

MRIS.bd.2.4 KI-Agenten

MRIS.bd: Benutzerdefinierte Schicht — Mythos-resistente Informationssicherheit

Systembaustein der Bausteinfamilie MRIS.bd

Benutzerdefinierter Baustein — keine offizielle Anforderung des BSI

Version 2.0 | Stand: 19.09.2026 | Autor: Richard Peddi

Lizenz: CC BY-SA 4.0 — die Verwendung in IT-Sicherheitskonzepten nach IT-Grundschutz ist ausdrücklich erwünscht; Bearbeitungen und Weitergabe erfolgen unter denselben Bedingungen

Inhaltsverzeichnis

1 Beschreibung	3
1.1 Einleitung	3
1.2 Zielsetzung.....	3
1.3 Abgrenzung und Modellierung	3
2 Gefährdungslage.....	4
2.1 Übernahme von KI-Agenten durch Prompt-Injection und manipulierte Werkzeugdefinitionen	4
2.2 Pauschale Rechte und fehlende Zuordenbarkeit.....	4
2.3 Unkontrollierte agentische Komponenten	4
3 Anforderungen	5
3.1 Basis-Anforderungen	5
MRIS.bd.2.4.A1 Einrichtung einer begrenzten Identität für den KI-Agenten (B).....	5
MRIS.bd.2.4.A2 Protokollierung der Aktionen des KI-Agenten (B).....	5
MRIS.bd.2.4.A3 Sicherstellung der menschlichen Kontrolle über den KI-Agenten (B).....	5
MRIS.bd.2.4.A4 Beschränkung auf freigegebene agentische Komponenten (B)	5
3.2 Standard-Anforderungen	5
MRIS.bd.2.4.A5 Prüfung der Agentensteuerung vor der Produktivsetzung (S).....	6
3.3 Anforderungen bei erhöhtem Schutzbedarf	6
4 Weiterführende Informationen.....	6

1 Beschreibung

1.1 Einleitung

KI-Agenten führen auf Basis von KI-Modellen eigenständig Aktionen aus: Sie lesen und schreiben Daten, rufen Werkzeuge und Schnittstellen auf und stoßen Abläufe in anderen Systemen an. Beispiele sind Coding-Agenten in der Softwareentwicklung, agentische Workflows in Fachanwendungen sowie Agenten, die als Software-as-a-Service oder über Low-Code- und No-Code-Plattformen betrieben werden. Als KI-Agent im Sinne dieses Bausteins gilt auch ein agentischer Ablauf.

Durch ihre Rechte werden KI-Agenten zu privilegierten Zielobjekten. Wird ein Agent übernommen, handelt der Angreifer mit dessen Rechten innerhalb der legitimen Infrastruktur. Wirksam sind strukturelle Barrieren: eine eigene, begrenzte Identität je Agent, eine lückenlose und zuordenbare Protokollierung, menschliche Kontrolle über irreversible Aktionen und die Beschränkung auf freigegebene Komponenten.

Dieser Baustein ist Teil der Bausteinfamilie MRIS.bd. Er setzt die Regelungen des Prozessbausteins MRIS.bd.1 Mythos-resistente Härting des ISMS zum geregelten Einsatz von KI-Agenten technisch am einzelnen Agenten um.

1.2 Zielsetzung

Ziel dieses Bausteins ist es, jeden produktiv eingesetzten KI-Agenten so abzusichern, dass seine Rechte begrenzt, seine Aktionen nachvollziehbar und seine Wirkung jederzeit durch Menschen beherrschbar sind.

1.3 Abgrenzung und Modellierung

Der Baustein MRIS.bd.2.4 KI-Agenten ist auf jeden produktiv eingesetzten KI-Agenten und jeden agentischen Ablauf anzuwenden. Gleichartige KI-Agenten können nach den Regeln der Strukturanalyse zu Gruppen zusammengefasst werden.

Das IT-Grundschutz-Kompendium enthält keinen Baustein für KI-Agenten; Berührungspunkte bestehen zu ORP.4 Identitäts- und Berechtigungsmanagement und OPS.1.1.5 Protokollierung.

Die Anforderungen gelten unabhängig davon, ob die Institution für das zugrunde liegende KI-System Anbieter oder Betreiber im Sinne der KI-Verordnung ist. Wo sich die Umsetzung unterscheidet, knüpft die Anforderung an die Kontrolle über die Steuerungskomponenten des Agenten an (siehe MRIS.bd.2.4.A5). Rolle und Einstufung werden im Inventar nach MRIS.bd.1.A6 erfasst.

