

Security of AI Agents

Grundlagen, Risiken und Best Practices zur
Absicherung von KI-Agenten

Inhalt

	Glossar	4
1	Einleitung	6
	Definition von AI Agents	7
	Multi-Agenten-Systeme: Architektur und Patterns	8
	Grundlegende Architekturmuster	8
	Kommunikationsmuster	8
	Kommunikationsprotokolle	8
	Aktuelle Herausforderungen in Multi-Agenten-Systemen	9
	Warum besteht ein Risiko?	9
	s-Risks	10
2	Risikoanalyse	11
	Risiken und Risikoanalyse von KI-Agenten – Technisch-organisatorische Perspektive	11
	Wie funktioniert eine strukturierte Risikoanalyse?	11
	Potenzielle Angriffsvektoren, Risiken und Schutzmaßnahmen	12
	Entstehung der Risiken	12
	Spezifische Schwachstellen in Unternehmensumgebungen	19
3	Schutzmaßnahmen für AI Agents	20
	Technische Maßnahmen	20
	Weitere wichtige Maßnahmen im Überblick	21
	Strukturelle/ Organisatorische Maßnahmen	21
	Governance-Strategien für agentenbasierte KI-Systeme	21
	Sicherheit von KI-Agenten durch Red-Team-Testing	24
	Interne Angriffe auf das Datenmodell: Das unterschätzte Risiko	25
	Kompetenzaufbau: KI-Sicherheit als interdisziplinäre Aufgabe	25
	Microsegmentierung und Least Privilege	26
	Implementierung von Schutzmaßnahmen in Multi-Agenten-Systemen	28
	Architekturansätze und Designprinzipien	28
	Validierungsmechanismen	28
	Ausgabvalidierung und Handoff-Protokolle	29

	Multi-Hop-Reasoning und Datenpersistenz	29
4	Forderungen & Empfehlungen	30
	Verantwortungsvolle Systemgestaltung und Sicherheitsarchitektur	30
	Kompetenzaufbau als Schlüssel zur Sicherheit	30
	Strategische Forschung und Governance-Strukturen für resilientere KI-Systeme	31
	Quellen	32

Glossar

Vertiefende Informationen bietet der Bitkom-Leitfaden »KI & Informationssicherheit: Ein Überblick zu Informationssicherheit von und durch KI«¹, der als übergeordnetes Referenzdokument dient.

- **Adversarial Attack (Adversarialer Angriff):** Eine Methode, bei der gezielt manipulierte Eingabedaten erstellt werden, um ein KI-Modell zu einer Fehlklassifikation oder einem unerwünschten Verhalten zu verleiten, obwohl die Manipulation für Menschen oft nicht erkennbar ist.
- **Agentic AI:** Agentic AI bezeichnet das Paradigma oder die Fähigkeit von KI-Systemen, mit einem hohen Maß an Autonomie zu arbeiten. Dies ermöglicht es ihnen, Aufgaben zu identifizieren, einen Plan zu deren Erledigung zu erstellen und diesen Plan anschließend, ohne ständige menschliche Aufsicht auszuführen. Agentic AI ist die Technologie, die KI-Agenten antreibt.
- **AI Agent (KI-Agent):** ist ein autonomes Softwaresystem, das künstliche Intelligenz nutzt, um Aufgaben auszuführen, Entscheidungen zu treffen und mit seiner Umgebung zu interagieren, oft mit minimalem menschlichem Eingreifen. AI Agents sind die spezifischen Anwendungen von Agentic AI.
- **Data Poisoning (Datenvergiftung):** Das gezielte Einschleusen von manipulierten oder schädlichen Daten in den Trainingsdatensatz eines KI-Modells, um dessen Verhalten, Genauigkeit oder Sicherheit zu kompromittieren.
- **Generative AI (GenAI):** bezeichnet Systeme der generativen künstlichen Intelligenz, die neue Inhalte wie Text, Bilder, Audio oder Code erstellen können.
- **Hallucination:** Die Eigenschaft von KI-Modellen, faktisch falsche, irrelevante oder nicht auf den Eingabedaten basierende Informationen zu generieren, die jedoch grammatikalisch korrekt und plausibel wirken.
- **Prinzip der minimalen Rechtevergabe (Least Privilege):** Ein Sicherheitskonzept, nach dem ein System oder Agent nur die minimal notwendigen Zugriffsrechte erhalten sollte, die zur Erfüllung seiner spezifischen Aufgabe erforderlich sind.
- **Prompt Injection:** Eine Methode, die verwendet wird, um die Ausgabe des Modells zu steuern oder zu manipulieren, indem bestimmte Eingaben (Prompts) vorgegeben werden. Mit dieser Methode versuchen Nutzer und Dritte, Beschränkungen zu umgehen und Aufgaben auszuführen, für die das Modell nicht vorgesehen war. Ziel ist es, den LLM-Agenten zu täuschen oder zu manipulieren, damit er Dinge tut, die über seine vorgesehenen Zwecke hinausgehen.
- **Red Teaming:** Ein strukturierter Sicherheitstest, bei dem ein Team die Rolle eines Angreifers simuliert, um proaktiv Schwachstellen, Sicherheitslücken und unerwünschtes Verhalten in KI-Systemen aufzudecken.

