AI Security Baseline

Technische AI Security Baseline für Finanzinstitute

Dokumenttyp: Technische AI Security Baseline für Finanzinstitute

Version: V2.0 · Status: Published Version · Datum: 12.07.2026

Owner: Martin Bischoff · 4T2 Solutions GmbH

Public · Frei verfügbares Rahmenwerk

Dokumentkontrolle

Feld	Wert
Dokumentname	AI Security Baseline
Version	V2.0
Status	Published Version
Vertraulichkeit	Public · Frei verfügbares Rahmenwerk
Owner	Martin Bischoff
Review-Zweck	Publizierte Version

Inhaltsverzeichnis

1. Zweck, Zielgruppe und Verwendung

Zweck

Die Security Baseline für Artificial Intelligence (AI) beschreibt die technischen Mindestanforderungen, die ein AI-System vor Betriebsfreigabe, Nutzungsfreigabe und produktivem Betrieb erfüllen muss. Sie gilt für AI-Plattformen, Modelle und Systeme mit Retrieval-Augmented Generation (RAG). Sie gilt auch für Agenten, Tool-Integrationen, Server des Model Context Protocol (MCP), AI-Funktionen in Software-as-a-Service-Produkten (SaaS-Produkten) und technische Komponenten, die Daten, Rechte, Ausgaben oder Aktionen einer AI-Nutzung beeinflussen.[1]

Zielgruppe

Security Engineers, Plattformteams, Solution Architects und technische System Owner verwenden die Baseline für Architektur, Implementierung, Security Review und Betrieb. Sie schafft zugleich eine belastbare technische Nachweisbasis für den Chief Information Security Officer (CISO), Second Line und Interne Revision.

Beantwortete Managementfrage

Schützen Plattform und technische Umsetzung die freigegebene AI-Nutzung ausreichend, und ist sie unter den nachgewiesenen technischen Voraussetzungen kontrolliert betreibbar?

Beziehung zur Dokumentenfamilie

Das AI Governance Framework for Financial Institutions definiert Führungsmodell, Rollen, AI Governance Lifecycle und Freigabelogik. Der AI Control Standard macht die Einhaltung durch Safeguards prüfbar. Das AI Operations Handbook führt die zugehörigen Workflows aus. Die Security Baseline konkretisiert ausschliesslich die technischen Mindestanforderungen und liefert die dafür benötigte Evidence.

Erwartete Ergebnisse und Abgrenzung

Die Anwendung ergibt eine zugeordnete technische Baseline, implementierte Requirements, nachvollziehbare Evidence und dokumentierte Abweichungen. Diese Ergebnisse unterstützen Betriebsfreigabe, Safeguard Testing, Revalidierung und Incident Response.

Die Baseline entscheidet weder über den fachlichen Nutzen noch über die Zulässigkeit eines AI Use Case. Sie ersetzt keine Risikoklassifikation, Nutzungsfreigabe, Supplier-Entscheidung oder fachliche Modellvalidierung.

Sie ergänzt die verbindlichen Enterprise Controls des Instituts und dupliziert sie nicht. Für jeden AI Use Case werden die folgenden Abhängigkeiten im Security Evidence Package referenziert und auf Anwendbarkeit geprüft:

1 Die Baseline verwendet einen technisch breiten Scope. Ob eine Komponente rechtlich ein AI-System nach Art. 3 № 1 AI Act ist, wird im Governance- und Klassifikationsprozess entschieden.

Enterprise-Control-Abhängigkeit	Erwartete Referenz
Identity und Privileged Access	verbindlicher IAM-, Multi-Faktor-Authentisierungs- und PAM-Standard
Kryptografie und Schlüssel	Verschlüsselungs- und Key-Management-Standard
Vulnerability Management	zentraler Scan-, Priorisierungs- und Behebungsprozess
Logging und Zeitbasis	zentrale Logplattform und verlässliche Zeitsynchronisation
Backup und Resilienz	Backup-, Business-Continuity-, Disaster-Recovery- und DORA-Teststandard

Methodische Grundlage

Erklärende Passagen folgen Digital.gov und der 4T2 Professional Prose. Technische Requirements trennen nach dem offenen IT-Grundschutz-Muster des Bundesamts für Sicherheit in der Informationstechnik (BSI) den erforderlichen Sicherheitszustand von möglichen Umsetzungshinweisen. Die Special Publications 800-53 und 800-53A des U.S. National Institute of Standards and Technology (NIST) dienen als Muster für atomare Ergebnisse und Verifikation; die Open Security Controls Assessment Language (OSCAL) für stabile IDs und getrennte Felder. Die methodische Nutzung begründet weder eine vollständige Katalogübernahme noch eine Zertifizierung.

Verifikation und Fail Criteria

Jedes Requirement beschreibt einen technischen Sollzustand. Die angegebene Evidence wird mindestens durch **Examine** geprüft; dazu gehören Dokumente, Konfigurationen und Logs. Wo das Requirement technisches Verhalten verlangt, kommt ein reproduzierbarer **Test** hinzu. Ein **Interview** kann Verantwortlichkeiten und Abläufe bestätigen, ersetzt aber weder Examine noch einen erforderlichen technischen Test. Ein Requirement besteht die Prüfung nicht, wenn der Sollzustand nicht erfüllt, die Evidence nicht aktuell oder die behauptete Wirkung nicht reproduzierbar ist.