Folgende Inhalte sind ebenfalls von Bedeutung und werden an anderer Stelle behandelt:

- Inventar, Rolle und Einstufung nach der KI-Verordnung, verantwortliche Personen und die Freigabeliste für agentische Komponenten (siehe MRIS.bd.1.A6 Geregelter Einsatz von KI-Agenten),
- Workload-Identitäten und identitätsbasierte Zugriffskontrolle auf der Plattform (siehe MRIS.bd.2.3 Ausführungsplattformen) sowie
- die Einhaltung rechtlicher Anforderungen aus der KI-Verordnung (siehe ORP.5 Compliance Management (Anforderungsmanagement)).

Dieser Baustein behandelt nicht

- die Sicherheit eigener KI-Produkte gegenüber Endkunden (Produktsicherheit) sowie
- die Bewertung der Vertrauenswürdigkeit von KI-Modellen und KI-Systemen.

2 Gefährdungslage

Da IT-Grundschutz-Bausteine nicht auf individuelle Informationsverbünde eingehen können, werden zur Darstellung der Gefährdungslage typische Szenarien zugrunde gelegt. Die folgenden spezifischen Bedrohungen und Schwachstellen sind für den Baustein MRIS.bd.2.4 KI-Agenten von besonderer Bedeutung.

2.1 Übernahme von KI-Agenten durch Prompt-Injection und manipulierte Werkzeugdefinitionen

Ein übernommener KI-Agent handelt mit seinen Rechten innerhalb der legitimen Infrastruktur, ohne dass der Angreifer eigene Zugangsdaten benötigt.

KI-Agenten verarbeiten Eingaben aus vielen Quellen: Dokumente, Webseiten, Tickets, Code und Antworten anderer Werkzeuge. Enthält eine dieser Quellen versteckte Anweisungen (Prompt-Injection) oder wird eine Werkzeugdefinition manipuliert, führt der Agent Aktionen im Sinne des Angreifers aus. Wenn der Agent dabei weitreichende Rechte besitzt und irreversible Aktionen ohne menschliche Freigabe ausführen darf, dann entsteht der Schaden, bevor ihn jemand bemerkt. Besonders kritisch ist der Fall des verwirrten Stellvertreters (Confused Deputy): Der Agent setzt seine legitimen Rechte für eine Anfrage ein, die er nicht als fremdgesteuert erkennt.

2.2 Pauschale Rechte und fehlende Zuordenbarkeit

KI-Agenten mit pauschalen Rechten unter persönlichen Benutzerkonten und ohne Audit-Trail sind weder begrenzt noch sauber zuordenbar.

In der Praxis laufen KI-Agenten häufig unter persönlichen Benutzerkonten der Entwickelnden oder mit pauschal vergebenen Rechten, weil das die Einrichtung vereinfacht. Wenn ein Agent unter einem persönlichen Konto handelt, dann lassen sich seine Aktionen weder begrenzen noch eindeutig von denen der Person unterscheiden. Fehlt zudem ein Audit-Trail, lässt sich nach einem Vorfall nicht rekonstruieren, welche Aktionen der Agent ausgeführt und wer sie ausgelöst hat.

2.3 Unkontrollierte agentische Komponenten

MCP-Server, IDE-Erweiterungen und Skills aus ungeprüften Quellen erweitern KI-Agenten um Code, der außerhalb jeder Sicherheitsbetrachtung liegt.

Die Fähigkeiten eines KI-Agenten lassen sich über Komponenten wie MCP-Server, IDE-Erweiterungen und Skills erweitern, die oft aus öffentlichen Quellen stammen und sich selbstständig aktualisieren. Wenn solche Komponenten ohne Prüfung eingebunden werden, dann erhält fremder Code Zugriff auf alles, worauf der Agent zugreifen kann. Eine kompromittierte Komponente kann Daten abfließen lassen oder Werkzeugdefinitionen so verändern, dass der Agent unbemerkt fremde Anweisungen ausführt.

3 Anforderungen

Im Folgenden sind die spezifischen Anforderungen des Bausteins MRIS.bd.2.4 KI-Agenten aufgeführt. Der ISB ist dafür zuständig, dass alle Anforderungen gemäß dem festgelegten Sicherheitskonzept erfüllt und überprüft werden. Bei strategischen Entscheidungen ist der ISB stets einzubeziehen.

Im IT-Grundschutz-Kompendium sind darüber hinaus weitere Rollen definiert. Sie sollten besetzt werden, insofern dies sinnvoll und angemessen ist.

Zuständigkeiten	Rollen
Grundsätzlich zuständig	IT-Betrieb
Weitere Zuständigkeiten	Entwickler, Fachverantwortliche

Genau eine Rolle sollte Grundsätzlich zuständig sein. Darüber hinaus kann es noch Weitere Zuständigkeiten geben. Falls eine dieser weiteren Rollen für die Erfüllung einer Anforderung vorrangig zuständig ist, dann wird diese Rolle hinter der Überschrift der Anforderung in eckigen Klammern aufgeführt.