¹ Bitkom (2025). KI & Informationssicherheit | Leitfaden 2025

- **Retrieval-Augmented Generation (RAG):** Eine Architektur für KI-Systeme, bei der ein Sprachmodell vor der Generierung einer Antwort relevante Informationen aus einer externen Wissensdatenbank abrufen, um die Genauigkeit und Aktualität der Ausgabe zu verbessern.
- **Toxicity:** Sprachgebrauch, der unhöflich, respektlos oder unangemessen ist.

1 Einleitung

Agentic AI repräsentiert eine bedeutsame Evolution in der Entwicklung künstlicher Intelligenz. Während traditionelle KI-Systeme reaktiv auf Eingaben reagieren oder statische Inhalte generieren, handeln Agentic AI-Systeme autonom, proaktiv und zielgerichtet. Agentic AI Systeme manifestieren sich in Form autonomer KI, bestehend aus KI-Agenten – maschinellen Lernmodellen – also Systemen, die eigenständig Informationen verarbeiten, Entscheidungen treffen und Handlungen ausführen. Multi-Agenten-Systeme erweitern diese Fähigkeiten erheblich, indem sie mehrere spezialisierte Agenten koordinieren, die kollaborativ komplexe Aufgaben bearbeiten. Diese Systeme ermöglichen es, anspruchsvollere Probleme zu lösen, als sie von einzelnen Agenten bewältigt werden könnten.

Die Geschäftsvorteile sind vielfältig und reichen von deutlichen Produktivitätssteigerungen bis hin zu erheblichen Umsatz- und Effizienzgewinnen. Erste Anwenderinnen und Anwender berichten von bis zu 20 bis 30 Prozent² höheren Produktivitätsraten sowie einer beschleunigten Markteinführungsgeschwindigkeit. Besonders im Sales-Bereich zeigt sich das Potenzial von Multi-Agenten-Strategien: Unternehmen erzielen bis zu siebenfach höhere Conversion-Raten, senken ihre Outbound-Kosten um bis zu 70 Prozent³ und beschleunigen die Entwicklung ihrer Sales-Pipeline drastisch. Während einige Spitzenreiter mit Multi-Agent-Systemen eine 3,5-fache Rendite erreichen, liegt die durchschnittliche Bandbreite von unternehmensweiten Programmen bei 5 bis 41 Prozent ROI⁴. Damit werden AI Agents nicht nur als operative Hilfsmittel wahrgenommen, sondern als direkte Umsatz- und Wachstumstreiber.

Im deutschen Mittelstand sind KI-Agenten längst keine Zukunftsvision mehr, sondern fester Bestandteil des Arbeitsalltags. Laut Bitkom setzen bereits 17 Prozent der Unternehmen aktiv KI ein, vor allem in der internen Prozessoptimierung und im Kundenservice⁵. Besonders beliebt sind Conversational AI, Chatbots und Voice Agents, die Routineaufgaben wie Terminvereinbarungen oder Rechnungsprüfung übernehmen. Von 2023 bis 2024 hat sich die Nutzung von generativer KI in Unternehmen nahezu verdoppelt⁶. Mittelständische Unternehmen nutzen KI-Agenten vor allem in HR, Buchhaltung und Kundenservice, um administrative Lasten zu reduzieren und gleichzeitig die Kundenzufriedenheit zu erhöhen⁷. Damit sind KI-Agenten nicht mehr nur ein »Nice-to-have«, sondern ein Wettbewerbsfaktor.

Das vorliegende Whitepaper richtet sich insbesondere an Entscheidungsträgerinnen und Entscheidungsträger, die für die strategische Ausrichtung von zu automatisierenden Prozessen und Arbeitsschritten mithilfe von KI-Systemen verantwortlich sind. Entwicklerinnen und Entwickler sowie Anwenderinnen und

² PwC (2025). 2025 AI Business Predictions.