Technische Anwendungslogik

Die Profilklasse B, E, C oder H bestimmt die erforderliche Kontrolltiefe.[2] Anschliessend zeigt die technische Architektur, welche Komponenten des AI Use Case tatsächlich betroffen sind. Die folgende Tabelle verbindet diese Komponenten mit der technischen Oberfläche, die in den nachfolgenden Requirement-Familien geprüft wird. Ein Team wählt daraus die einschlägigen Requirements und dokumentiert Abweichungen nach dem Control Standard. X ist keine Profilklasse für einen zulässigen Betrieb, sondern führt zu Blockierung, Isolation oder kontrolliertem Abbau.

Technisches Objekt	Zu prüfende technische Oberfläche
AI-Anwendung	Benutzeroberfläche, Geschäftslogik, Prompt-Schicht, Integrationen und Logs
Model Service	Gehostetes, lokales oder angepasstes Modell sowie Model Gateway oder Application Programming Interface (API)

2 B, E, C und H bezeichnen die aufeinander aufbauenden technischen Kontrolltiefen Basic, Enhanced, Critical und High Risk. X bezeichnet den Ausschlussstatus Excluded / prohibited und keine freigabefähige Betriebsstufe. Die vollständige Zuordnung und die Massnahmen bei X folgen in Kapitel 16.

Technisches Objekt	Zu prüfende technische Oberfläche
RAG Service	Datenquelle, Embedding Pipeline, Vector Database, Retrieval-Schicht und Quellenangaben
Agent Service	Tool Registry, Berechtigungen, Ausführungsgrenzen, Freigaben und Stop Controls
Integration Layer	API, Connector, MCP Server, Plugin, technische Identität und Secrets
Monitoring Layer	Logs, Telemetrie, Warnungen, Drift- und Missbrauchssignale sowie forensische Nachweise

2. Security Architecture Principles für AI

Die Architekturübersicht zeigt, wie Informationen durch das AI-System fliessen und an welchen Stellen Sicherheitskontrollen eingreifen. Sie bildet damit die technische Grundlage für Freigaben, Incident Response und Revalidierung.

Besondere Aufmerksamkeit gilt den Übergängen zwischen Modell, Datenquellen und angebundenen Funktionen. An diesen Vertrauensgrenzen muss erkennbar sein, welche Zugriffe zulässig sind, wie sie kontrolliert werden und wie das Institut bei einem Sicherheitsereignis eingreifen kann.

ID	Requirement	Applicability	Evidence
AI-SB-ARCH-01	Das AI-System hat eine aktuelle Architekturübersicht mit Nutzern, Anwendungen, Modellen, Datenquellen, Tools, externen Diensten, Netzgrenzen und Logging-Zielen.	B, E, C, H	Architecture diagram, data-flow diagram
AI-SB-ARCH-02	Architekturgrenzen trennen Entwicklung, Test, Pilot und Produktion. Produktive Daten dürfen nur in dafür freigegebenen Umgebungen genutzt werden.	B, E, C, H	Environment overview, data handling statement
AI-SB-ARCH-03	Kritische Kontrollpunkte für Zugriff, Retrieval, Tool-Aufruf, Ausgabe, Logging und Stop-Möglichkeit sind dokumentiert.	E, C, H	Control point map
AI-SB-ARCH-04	Kritische oder wichtige Nutzungen haben dokumentierte Fallback-, Degradation- oder Stop-Optionen.	C, H	Resilience and stop plan
AI-SB-ARCH-05	Architekturänderungen mit Wirkung auf Daten, Rechte, Modelle, Tools oder Anbieter lösen Revalidierung aus.	B, E, C, H	Change record, revalidation trigger log
AI-SB-ARCH-06	Technische Komponenten, Modelle, Integrationen und Logs verwenden die eindeutige AI-System- oder Use-Case-ID. Bei gesetzlicher Registrierung wird die EU-Datenbank-ID oder Provider-URL als Metadatum referenziert.	B, E, C, H; Registrierungsmetadatum bei bestätigtem Art.-49-Trigger	Configuration inventory, AI Register link, registration metadata

ID	Requirement	Applicability	Evidence
AI-SB-ARCH-07	Kritische AI-Nutzungen verweisen auf Business Impact Analysis, Wiederanlaufzeit, maximal tolerierbaren Datenverlust und einen getesteten Wiederanlaufplan. Bestehende DORA- oder BCM-Evidence darf referenziert werden.	C, H oder kritische/wichtige Funktion	BIA link, RTO/RPO record, recovery test evidence

3. Plattform- und Deployment-Baseline

Die Plattform muss kontrollierten Betrieb ermöglichen. Das gilt für interne Plattformen, Managed AI Services, SaaS-AI, externe AI-APIs und hybride Architekturen.

Modell- und Serviceänderungen können das Verhalten einer AI-Nutzung verändern, ohne dass Anwendungscode angepasst wird. Eine belastbare Plattform-Baseline hält deshalb Umgebung, Konfiguration, Versionen und Betriebsgrenzen nachvollziehbar.

ID	Requirement	Applicability	Evidence
AI-SB-PLT-01	Jede AI-Plattform hat einen technischen Owner, eine Betriebsumgebung, eine Datenzone und einen Supportpfad.	B, E, C, H	Platform record
AI-SB-PLT-02	Produktionsbetrieb erfolgt nur auf freigegebenen Plattformen oder Services mit dokumentierter Betriebsfreigabe.	B, E, C, H	Betriebsfreigabe Evidence
AI-SB-PLT-03	Plattformkonfiguration, Modellversionen, Promptversionen und Integrationsversionen sind nachvollziehbar.	E, C, H	Configuration inventory
AI-SB-PLT-04	Nicht freigegebene AI-Services sind technisch zu blockieren oder über akzeptierte Restrisiken zu behandeln.	B, E, C, H	Access policy, exception record
AI-SB-PLT-05	Ressourcenverbrauch, Kosten und Rate Limits sind für produktive AI-Nutzung technisch begrenzt oder überwacht.	E, C, H	Quota settings, consumption report

4. Identity, Access und Privileged Operations

AI-Zugriffe müssen menschliche Identitäten, technische Identitäten und vom System ausgelöste Aktionen trennen. Agenten und Integrationen dürfen nur Rechte erhalten, die für die freigegebene Nutzung erforderlich sind.