3.1 Basis-Anforderungen

Die folgenden Anforderungen MÜSSEN für diesen Baustein vorrangig erfüllt werden.

MRIS.bd.2.4.A1 Einrichtung einer begrenzten Identität für den KI-Agenten (B)

Der KI-Agent MUSS eine technische, auf die erforderlichen Rechte begrenzte Identität erhalten. Falls die eingesetzte Plattform keine agentspezifische Identität unterstützt, dann MUSS eine kontrollierte Delegation mit minimalen Rechten über ein eigenes Dienstkonto erfolgen. Die Berechtigungsgrenzen dieser Delegation MÜSSEN dokumentiert werden.

MRIS.bd.2.4.A2 Protokollierung der Aktionen des KI-Agenten (B)

Aktionen des KI-Agenten mit Schreib- oder Netzwerkrechten MÜSSEN lückenlos und der auslösenden menschlichen Identität zuordenbar protokolliert werden. Die Aufbewahrungsdauer der Protokolle SOLLTE unter Berücksichtigung gesetzlicher, vertraglicher und forensischer Anforderungen festgelegt werden.

MRIS.bd.2.4.A3 Sicherstellung der menschlichen Kontrolle über den KI-Agenten (B)

Der KI-Agent MUSS durch die verantwortliche Person abgeschaltet werden können. Irreversible Aktionen des KI-Agenten MÜSSEN eine menschliche Freigabe erfordern.

MRIS.bd.2.4.A4 Beschränkung auf freigegebene agentische Komponenten (B)

Für den KI-Agenten DÜRFEN NUR agentische Komponenten eingesetzt werden, die in der Freigabeliste geführt sind. Diese Beschränkung SOLLTE technisch durchgesetzt werden.

3.2 Standard-Anforderungen

Gemeinsam mit den Basis-Anforderungen entsprechen die folgenden Anforderungen dem Stand der Technik für diesen Baustein. Sie SOLLTEN grundsätzlich erfüllt werden.

MRIS.bd.2.4.A5 Prüfung der Agentensteuerung vor der Produktivsetzung (S)

Falls Prompts, Werkzeugdefinitionen, Retrieval-Pipelines oder Eskalationslogik des KI-Agenten unter der Kontrolle der Institution stehen, dann SOLLTEN diese Komponenten wie Produktivcode einem Code-Review unterzogen werden. Falls diese Komponenten vom Hersteller kontrolliert werden, dann SOLLTEN stattdessen Nachweise des Herstellers zu ihrer Absicherung eingefordert werden. Vor der Produktivsetzung SOLLTE der KI-Agent auf Anfälligkeit für Prompt-Injection und Confused-Deputy-Angriffe getestet werden.

3.3 Anforderungen bei erhöhtem Schutzbedarf

Für diesen Baustein sind keine Anforderungen für einen erhöhten Schutzbedarf definiert.

4 Weiterführende Informationen

Das Framework „Mythos-resistente Informationssicherheit (MRIS)“ (Version 2.0) und der „MRIS Implementation Guide“ (Version 2.0) beschreiben das Mythos-Härtungs-Control MHC-13, das diesem Baustein zugrunde liegt, mit Reifegrad-Pfad, Kennzahlen und typischen Fehlern.

Die Verordnung (EU) 2024/1689 (KI-Verordnung) in der Fassung der Verordnung (EU) 2026/1744 legt für Hochrisiko-KI-Systeme unter anderem Anforderungen an die technische Dokumentation (Art. 11 i. V. m. Anhang IV), die Protokollierung (Art. 12), die menschliche Aufsicht (Art. 14) sowie Genauigkeit, Robustheit und Cybersicherheit (Art. 15) fest; für ihre Einhaltung ist der Anbieter verantwortlich. Betreiber müssen die menschliche Aufsicht geeigneten Personen übertragen (Art. 26 Abs. 2) und automatisch erzeugte Protokolle mindestens sechs Monate aufbewahren (Art. 26 Abs. 6). Diese Rechtspflichten sind keine Anforderungen dieses Bausteins; die maschinenlesbare Fassung der Bausteinfamilie ordnet sie informativ den einzelnen Anforderungen zu.

Der vom BSI am 6. Juli 2026 als Community Draft veröffentlichte Prüfkatalog „AI Audit and Assurance Assessment Architecture“ (A5) bewertet die Vertrauenswürdigkeit von KI-Systemen entlang ihres gesamten Lebenszyklus. Dieser Baustein beschränkt sich demgegenüber auf die Absicherung produktiv eingesetzter KI-Agenten als privilegierte Zielobjekte im Informationsverbund. Beide Werke schließen sich nicht aus, sondern ergänzen sich.