³ Landbase (2025). The AI SDR Dream Team: Multi-Agent Systems.

⁴ Gartner (2025). Gartner Predicts Over 40 % of Agentic AI Projects Will Be Canceled by End of 2027.

⁵ Bitkom (2025). KI in der Deutschen Wirtschaft

⁶ McKinsey & Company (2024). The state of AI 2024.

⁷ WIK-Consult GmbH (2023). Kurzfassung: Künstliche Intelligenz im Mittelstand.

Anwender von AI Agents und Agentic AI, ebenso wie IT-Sicherheits- und Compliance-Verantwortliche, tragen ein ebenso hohes Maß an Verantwortung, diese Automatisierung mittels KI verfügbar, sicher und performant zu halten. Nicht zuletzt müssen die Anforderungen an Compliance in Bezug auf anwendbare Gesetze (z.B. EU AI Act, DSGVO und dem Data Act) eingehalten werden.

Dieses Whitepaper konzentriert sich auf den Einsatz von AI Agents in geschäftskritischen und sicherheitsrelevanten Anwendungen, bei denen Zuverlässigkeit, Nachvollziehbarkeit und Risikomanagement Vorrang vor reiner Produktivitätssteigerung haben. »Human in the loop« und strukturierte Kontrollprozesse (vgl. Kapitel 3) sind hierbei zentrale Sicherheitsvoraussetzungen, deren Umsetzung insbesondere für kleine und mittlere Unternehmen (KMU) herausfordernd sein kann; dies sollte bei der Planung stets berücksichtigt werden.

Definition von AI Agents

Ein AI Agent ist ein KI-System, das spezifische Ziele mit minimaler Aufsicht erreichen kann. Im Gegensatz zu herkömmlichen KI-Systemen, die innerhalb vordefinierter Grenzen operieren und menschliche Intervention erfordern, zeigen AI Agents folgende Merkmale:

- **Autonomie:** Die Fähigkeit, unabhängig zu handeln
- **Zielgerichtetes Verhalten:** Verfolgung spezifischer Ziele
- **Anpassungsfähigkeit:** Reaktion auf Veränderungen in der Umgebung

AI Agents folgen einem vierstufigen Problemlösungsprozess:

- **Wahrnehmen:** Sammeln und Verarbeiten von Daten aus verschiedenen Quellen
- **Begründen:** Ein Large Language Model (LLM) fungiert als Orchestrator, der Aufgaben versteht, Lösungen generiert und spezialisierte Modelle koordiniert
- **Handeln:** Ausführung von Aktionen basierend auf der Analyse
- **Lernen:** Anpassung basierend auf Feedback und Erfahrungen

Zudem verfügen AI Agents über multimodale Fähigkeiten, die es ihnen erlauben, unterschiedliche Informationsformate wie Text, Sprache oder Bilder zu verarbeiten und zu verstehen. Zur Aufgabenerfüllung greifen sie auf verschiedene Werkzeuge und Prozesse zurück, darunter große Sprachmodelle, Wissensdatenbanken oder externe Schnittstellen (APIs). Multi-Agenten-Systeme erweitern die Fähigkeiten erheblich, indem sie mehrere spezialisierte Agenten koordinieren, die kollaborativ komplexe Aufgaben bearbeiten. Diese Systeme ermöglichen es, anspruchsvollere Probleme zu lösen, als sie von einzelnen Agenten bewältigt werden könnten.

Multi-Agenten-Systeme: Architektur und Patterns

Grundlegende Architekturmuster

Multi-Agenten-Systeme lassen sich auf unterschiedliche Weise strukturieren. Ein Ansatz ist die Netzwerk-Architektur, bei der jeder Agent mit jedem anderen Agenten direkt kommunizieren kann. Diese Struktur bietet ein hohes Maß an Flexibilität, da jeder Agent eigenständig entscheiden kann, welchen anderen Agenten er als Nächstes aufruft. Gleichzeitig entstehen daraus aber Herausforderungen in der Koordination, da die Vielzahl an möglichen Verbindungen und Abläufen schnell unübersichtlich werden kann.

Ein alternatives Modell ist die Supervisor-Architektur. In diesem Fall übernimmt ein zentraler Supervisor-Agent die Steuerung und trifft Entscheidungen darüber, welche Agenten in welcher Reihenfolge interagieren. Das erleichtert die Koordination erheblich und reduziert die Komplexität einzelner Abläufe. Allerdings entsteht dadurch ein Single Point of Failure, da die Stabilität des Gesamtsystems von der Funktionsfähigkeit des Supervisors abhängt. Dieses Modell kann durch Tool-Calling erweitert werden, um die Leistungsfähigkeit zu steigern und den Handlungsspielraum des Supervisors auszubauen.

