, making the phase-specific focus areas visible at a glance. The figure serves as a guide for auditors when planning and conducting an AI audit. **The highlighted focus areas indicate in which life cycle phase a risk area is typically most relevant, based on experience. However, they do not claim to be exhaustive**. In practice, any risk area may be significant in any phase: the nature, complexity and operational context of the AI application under assessment ultimately determine which risk areas should be prioritised for assessment in a specific case. Before commencing an AI audit, the scope of the audit must therefore be explicitly defined: is the focus on a specific AI application, is the audit conducted at organisational level, or both (see section2.3 Structure of the guide). Auditors are expressly encouraged to critically examine the key areas and to adapt the guide flexibly to the specific characteristics of their audit. The diagram does not replace an in-depth examination of the individual risk areas, as described in sections 3 to 5, but serves as a visual introduction to the assessment framework.

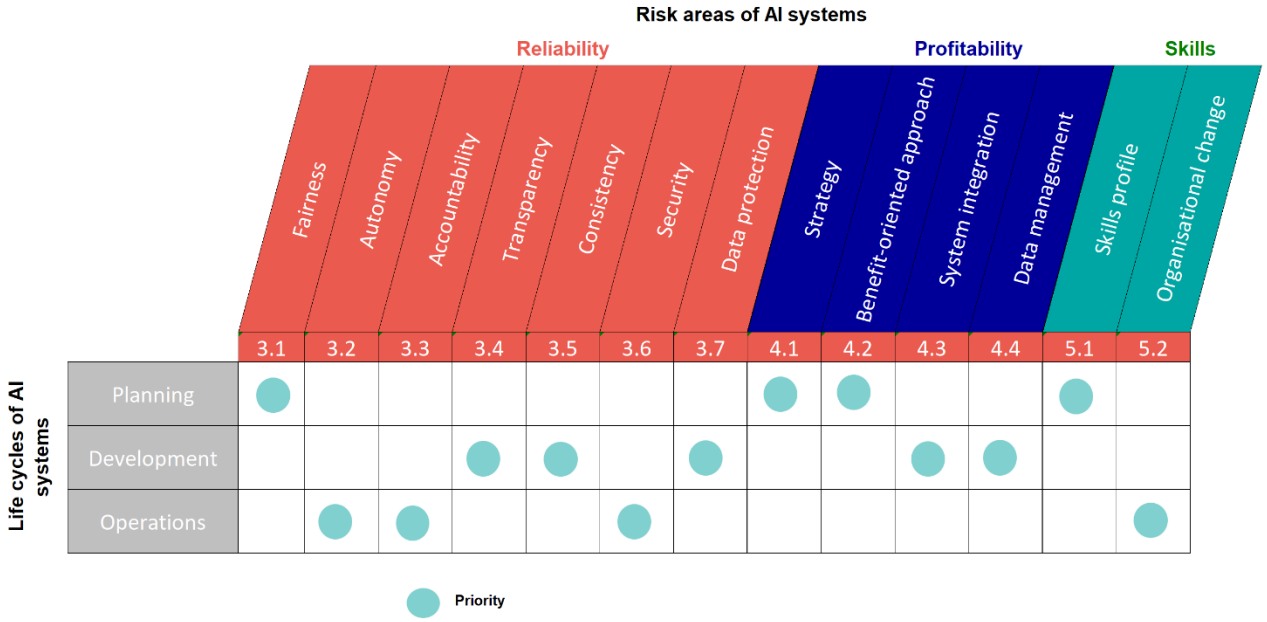

Figure 3 : Overview of the risk areas across the three risk dimensions throughout the lifecycle of an AI application.

Auditors are not required to examine all risk areas in equal depth. Rather, the extent of the audit depends on the presumed risk: risk areas in which significant vulnerabilities are suspected – based on the specific characteristics of the AI application under audit, the institutional context or initial findings – must be examined in greater depth. Risk areas presenting a low level of risk may be deferred following a reasoned assessment. This prioritisation decision must be documented transparently and forms an integral part of the audit planning.

The audit of an AI application should ideally be carried out by a multidisciplinary team that brings specialist knowledge to the relevant audit focus areas. This team may, for example, comprise experts from the fields of business efficiency, project management, IT security, artificial intelligence, compliance and law. Should uncertainties arise when addressing certain questions, it may be helpful to draw on additional expertise to ensure a well-founded and comprehensive audit.

ANHANG 1 – FURTHER READING

1. Federal Council. (2020). 'Artificial Intelligence' Guidelines for the Federal Government – Guidance on the use of AI in the Federal Administration (german). Retrieved from https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-81319.html
2. Algemene Rekenkamer (Netherlands Court of Audit). (2022). An Audit of Algorithms – Nine Algorithms used by the Dutch Government. Retrieved from https://english.rekenkamer.nl/publications/reports/2022/05/18/an-audit-of-9-algorithms-used-by-the-dutch-government
3. National Audit Office (NAO). (2024). Use of Artificial Intelligence in Government. Retrieved from https://www.nao.org.uk/reports/use-of-artificial-intelligence-in-government/
4. SAIs of Finland, Germany, the Netherlands, Norway, and the United Kingdom. (2020). Auditing machine learning algorithms – A white paper for public auditors. Retrieved from https://www.auditingalgorithms.net/
5. Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS). (2021). Guidelines for designing trustworthy artificial intelligence. Retrieved from https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefkatalog/202107_KI-Pruefkatalog.pdf
6. European Commission, High-Level Expert Group on Artificial Intelligence. (2019). The Assessment List for Trustworthy Artificial Intelligence – for self-assessment. Retrieved from https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342
7. National Audit Office (NAO). (2022). Good Practice Guidance – Framework to Review Models. Retrieved from https://www.nao.org.uk/wp-content/uploads/2016/03/11018-002-Framework-to-review-models_External_4DP.pdf
8. Institute of Public Auditors in Germany (IDW). (2022). IDW Auditing Standard 861 – Audit of AI Systems. Retrieved from https://www.idw.de/IDW/IDW-Verlautbarungen/IDW-PS/IDW-EPS-861-02-2022.pdf
9. The Institute of Internal Auditors (IIA). (2022). The IIA's Artificial Intelligence Auditing Framework. Retrieved from https://www.nist.gov/system/files/documents/2021/10/04/GPI-Artificial-Intelligence-Part-II.pdf
10. OECD (2023), "Advancing Accountability in AI: Governing and Managing Risks Throughout the Lifecycle for Trustworthy AI", OECD Digital Economy Papers, No. 349, OECD Publishing, Paris, Retrieved from https://www.oecd.org/en/publications/advancing-accountability-in-ai_2448f04b-en.html