Diese Trennung ist für AI besonders wichtig, weil ein Modell Aktionen über mehrere Identitäten und Tools anstossen kann. Eindeutige Zuordnung und Least Privilege begrenzen die Reichweite fehlerhafter oder manipulierter Ausgaben.

ID	Requirement	Applicability	Evidence
AI-SB-IAM-01	Menschliche Nutzer, Service Accounts, Agentenidentitäten und technische Integrationen sind eindeutig identifizierbar.	B, E, C, H	Identity inventory

ID	Requirement	Applicability	Evidence
AI-SB-IAM-02	Zugriff auf Modelle, Datenquellen, Prompts, Logs, Tools und Admin-Funktionen folgt Least Privilege.	B, E, C, H	Access matrix
AI-SB-IAM-03	Privilegierte Operationen sind stärker geschützt, protokolliert und regelmässig rezertifiziert.	E, C, H	Admin access review
AI-SB-IAM-04	Tool- und Agentenrechte sind an AI Use Case, Nutzerrolle und Freigabeumfang gebunden.	E, C, H	Tool permission record
AI-SB-IAM-05	Produktionsrechte sind von Entwicklungs- und Testrechten getrennt.	B, E, C, H	Segregation evidence
AI-SB-IAM-06	Berechtigungen werden bei Stilllegung, Supplier Exit oder Nutzungsentzug zeitnah entfernt.	B, E, C, H	Deprovisioning log

5. Data Security für Prompts, Trainingsdaten, Logs und Embeddings

Die fachliche Datenfreigabe und die datenschutzrechtliche Bewertung erfolgen in den dafür zuständigen Verfahren. Diese Baseline regelt, wie freigegebene Daten innerhalb eines AI-Systems technisch geschützt und kontrolliert werden.

Prompts, Embeddings und Modellantworten können vertrauliche Informationen aus ihrer ursprünglichen Ablage herauslösen und an weiteren Stellen speichern. Ihr Schutz muss deshalb den gesamten technischen Lebensweg abdecken: von der Übertragung bis zur nachweisbaren Löschung.

ID	Requirement	Applicability	Evidence
AI-SB-DATA-01	Datenarten für Prompt, Antwort, Training, Evaluation, Retrieval, Logs, Telemetrie und Embeddings sind klassifiziert.	B, E, C, H	Data classification record
AI-SB-DATA-02	Vertrauliche Daten werden durch Verschlüsselung, Zugriffskontrolle und geeignete Speicherorte geschützt.	B, E, C, H	Encryption and storage evidence
AI-SB-DATA-03	Training, Fine-Tuning oder Anbietertraining mit Institutsdaten ist deaktiviert oder explizit freigegeben.	B, E, C, H	Training exclusion evidence
AI-SB-DATA-04	Produktive Daten in Test, Pilot oder Evaluation sind maskiert, synthetisch ersetzt oder durch Freigabe gedeckt.	E, C, H	Test data approval
AI-SB-DATA-05	Aufbewahrung und Löschung sind für Prompts, Antworten, Logs, Embeddings, Indizes und Modellartefakte definiert.	B, E, C, H	Retention and deletion settings
AI-SB-DATA-06	Datenabfluss über Antworten, Logs, Anbieter, Plugins oder Connectoren wird durch technische Regeln begrenzt.	E, C, H	Leakage control evidence
AI-SB-DATA-07	Ein Enforcement Point prüft GPAI-Eingaben vor externer Übertragung auf Secrets und unzulässige Datenklassen. Er erlaubt, maskiert, blockiert oder hält sie zurück.	externe GPAI-Nutzung	Ingress policy and blocking test

6. RAG und Vector Database Security

RAG-Systeme[3] übernehmen Risiken aus Quellen, Berechtigungen und Aktualisierungspfaden. Die technische Mindestlinie muss verhindern, dass Retrieval Datenzugriffe ausweitet oder manipulierte Inhalte als vertrauenswürdiger Kontext wirken.

Die Sicherheitsgrenze umfasst damit nicht nur Modell und Anwendung, sondern auch Ingestion, Index, Vector Store und Retrieval. Quellenherkunft und Berechtigungen müssen in dieser Kette erhalten bleiben.

ID	Requirement	Applicability	Evidence
AI-SB-RAG-01	Jede Wissensbasis hat Owner, Zweck, Datenklassifikation, Aktualisierungsweg und Löschweg.	E, C, H	Knowledge-base record
AI-SB-RAG-02	Retrieval respektiert Quellberechtigungen oder setzt eine gleichwertige technische Zugriffskontrolle durch.	E, C, H	Permission inheritance test
AI-SB-RAG-03	Quellenintegrität, Aktualität und Herkunft sind vor Aufnahme in Index oder Vector Store geprüft.	E, C, H	Source validation record
AI-SB-RAG-04	Embeddings und Vector Stores sind gegen unbefugten Zugriff, Export und Cross-Tenant Leakage geschützt.	E, C, H	Vector database access review
AI-SB-RAG-05	Retrieval Poisoning, unerwartete Quellenmischung und veraltete Quellen sind testbar und überwachbar.	C, H	RAG abuse test
AI-SB-RAG-06	Antworten aus RAG-Systemen können relevante Quellen oder Retrieval-Bezüge nachvollziehbar machen.	E, C, H	Citation or retrieval trace sample

7. Model, Prompt und Output Security

Modelle, Prompts und Outputs sind steuernde technische Elemente. Sie brauchen Schutz, Versionierung und Tests gegen Manipulation, Missbrauch und unbeabsichtigte Offenlegung.