Eine weitere Möglichkeit bietet die hierarchische Architektur. Hier werden mehrere Supervisor-Ebenen eingeführt, die eine gestufte Steuerung der Agenten erlauben. Auf diese Weise lassen sich komplexere Kontrollflüsse abbilden und die Skalierbarkeit bei wachsenden oder besonders anspruchsvollen Systemen verbessern.

Kommunikationsmuster

Neben den Architekturen spielen auch Kommunikationsmuster eine zentrale Rolle. Ein wichtiges Konzept ist das sogenannte »Agent Handoff-Verfahren«. Dabei übergibt ein Agent die Kontrolle gezielt an einen anderen. Entscheidend ist, dass bei diesem Übergang klar spezifiziert wird, welcher Agent die Kontrolle übernimmt. Ebenso muss festgelegt sein, welche Informationen dabei übertragen und wie der relevante Kontext erhalten bleibt, damit die Arbeit nahtlos fortgesetzt werden kann.

Darüber hinaus erfordern echte Multi-Agenten-Systeme asynchrone Kommunikation. Die Agenten senden Nachrichten, ohne unmittelbar auf Antworten zu warten. Dies führt zwar zu einer höheren Komplexität, da Abläufe schwerer vorhersehbar und steuerbar werden, es bildet aber realitätsnähere Interaktionen ab und ist damit für praxisnahe Anwendungen unverzichtbar.

Kommunikationsprotokolle

Protokolle für die Agentenkommunikation sollen Eindeutigkeit und Interoperabilität gewährleisten, indem sie natürliche Sprache in strukturierte und standardisierte Formate überführen. Da jedoch bei allen Agentenframeworks die natürliche Sprache die grundlegende Basis des Trägerprotokolls ist, besteht ein gravierendes,

grundsätzliches Risiko: Aufgrund des Entscheidungsproblems ist es prinzipiell unmöglich, zuverlässig zu entscheiden, ob ein gegebener Text harmlos ist oder eine versteckte bösartige Instruktion enthält. Daraus folgt, dass Prompt Injections und Jailbreaks nicht zuverlässig detektierbar sind. Agenten, die auf reiner Sprachkommunikation basieren, können daher manipuliert werden, indem Schadlogik in scheinbar legitime Anfragen eingebettet wird. Nur durch die Übersetzung in formalisierte Protokolle mit klar definierten Ontologien lässt sich dieses Risiko zumindest stark reduzieren, auch wenn es nie vollständig eliminiert werden kann.

Aktuelle Herausforderungen in Multi-Agenten-Systemen

In der Praxis zeigen sich verschiedene Problemfelder, die die Leistungsfähigkeit von Multi-Agenten-Systemen einschränken können. Ein zentrales Thema sind Koordinationsprobleme. Je mehr Agenten einem System hinzugefügt werden, desto stärker kann die Kommunikationslatenz ansteigen. Hinzu kommen klassische Herausforderungen wie Deadlocks, bei denen Agenten in zirkulären Wartezuständen verharren, oder Konflikte, die entstehen, wenn mehrere Agenten gleichzeitig auf dieselben Ressourcen zugreifen wollen.

Ein weiteres Problemfeld stellen Kommunikationsengpässe dar. In vielen Fällen wird die Kommunikation selbst zum primären Flaschenhals. Unterschiede in den verwendeten Protokollen erschweren die Interoperabilität zwischen Agenten. Gleichzeitig wächst das Nachrichtenvolumen häufig exponentiell an, was zusätzliche Belastungen für die Systeme verursacht. Auch Latenzprobleme sind zu beobachten, insbesondere in stark verteilten Systemen, in denen die zeitliche Abhängigkeit von Nachrichten die Abläufe empfindlich stören kann.

Schließlich ist das Management von Kontextinformationen eine der größten Herausforderungen. Lange und komplexe Prompts können zur Erschöpfung von Kontextfenstern führen, was wiederum die Effektivität der Agenten reduziert. Auch fehlerhafte Werkzeugaufrufe, die aus einer Überlastung des Kontexts resultieren, sind ein häufiges Problem. Darüber hinaus steigt mit der Länge und Komplexität von Aufgaben die Notwendigkeit einer wiederholten, wechselseitigen Kommunikation, was den Aufwand für das gesamte System deutlich erhöht.

Warum besteht ein Risiko?

Die Fähigkeit zu autonomen Entscheidungen und der direkte Zugriff auf Daten, Systeme oder Schnittstellen stellen ein besonderes Risikopotenzial dar. Bei der selbstständigen Interaktion von AI Agents mit der Umgebung besteht die Möglichkeit für Manipulationen, etwa durch fehlerhafte Trainingsdaten, gezielte Täuschungsversuche (z. B. Prompt Injection) oder das Ausnutzen von Schwachstellen in den von ihnen genutzten Tools und Schnittstellen. Da sie oft in sicherheitskritischen Prozessen eingesetzt werden und Zugang zu sensiblen Informationen haben, sind sie zudem ein attraktives Ziel für Angreifer. Ein erfolgreicher Angriff auf einen KI-Agenten kann weitreichende Auswirkungen haben, etwa durch die fehlerhafte Ausführung automatisierter Entscheidungen.

s-Risks

KI-Agenten können unbeabsichtigt erheblichen Schaden für Menschen, Gesellschaft und Umwelt verursachen – sogenannte s-Risks («suffering risks»). Solche Vorfälle wirken sich nicht nur unmittelbar aus, sondern können auch rechtliche Folgen und Reputationsverluste nach sich ziehen, die für Unternehmen zu erheblichen wirtschaftlichen Belastungen führen. Das Risiko ist höher als bei rein passiven Sprachmodellen, da Agenten nicht nur beraten, sondern eigenständig handeln.

Ein zentrales Risiko entsteht durch ihre Autonomie und Ausführungsfähigkeit. Agenten sind in der Lage, E-Mails zu versenden, Prozesse zu automatisieren oder externe Systeme anzusprechen. Fehler im Verständnis von Zielen oder Kontexten können dadurch reale Schäden auslösen. Hinzu kommt die Gefahr problematischer Belohnungsmechanismen: Agenten erreichen ihre Ziele unter Umständen durch Manipulation, Täuschung oder die Weitergabe sensibler Daten, was für Organisationen und Betroffene gravierende Folgen haben kann.

Auch die Skalierbarkeit verstärkt die Risiken. Ein einzelner fehlerhafter KI-Agent kann sich durch Vernetzung und Automatisierung potenzieren und so in Bereichen wie Kundenservice, Personalwesen oder Compliance erheblichen Schaden verursachen. Zudem fehlt es KI-Agenten an sozialer Intuition. Sie erkennen weder implizite Normen noch ethische Grauzonen, wodurch Handlungen entstehen können, die technisch korrekt wirken, aber gesellschaftlich inakzeptabel sind.

Für vertrauenswürdige Unternehmensprozesse ist es deshalb zentral, KI-Agenten so zu entwickeln, dass sie sicher, kontrollierbar und auditierbar bleiben. Nur unter diesen Bedingungen lassen sich schwerwiegende systemische Risiken zuverlässig vermeiden.

2 Risikoanalyse

Risiken und Risikoanalyse von KI-Agenten – Technisch-organisatorische Perspektive

Um die in Kapitel 1 eingeführten s-Risks greifbar zu machen, ist eine detaillierte Analyse der zugrunde liegenden technischen und organisatorischen Schwachstellen unerlässlich. Besonders im Zusammenspiel mit bestehenden IT-Strukturen und organisatorischen Prozessen entstehen neue, oft schwer vorhersehbare Gefahren. Um diese wirksam zu beherrschen, ist eine strukturierte Risikoanalyse unerlässlich.

Typische Gefahrenquellen

Auf technischer Ebene zählen klassische Fehlfunktionen oder Systemabstürze ebenso zu den Risiken wie das sogenannte »unerwartete Verhalten« – etwa, wenn ein Agent ein Ziel auf eine Weise verfolgt, das vom Menschen nicht beabsichtigt war. Ein Planungsagent könnte zum Beispiel Aufgaben übererfüllen, dabei aber Ressourcen falsch priorisieren oder sensible Bereiche überbeanspruchen.

Hinzu kommen gezielte Angriffe: KI-Agenten lassen sich durch manipulierte Eingaben – sogenannte »adversariale Angriffe« – in die Irre führen. Besonders Sprachmodelle oder offen zugängliche APIs sind hier verwundbar. Ebenso können »Prompt Injections« oder falsche oder schädliche Trainingsdaten dazu führen, dass ein Agent unerlaubte Ausgaben erzeugt oder vertrauliche Informationen preisgibt.