Da probabilistische Ausgaben nicht allein durch klassische Codeprüfung abgesichert werden können, verbindet dieser Block Artefaktkontrolle mit risikobasierten Tests und validierbaren Ausgabegrenzen.

ID	Requirement	Applicability	Evidence
AI-SB-MOD-01	Modellherkunft, Modellversion, Anbieter, Lizenz- oder Nutzungsbedingungen und bekannte Einschränkungen sind dokumentiert.	B, E, C, H	Model record

3 Retrieval-Augmented Generation verbindet generative Modelle mit abgerufenen Inhalten aus externen oder internen Wissensquellen. Die Berechtigungs- und Herkunftskontrolle der Quellen bleibt deshalb Teil der Sicherheitsgrenze.

ID	Requirement	Applicability	Evidence
AI-SB-MOD-02	System Prompts, Developer Instructions und Policy Prompts sind geschützt, versioniert und change-kontrolliert.	E, C, H	Prompt version history
AI-SB-MOD-03	Direkte und indirekte Prompt Injection[4] wird vor produktiver Nutzung risikobasiert getestet.	E, C, H	Prompt injection test report
AI-SB-MOD-04	Output-Filter, Validierungsregeln oder Review-Schritte begrenzen schädliche, vertrauliche oder fachlich ungeprüfte Ausgaben.	E, C, H	Output control evidence
AI-SB-MOD-05	Fine-Tuning, LoRA oder vergleichbare Modellanpassungen haben getrennte Daten-, Modell- und Freigabenachweise.	C, H	Tuning record
AI-SB-MOD-06	Modellwechsel oder relevante Promptänderungen lösen Regressionstest und Revalidierung aus.	E, C, H	Regression test record
AI-SB-MOD-07	Bei bestätigtem Art.-50-Abs.-2-Trigger werden synthetische Audio-, Bild-, Video- und Textausgaben technisch machbar maschinenlesbar markiert; Erkennbarkeit, Interoperabilität, Robustheit und Zuverlässigkeit werden getestet.	Provider Activation und Art.-50-Abs.-2-Trigger	Marking design, detection and interoperability test
AI-SB-MOD-08	Technische Benutzeroberflächen setzen freigegebene Interaktions-, Expositions- und Disclosure-Hinweise zum vorgeschriebenen Zeitpunkt zugänglich um.	bestätigter Art.-50-Trigger mit technischer Anzeige	Notice placement and accessibility test

8. Agenten, Tool-Calling und MCP Security

Agenten, Tools und MCP-Server[5] verbinden AI-Systeme mit Daten und Aktionen. Die technische Baseline setzt enge Grenzen für Registrierung, Rechte, Protokollierung und Abschaltung.

Aus Modellantworten werden damit ausführbare Operationen. Tool-Schemas, Transaktionsgrenzen und Stop-Mechanismen begrenzen den möglichen Wirkungsradius und erhalten die technische Eingriffsfähigkeit.

ID	Requirement	Applicability	Evidence
AI-SB-AGT-01	Jeder Agent, jedes Tool und jeder MCP-Server ist registriert, einem Owner zugeordnet und einem freigegebenen AI Use Case zugeordnet.	E, C, H	Tool and MCP inventory
AI-SB-AGT-02	Tool-Schemas beschränken erlaubte Aktionen, Parameter, Datenbereiche und Transaktionsgrenzen.	E, C, H	Tool schema review

4 Prompt Injection bezeichnet die Manipulation des Modellverhaltens durch direkte Eingaben oder indirekt eingebettete Instruktionen. OWASP führt sie als LLM01:2025.

5 Das Model Context Protocol (MCP) verbindet AI-Systeme mit Kontextquellen und Tools; jeder MCP-Server bleibt eine sicherheitsrelevante Integration.

ID	Requirement	Applicability	Evidence
AI-SB-AGT-03	Kritische Aktionen benötigen technische Human-in-the-Loop-Grenzen oder vergleichbare Freigabekontrollen.	C, H	Approval control evidence
AI-SB-AGT-04	Tool-Aufrufe, Parameter, Ergebnisse, Nutzerkontext und Fehler sind vollständig protokolliert.	E, C, H	Tool-call log sample
AI-SB-AGT-05	Agenten und MCP-Server haben Stop-, Disable- oder Quarantine-Mechanismen.	E, C, H	Kill-switch test
AI-SB-AGT-06	Externe Tools und MCP-Server werden wie sicherheitsrelevante Integrationen geprüft.	E, C, H	Connector security review

9. API, Integration und Secrets Security

AI-Integrationen dürfen keine unkontrollierten Geheimnisse, Schattenzugriffe oder unklaren Datenpfade erzeugen. Secrets und Tokens sind technische Hochrisikoartefakte.

Externe Modellendpunkte und flexible Connectoren erweitern klassische API-Risiken um dynamisch erzeugte Inhalte. Kontrollierte Egress-Pfade und verwaltete Credentials halten diese Verbindungen innerhalb der freigegebenen Systemgrenze.