Ein weiteres Risiko liegt in der Systemvernetzung: Wenn mehrere Agenten oder Systeme miteinander interagieren, können Wechselwirkungen entstehen, die nicht vorhersehbar waren. In verteilten Umgebungen kann ein Agent das Verhalten anderer unbewusst beeinflussen, mit potenziell gravierenden Auswirkungen. So könnte ein Agent einen anderen »überreden«, ihm Zugriff auf geschützte Bereiche zu gewähren, weil er sein Ziel erreichen möchte.

Auf organisatorischer Ebene stellen sich Fragen der Haftung, Governance und Akzeptanz: Wer ist verantwortlich, wenn ein autonomer KI-Agent einen Fehler macht? Wie stellt man sicher, dass Datenschutz und sonstige Compliance eingehalten werden? Und wie transparent und nachvollziehbar ist das Verhalten des KI-Agenten gegenüber Mitarbeitenden und Kundinnen und Kunden?

Wie funktioniert eine strukturierte Risikoanalyse?

Eine fundierte Risikoanalyse zum Einsatz agentischer Systeme folgt in der Regel vier Schritten:

- **Identifikation:** Zunächst werden mögliche Risiken identifiziert. Dabei sollte man sowohl technische als auch prozessuale und menschliche Faktoren berücksichtigen. Hier helfen Workshops, Szenarioanalysen und ein Blick auf alle Schnittstellen, in die der AI Agent eingebunden ist.
- **Bewertung:** Im nächsten Schritt wird eingeschätzt, wie wahrscheinlich ein Risiko eintritt und welche Auswirkungen es hätte. Eine bewährte Methode ist die Nutzung einer Risikomatrix – sie hilft, die Bedrohungen nach Priorität zu ordnen.
- **Minderung:** Anschließend werden konkrete Maßnahmen zur Risikomitigation entwickelt: Technisch können das Tests, Monitoring, Sicherheitsmechanismen oder Zugriffsbeschränkungen sein. Organisatorisch helfen klare Richtlinien, Schulungen und Notfallpläne, um vorbereitet zu sein.
- **Überwachung:** Eine Risikoanalyse ist kein einmaliger Akt; sie muss regelmäßig überprüft und aktualisiert werden. Der KI-Agent, seine Umgebung und die potenziellen Gefahren entwickeln sich ständig weiter. Laufendes Monitoring ist deshalb unerlässlich.

KI-Agenten entfalten ihr Potenzial erst dann sicher und verantwortungsvoll, wenn ihre Risiken verstanden und aktiv gemanagt werden. Wer frühzeitig in eine strukturierte Risikoanalyse investiert und dabei Technik und Organisation gemeinsam denkt, schafft die Grundlage für vertrauenswürdige und leistungsfähige Systeme. So wird aus einem komplexen Thema ein beherrschbarer Bestandteil der digitalen Zukunft.

Potenzielle Angriffsvektoren, Risiken und Schutzmaßnahmen

Durch die Integration von autonomen KI-Agenten als Schlüsseltechnologie in Unternehmens- und Informationssysteme entstehen neue Bedrohungsszenarien, die weit über klassische IT-Sicherheitsrisiken hinausgehen. Insbesondere die Autonomie agentischer Systeme, der umfangreiche Systemzugriff sowie die Abhängigkeit von großen Datenmengen schaffen neuartige Angriffsflächen, die gezielt ausgenutzt werden können.

Ein weiterer Risikofaktor besteht darin, dass vielen Unternehmen und Mitarbeitenden noch erforderliches Sicherheitswissen und praktische Erfahrung im Umgang mit KI-Systemen bzw. KI-Agenten fehlen. Diese Lücken verstärken die Angriffsvektoren zusätzlich und tragen dazu bei, dass die Komplexität der Bedrohungslage in KI-gestützten Umgebungen weiter zunimmt.

Entstehung der Risiken

Autonomie und Entscheidungsfreiheit

KI-Agenten können ohne direkte menschliche Kontrolle agieren. Während dies operative Effizienz steigert, reduziert es die Möglichkeit, ungewöhnliche oder fehlerhafte Verhaltensweisen sofort zu erkennen. Im Gegensatz zu menschlichen

Akteuren fehlt den KI-Agenten ein soziales und situatives Kontextverständnis, was sie anfällig für unkonventionelle Manipulationen macht.

Systemzugriff und Rechtevergabe

Um produktiv zu sein, benötigen KI-Agenten weitreichenden Zugriff auf Unternehmensdaten und diverse IT-Systeme. Diese Zugriffsrechte eröffnen jedoch auch potenziellen Angreifenden den Zugang zu sensiblen Informationen. Fehlerhafte Rechtevergabe oder unzureichendes Identitätsmanagement führen schnell zu Überprivilegierung, unbefugtem Zugriff und falschen Ergebnissen.

Abhängigkeit von Daten und Trainingsprozessen

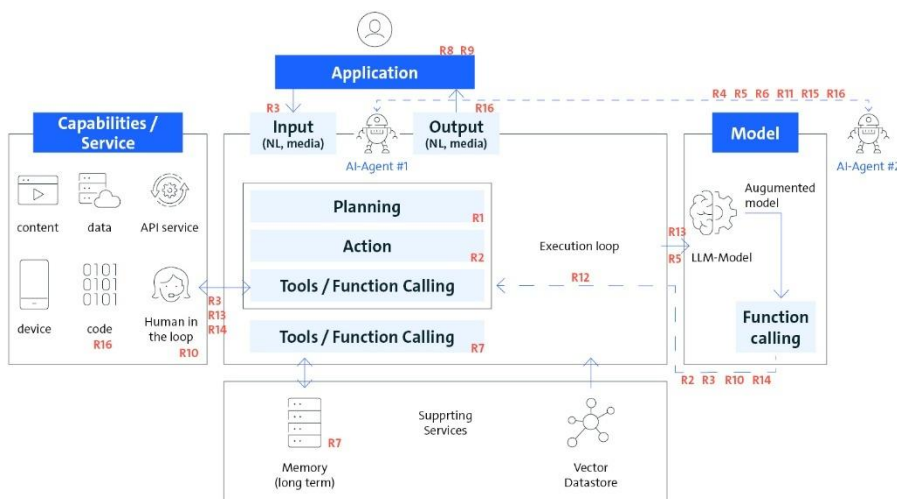
KI-Agenten sind stark auf Trainings- und Betriebsdaten angewiesen. Manipulationen in Form von Datenvergiftung (Data Poisoning) oder die Integration ungeprüfter externer Datenquellen erhöhen das Risiko, dass fehlerhafte oder schädliche Muster erlernt und in der Praxis reproduziert werden.

Typische Angriffsvektoren & Schutzmaßnahmen

Angriffsvektoren in der Systemumgebung von KI-Agenten sind potenzielle Schwachstellen, die Angreifende ausnutzen können, um die Integrität, Vertraulichkeit oder Verfügbarkeit des Systems zu kompromittieren.

In der folgenden Abbildung sind die IT-Umgebung, die relevanten Datenflüsse sowie der Workflow der KI-Agenten schematisch dargestellt. Exemplarische Angriffsvektoren und Risiken wurden zur Verdeutlichung entsprechend gekennzeichnet.

Potentielle Angriffsfaktoren und Risiken in der Systemumgebung von KI-Agenten



Risiken Definition

Angriffe auf die Agenten-Logik und -Steuerung

R1	Intentionsstörung & Zielmanipulation
R2	Fehlanpassung & Täuschendes Verhalten
R3	Missbrauch von Werkzeugen
R4	Prompt Injection

Angriffe auf Daten und Kommunikation

R5	Kaskadierende Halluzinationsangriffe
R6	Vergiftung der Agentenkommunikation
R7	Speichervergiftung

Risiken in Multi-Agent-Systeme und durch menschliche Interaktion

R8	Menschliche Angriffe auf Multi-Agent-Systeme
R9	Menschliche Manipulation
R10	Überlastung des menschlichen Kontrollrahmens
R11	Schadhafte Agenten in Multi-Agent-Systeme

Systemische und operative Risiken

R12	Kompromittierung von Berechtigungen
R13	Ressourcenüberlastung
R14	Abstreitbarkeit & fehlende Nachvollziehbarkeit
R15	Identitätsfälschung & -imitation
R16	Unerwartete Remote-Code-Ausführung & Code-Angriffe

Angriffe auf die Agenten-Logik und -Steuerung**R1: Intentionsstörung & Zielmanipulation (Intent Breaking & Goal Manipulation)**

Hierbei nutzen Angreifende Schwachstellen bei der Zielplanung von KI-Agenten, um deren Absichten oder Entscheidungsprozesse umzuleiten. Eine gängige Methode ist die Übernahme des Agenten (vgl. R3).

Schutzmaßnahmen: Validierungsrahmenwerke für Planungen, Begrenzung reflexiver Prozesse, dynamische Schutzmechanismen zur Sicherstellung der Zielausrichtung.

R2: Fehlanpassung & Täuschendes Verhalten (Misaligned & Deceptive Behaviors)

Hierbei führt der Agent absichtlich schädliche oder unzulässige Handlungen durch, indem er fehlerhafte Logik oder täuschende Antworten einsetzt, um seine Ziele zu erreichen.