ID	Requirement	Applicability	Evidence
AI-SB-API-01	AI-APIs und Connectoren sind registriert und haben definierte Netzwerk-, Daten- und Authentisierungsgrenzen.	B, E, C, H	Integration record
AI-SB-API-02	API-Schlüssel, Tokens und Secrets liegen in freigegebenem Secrets Management und werden rotiert.	B, E, C, H	Secrets management evidence
AI-SB-API-03	Service Accounts nutzen kurze, zweckgebundene oder kontrolliert rotierte Credentials.	E, C, H	Credential review
AI-SB-API-04	Externe Datenübertragung ist auf freigegebene Endpunkte, Regionen und Datenklassen begrenzt.	E, C, H	Egress control evidence
AI-SB-API-05	Integrationen haben Timeout-, Retry-, Rate-Limit- und Fehlerbehandlungsregeln.	E, C, H	API configuration evidence
AI-SB-API-06	Plugins und Connectoren können bei Incident, Policy-Verstoss oder Supplier Exit deaktiviert werden.	E, C, H	Disable procedure test
AI-SB-API-07	Externe GPAI-Zugriffe werden, soweit technisch durchsetzbar, über freigegebene Gateways, Proxies oder gleichwertige Enforcement Points geführt. Direkte Umgehung wird verhindert oder als genehmigte Ausnahme überwacht.	externe GPAI-Nutzung	Gateway architecture and bypass test

10. Logging, Monitoring und Detection

AI-Systeme müssen beobachtbar sein, ohne unnötig vertrauliche Inhalte in Logs zu verlagern. Monitoring liefert Evidence für Betrieb, Incident Management, Assurance und Revalidierung.

Neben technischen Fehlern sind auch Drift, missbräuchliche Prompts und unerwartete Tool-Aktivität relevant. Telemetrie muss diese Signale erkennbar machen, ohne selbst einen zusätzlichen Datenabfluss zu erzeugen.

ID	Requirement	Applicability	Evidence
AI-SB-LOG-01	Nutzung, Systemereignisse, Modellaufrufe, Tool-Aufrufe, Fehler und sicherheitsrelevante Entscheidungen werden risikobasiert protokolliert.	B, E, C, H	Logging configuration
AI-SB-LOG-02	Logs enthalten keine unnötigen vertraulichen Inhalte und sind gegen unbefugten Zugriff geschützt.	B, E, C, H	Log data review
AI-SB-LOG-03	Monitoring erkennt Missbrauchssignale, ungewöhnliche Nutzung, Datenabflussrisiken, Fehlerraten und Tool-Anomalien.	E, C, H	Monitoring rule set
AI-SB-LOG-04	Kritische Alerts haben definierte Empfänger, Eskalationswege und Testnachweise.	C, H	Alert test
AI-SB-LOG-05	Drift-, Performance- und Qualitätsmetriken werden für relevante AI Use Cases erhoben.	E, C, H	Monitoring dashboard sample
AI-SB-LOG-06	Forensisch relevante Logs werden für Incident Response und Audit ausreichend lang geschützt aufbewahrt.	C, H	Forensic retention evidence
AI-SB-LOG-07	Automatisch erzeugte Logs eines Hochrisiko-Systems unter Kontrolle des Instituts sind exportierbar und werden mindestens sechs Monate oder nach einer längeren anwendbaren Regel geschützt aufbewahrt.	Hochrisiko-Deployer oder Provider	Retention configuration, protected export test

11. Evaluation, Robustheit und Abuse Testing

Technische Evaluation prüft, ob das AI-System gegen relevante Fehlverhalten, Manipulation und Betriebsrisiken widerstandsfähig genug ist. Fachliche Modellvalidierung bleibt im Model-Risk-Kontext; diese Baseline liefert technische Testanforderungen.

Deterministische Funktionstests reichen dafür nicht aus. Abuse Cases, Negativfälle und Regressionstests prüfen, ob Schutzgrenzen auch bei variierenden Eingaben und geänderten AI-Artefakten tragen.

ID	Requirement	Applicability	Evidence
AI-SB-EVAL-01	Für produktive AI Use Cases existiert eine technische Testbasis mit Standardfällen, Negativfällen und Missbrauchsfällen.	E, C, H	Evaluation plan
AI-SB-EVAL-02	Prompt Injection, Jailbreak, Policy Bypass, Datenabfluss und unsichere Tool-Nutzung werden risikobasiert getestet.	E, C, H	Abuse test report
AI-SB-EVAL-03	Regressionstests prüfen relevante Änderungen an Modell, Prompt, Retrieval, Tool, Datenquelle und Anbieter.	E, C, H	Regression test report

ID	Requirement	Applicability	Evidence
AI-SB-EVAL-04	Kritische AI Use Cases haben definierte Schwellenwerte für Abbruch, Einschränkung oder Revalidierung.	C, H	Threshold record
AI-SB-EVAL-05	Testdaten, Testprompts und Ergebnisse werden nachvollziehbar versioniert.	E, C, H	Test artifact repository
AI-SB-EVAL-06	Für Hochrisiko-Systeme sind kontextbezogene Akzeptanzkriterien, erklärte Accuracy-Metriken, bekannte Leistungsgrenzen sowie Robustheits- und Cybersicherheitsergebnisse dokumentiert und freigegeben.	H; Provider-Deliverable bei bestätigter Rolle	Performance and limitation record, approval evidence

12. Secure AI Development und Change Control

AI-spezifische Änderungen können in Code, Prompt, Modell oder Retrieval entstehen. Sie können auch Konfiguration, Tool-Schema oder Anbieterleistung betreffen. Change Control muss diese Artefakte erkennen.

Der Release-Pfad muss deshalb mehr als Softwarestände erfassen. Erst die gemeinsame Versionierung der wirksamen AI-Artefakte ermöglicht reproduzierbare Tests, Rollback und gezielte Revalidierung.

ID	Requirement	Applicability	Evidence
AI-SB-SDLC-01	AI-Artefakte sind versioniert und einem kontrollierten Change-Prozess zugeordnet.	B, E, C, H	AI change record
AI-SB-SDLC-02	Entwicklung, Test, Pilot und Produktion haben getrennte Daten- und Zugriffskontrollen.	B, E, C, H	Environment segregation evidence
AI-SB-SDLC-03	Code, Prompt, Tool-Schema, Retrieval-Pipeline und Infrastrukturänderungen werden vor Release geprüft.	E, C, H	Release checklist
AI-SB-SDLC-04	Abhängigkeiten, Container, Libraries und Modellartefakte werden auf bekannte Schwachstellen geprüft.	E, C, H	Vulnerability scan
AI-SB-SDLC-05	Relevante Releases haben Rollback-, Disable- oder Fallback-Optionen.	E, C, H	Rollback evidence
AI-SB-SDLC-06	Bei bestätigter Hochrisiko-Providerrolle liefern Entwicklung und Betrieb versionierte technische Inputs für Anhang IV, Betriebsanleitung, Konformitätsdossier und Post-Market-Monitoring-Plan.	Provider-Aktivierungsentscheid positiv	Provider technical dossier index

ID	Requirement	Applicability	Evidence
AI-SB-SDLC-07	Bei bestätigter Hochrisiko-Providerrolle dokumentiert das Institut Designentscheidungen, Herkunft, Aufbereitung, Annahmen, Eignung, Repräsentativität, Fehler und Bias-Minderung für Trainings-, Validierungs- und Testdaten.	Provider Activation und eigene Datenverantwortung	Training, validation and test data governance record
AI-SB-SDLC-08	Modelle, Weights, Container und kritische Erweiterungen stammen aus freigegebenen Repositories. Herkunft und Integrität werden vor Deployment durch Signatur, Hash oder einen gleichwertigen unveränderbaren Nachweis geprüft.	E, C, H	Approved repository record, signature or hash verification

13. Supplier, Cloud und Managed AI Services

Supplier- und Vertragsentscheidungen liegen im AI Operations Handbook. Diese Baseline definiert technische Nachweise, die vor und während externer AI-Nutzung benötigt werden.

Bei Managed AI Services liegen zentrale Komponenten ausserhalb der eigenen technischen Kontrolle. Anbieterinformationen, Änderungswege und Exit-Fähigkeit schliessen diese Lücke für Security Review und laufenden Betrieb.

ID	Requirement	Applicability	Evidence
AI-SB-SUP-01	Managed AI Services liefern technische Informationen zu Datenverarbeitung, Logging, Training, Speicherort, Subprozessoren und Zugriffen.	B, E, C, H	Supplier technical evidence pack
AI-SB-SUP-02	Anbieterzugriffe auf Daten, Logs, Supportfälle und Modellartefakte sind begrenzt und nachvollziehbar.	E, C, H	Supplier access evidence
AI-SB-SUP-03	Externe Modell- oder Plattformänderungen haben Benachrichtigungs-, Test- und Revalidierungswege.	E, C, H	Supplier change evidence
AI-SB-SUP-04	Exit, Datenrückgabe, Datenlöschung und Export technischer Artefakte sind technisch unterstützt.	E, C, H	Exit evidence
AI-SB-SUP-05	Kritische Supplier haben Nachweise zu Resilienz, Security Testing, Incident Support und Auditfähigkeit.	C, H	Supplier assurance evidence

14. Incident Readiness und Forensics

Die Baseline stellt sicher, dass technische Eindämmung, Analyse und Wiederherstellung möglich sind. Der Incident Workflow selbst liegt im AI Operations Handbook.

AI-Incidents können Modell, Prompt, Retrieval und Tool-Ausführung gleichzeitig betreffen. Forensische Spuren und selektive Abschaltmöglichkeiten müssen diese Komponenten deshalb in ihrem tatsächlichen Zusammenspiel erfassen.

ID	Requirement	Applicability	Evidence
AI-SB-IR-01	AI-Systeme haben technische Kontaktpunkte und Logs für Security Incident Response.	B, E, C, H	Incident readiness record
AI-SB-IR-02	Agenten, Tools, Connectoren, RAG-Quellen und Modellzugriffe können bei Incident eingeschränkt oder deaktiviert werden.	E, C, H	Containment test
AI-SB-IR-03	Forensische Daten erfassen relevante Prompts, Antworten, Tool-Aufrufe, Retrieval-Spuren, Modellversionen und Nutzerkontext risikobasiert.	C, H	Forensic log sample
AI-SB-IR-04	Vor Rückkehr in den Normalbetrieb sind technische Korrektur, Regressionstest und Revalidierung nachgewiesen.	E, C, H	Recovery test record
AI-SB-IR-05	Datenabfluss, Prompt Injection, Retrieval Poisoning und Tool-Missbrauch haben erkennbare technische Triage-Signale.	E, C, H	Detection use-case list

15. Evidence und Nachweisführung

Technische Evidence muss so abgelegt sein, dass sie für Safeguard Testing, Betriebsfreigabe, Revalidierung, Incident Management und Audit nutzbar bleibt. Jede Evidence bleibt einem Requirement, einem AI-System und dem gültigen technischen Stand eindeutig zugeordnet.

Der AI Control Standard verwendet diese Nachweise zur Prüfung der Safeguards. Das AI Operations Handbook bindet sie als Ergebnis in die einschlägigen Workflows und Handoffs ein. Die Baseline führt keine parallele Ablage ein.

Evidence class	Examples	Primary consumer
Architecture evidence	Architecture diagram, data-flow diagram, control point map	AI Control Standard, Operations Handbook
Access evidence	Access matrix, admin review, tool permission record	AI Control Standard
Data evidence	Classification record, training exclusion, retention settings	AI Control Standard, Operations Handbook
RAG evidence	Knowledge-base record, permission inheritance test, retrieval trace	AI Control Standard, Operations Handbook
Agent evidence	Tool inventory, tool schema review, tool-call log, kill-switch test	AI Control Standard, Operations Handbook
Monitoring evidence	Log configuration, alert test, monitoring dashboard sample	AI Control Standard, Reporting Workflow im AI Operations Handbook

Test- und Supplier-Nachweise bilden eine eigene, bei Seitenwechsel vollständig wiederholbare Tabelle.

Evidence class	Examples	Primary consumer
Test evidence	Abuse test, regression test, vulnerability scan, threshold record	AI Control Standard

Evidence class	Examples	Primary consumer
Supplier evidence	Supplier technical evidence pack, supplier access review, exit evidence	Supplier Workflow im AI Operations Handbook, AI Control Standard

16. Baseline Levels und Ausschlussstatus

Die Profilklassen B, E, C und H werden im AI Control Standard geprüft und im AI Operations Handbook angewendet. Diese Baseline ordnet ihnen technische Mindesttiefen zu. X wird getrennt davon als Ausschlussstatus geführt.

Die Stufen beschreiben die erwartete Tiefe, nicht vier getrennte Architekturen. Höhere Profilklassen bauen auf den grundlegenden Kontrollen auf und ergänzen stärkere Evidence, Überwachung und Eingriffsmöglichkeiten dort, wo Wirkung und Angriffsfläche dies verlangen.

Profilklasse	Technical baseline level	Minimum technical expectation
B	Basic	Registrierte Nutzung, bekannte Datenklasse, kontrollierter Zugriff, minimale Logging- und Löschnachweise
E	Enhanced	Architektur, Datenfluss, Prompt- und Modellversion, Monitoring, Missbrauchstest und Revalidierungsauslöser
C	Critical	Erweiterte Resilienz, strengere Zugriffskontrolle, forensische Logs, technische Stop-Kontrollen und Supplier Evidence
H	High Risk	High-risk-relevante technische Nachweise, robuste Human Oversight Controls, enges Monitoring und erhöhte Änderungsprüfung

Technical area	Basic (B)	Enhanced (E)	Critical (C)	High Risk (H)
Architecture	basic record	data-flow diagram	control point map	high-risk evidence
Identity	standard access	role-based matrix	privileged review	strict approval controls
Data	classification	retention and leakage controls	forensic retention	high-risk data evidence
RAG	if used	source and permission checks	poisoning tests	high-risk source review
Agents	if used	tool inventory	kill-switch and Human in the Loop (HITL)	high-risk tool constraints
Monitoring	basic logging	alerting and metrics	forensic signals	high-risk monitoring
Testing	basic check	abuse and regression tests	threshold tests	high-risk evaluation

Technische Massnahmen bei Ausschlussstatus X

X darf weder durch eine höhere Kontrolltiefe noch durch Risk Acceptance in einen freigabefähigen Zustand überführt werden. Die folgenden Massnahmen sichern den Ausschlussentscheid technisch ab und verweisen auf einschlägige Requirements dieser Baseline.

ID	Verbindliche X-Massnahme	Referenzierte Requirements	Evidence
AI-SB-X-01	Das Institut blockiert Nutzung, Deployment und produktive Freigabe.	AI-SB-PLT-02, AI-SB-PLT-04	Block- oder Deny-Test, X Decision Record
AI-SB-X-02	Das Institut deaktiviert Nutzerzugänge, Agenten, Tools, APIs, Plugins und Connectoren.	AI-SB-IAM-06, AI-SB-AGT-05, AI-SB-API-06, AI-SB-IR-02	Deprovisioning Log, Disable- oder Containment-Test
AI-SB-X-03	Das Institut isoliert, löscht oder archiviert Daten und technische Artefakte kontrolliert.	AI-SB-DATA-05, AI-SB-SUP-04	Lösch-, Isolations- oder Archivierungsnachweis
AI-SB-X-04	Das Institut stoppt externe Übertragung und Supplier-Zugriffe und löst notwendige Exit-Schritte aus.	AI-SB-API-04, AI-SB-SUP-02, AI-SB-SUP-04	Egress Block Evidence, Supplier- oder Exit-Nachweis
AI-SB-X-05	Das Institut dokumentiert Containment und Stilllegung mit Owner, Frist und Restaufgaben.	AI-SB-IR-01, AI-SB-IR-04	Containment Record, Stilllegungs-Record
AI-SB-X-06	Das Institut aktualisiert das AI Register und lässt den technischen Abschluss unabhängig prüfen.	AI-SB-ARCH-05, AI-SB-LOG-01	Registerstatus, X Closure Review

17. Glossar

Dieses Glossar führt nur technische Begriffe, die für Umsetzung und Nachweisführung benötigt werden. Das AI Governance Framework for Financial Institutions bleibt für Rollen, Lifecycle, Freigaben und weitere Governance-Begriffe führend. Der AI Control Standard definiert Safeguards; das AI Operations Handbook definiert Workflows.

Term	Meaning in this baseline
AI artifact	Technisches Artefakt wie Modell, Prompt, Tool-Schema, Retrieval-Index, Log oder Evaluation Set
AI	Artificial Intelligence; in der Dokumentfamilie einheitlich verwendete Bezeichnung für künstliche Intelligenz.
Agent	AI-Komponente, die Tools auswählt oder Aktionen auslösen kann
Evidence	Nachweis, der Implementierung, Betrieb oder Prüfung eines Requirements belegt
MCP server	Integrationsschicht, die Modelle mit Tools, Daten oder Kontextquellen verbindet
RAG	Retrieval-Augmented Generation mit angebundener Wissensbasis oder Vector Database
Requirement	Verbindliche technische Mindestanforderung dieser Baseline
Examine	Prüfung von Dokumenten, Konfigurationen, Logs oder anderen Artefakten gegen das Requirement.
Interview	Befragung zuständiger Rollen zur Bestätigung von Verantwortung und gelebtem Ablauf.
Test	Reproduzierbare technische Prüfung, ob das verlangte Verhalten tatsächlich eintritt.
Vector database	Speicher- und Suchkomponente für Embeddings

Term	Meaning in this baseline
API	Application Programming Interface; technische Schnittstelle zwischen Systemen oder Services.
SaaS	Software as a Service; extern betriebene Softwareleistung.
HITL	Human in the Loop; vorgesehene menschliche Beteiligung an einer technischen Aktion oder Entscheidung.
CISO	Chief Information Security Officer; Führungsrolle für Informationssicherheit und zugehörige Sicherheitsrisiken.
Requirement-ID-Präfixe	ARCH Architektur, PLT Plattform, IAM Identität und Zugriff, DATA Daten, RAG Retrieval, MOD Modell, AGT Agenten, API Schnittstellen, LOG Protokollierung, EVAL Evaluation, SDLC Entwicklung, SUP Supplier, IR Incident Response und X Ausschlussmassnahmen.
Digital.gov Plain Language	Offene Methode der U.S. General Services Administration für verständliche technische Erklärungstexte.
4T2 Professional Prose	4T2 Schreibmethode für klare fachliche Erklärungen ausserhalb technischer Requirements.
BSI IT-Grundschutz	Offene BSI-Methode zur Trennung von Sicherheitsanforderung und Umsetzungshinweis.
NIST SP 800-53	Offener NIST-Kontrollkatalog, der als Muster für atomare Sicherheitszustände dient.
NIST SP 800-53A	Offene NIST-Prüfmethode, die als Muster für Evidence und Verifikation dient.
OSCAL	Offenes NIST-Modell für stabile und strukturierte Kontroll- und Anforderungsinformationen.

18. Quellen

Primary regulatory sources validated on 2026-07-11:

- Regulation (EU) 2024/1689 (AI Act), especially Articles 9-15 and 26: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng
- Council of the EU, final adoption of the Digital Omnibus on AI on 29 June 2026; technical applicability dates remain under regulatory watch until publication and entry into force: https://www.consilium.europa.eu/en/press/press-releases/2026/06/29/artificial-intelligence-council-gives-final-green-light-to-simplify-and-streamline-rules/
- Regulation (EU) 2022/2554, Digital Operational Resilience Act (DORA), especially Articles 5-14, 17-23 and 28-30: https://eur-lex.europa.eu/eli/reg/2022/2554/oj/eng
- Regulation (EU) 2016/679, General Data Protection Regulation (GDPR), especially Articles 25, 28 and 32: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

Supporting technical sources:

- Digital.gov Plain Language Guide Series: https://digital.gov/guides/plain-language
- BSI IT-Grundschutz, Anforderungen anpassen: https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Standards-und-Zertifizierung/IT-Grundschutz/Zertifizierte-Informationssicherheit/IT-Grundschutzschulung/

Online-Kurs-IT-Grundschutz/Lektion_5_Modellierung/Lektion_5_05/Lektion_5_05_node.html

- NIST SP 800-53 Rev. 5, NIST SP 800-53A Rev. 5 and NIST OSCAL: https://csrc.nist.gov/pubs/sp/800/53/r5/upd1/final, https://csrc.nist.gov/pubs/sp/800/53/a/r5/final and https://pages.nist.gov/OSCAL/
- NIST AI RMF 1.0 and current NIST AI RMF resources: https://www.nist.gov/itl/ai-risk-management-framework
- NIST AI 600-1, Generative AI Profile: https://doi.org/10.6028/NIST.AI.600-1
- Open Worldwide Application Security Project (OWASP), Top 10 for Large Language Model (LLM) Applications 2025: https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/
- BSI AI Finance Test Criteria Catalogue: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/AI-Finance_Test-Criteria.html

NIST marks AI RMF 1.0 as under revision. OWASP and provider guidance are supporting technical references and remain subject to periodic source watch; they are not treated as binding legal requirements.

19. Haftungs- und Anwendungshinweis

Diese Baseline wurde mit fachlicher Sorgfalt und gegen den angegebenen Rechtsstand erstellt. Eine Gewähr für Vollständigkeit, Fehlerfreiheit oder dauerhafte Aktualität wird nicht übernommen. Sie ist weder Rechtsberatung noch regulatorische Freigabe, Zertifizierung oder Konformitätsbescheinigung.

Die technische Anwendbarkeit und Wirksamkeit jedes Requirements muss für die konkrete Systemversion, Architektur und AI-Nutzung nachgewiesen werden. Die Baseline ersetzt weder die verbindlichen Enterprise Controls noch technische Tests und Freigaben des anwendenden Instituts.

Copyright und offene Lizenz

Transparenz zur Erstellung

Martin Bischoff hat dieses Dokument mit Unterstützung generativer AI erstellt. AI wurde für Formulierungsentwürfe, redaktionelle Überarbeitung, grafische Gestaltung sowie die strukturierte Prüfung auf inhaltliche Vollständigkeit und Konsistenz eingesetzt. Auswahl, fachliche Bewertung und Freigabe erfolgten durch Martin Bischoff. Die AI-Unterstützung ersetzt keine unabhängige Rechts-, Wirksamkeits- oder Konformitätsprüfung.