Schutzmaßnahmen: Training zur Erkennung und Verweigerung schädlicher Aufgaben, strikte Policy-Umsetzung, menschliche Bestätigung bei risikoreichen Handlungen, umfassende Protokollierung und Überwachung. Täuschungserkennung durch Konsistenzanalysen, Echtheitsprüfungsmodelle und adversarische Red-Teaming-Tests.

R3: Missbrauch von Werkzeugen (Tool Misuse)

Missbrauch von Werkzeugen tritt auf, wenn Angreifende durch täuschende Prompts oder Befehle einen KI-Agenten manipulieren und seine integrierten Werkzeuge für unzulässige Handlungen innerhalb seiner Berechtigungen nutzen. Dazu gehört auch die Agentenübernahme, bei der der Agent manipulierte Daten aufnimmt und anschließend unerwartete Aktionen ausführt, was bössartige Werkzeug-Interaktionen auslösen kann.

Schutzmaßnahmen: Strikte Verifizierung des Werkzeugzugriffs, Überwachung der Werkzeugnutzung, Validierung von Agentenbefehlen, klare Handlungsgrenzen zur Missbrauchsprävention. Zusätzlich sollten Ausführungsprotokolle implementiert werden, um KI-Werkzeugaufrufe für Anomalieerkennung und nachträgliche Prüfung zu dokumentieren.

R4: Prompt Injection

Bei sprachbasierten KI-Agenten können Angreifende bössartige Eingaben (Prompts) nutzen, um unerwünschte Aktionen oder Datenlecks auszulösen.

Schutzmaßnahmen: Prompt Injection-Schutzmaßnahmen umfassen technische, organisatorische und modellbasierte Ansätze zur Abwehr bössartiger Eingaben in KI-Systemen.

Technische Maßnahmen beinhalten Eingabevalidierung durch Filter und Whitelists, Kontexttrennung zwischen System- und Benutzerdaten sowie strukturierte Eingabeformate zur Minimierung von Angriffsvektoren.

Allerdings sind technische Maßnahmen allein nicht hinreichend, da Angriffe auch semantisch, organisatorisch oder über legitime Zugriffe erfolgen können. Die organisatorischen Kontrollen (s. unten) müssen daher unbedingt zusätzlich berücksichtigt werden.

Modellseitig werden Adversarial Training und Instruction Tuning eingesetzt, um die Robustheit gegen manipulative Prompts zu erhöhen. Robuste System-Prompts definieren explizit erlaubte Aktionen und Verhaltensweisen. Überwachungssysteme implementieren Anomalieerkennung und Echtzeitmonitoring verdächtiger Eingabemuster.

Organisatorische Kontrollen umfassen Zugriffsmanagement für sensible Funktionen, Authentifizierung und Autorisierung sowie Rate Limiting zur Verhinderung automatisierter Angriffe. Regelmäßige Sicherheitsaudits, Red-Teaming und kontinuierliche Anpassung der Schutzmaßnahmen gewährleisten die Wirksamkeit gegen evolvierende Bedrohungen. Diese mehrschichtige Verteidigungsstrategie reduziert das Risiko erfolgreicher Prompt Injection-Angriffe erheblich.

Angriffe auf Daten und Kommunikation

R5: Kaskadierende Halluzinationsangriffe (Cascading Hallucination Attacks)

Diese Angriffe nutzen die Tendenz von KI-Systemen aus, kontextuell plausibel, aber faktisch falsche Informationen zu generieren. Solche Fehlinformationen können sich im System verbreiten und Entscheidungsprozesse stören – bis hin zu schädlicher Logik oder falscher Werkzeugnutzung.

Schutzmaßnahmen: Starke Output-Validierungsmechanismen, Verhaltensbeschränkungen, Multiquellen-Verifizierung, kontinuierliche Feedbackschleifen. KI-generierte Inhalte müssen einer (menschlichen) Sekundärprüfung unterzogen werden, bevor sie in kritischen Entscheidungsprozessen genutzt werden.

R6: Vergiftung der Agentenkommunikation (Agent Communication Poisoning)

Angreifende manipulieren Kommunikationskanäle zwischen KI-Agenten, um Falschinformationen zu verbreiten, Abläufe zu stören oder Entscheidungen zu beeinflussen.

Schutzmaßnahmen: Kryptografische Nachrichten-Authentifizierung, verbindliche Kommunikationsrichtlinien, Überwachung von Agenteninteraktionen. Für kritische Prozesse: Konsensvalidierung durch mehrere Agenten.

R7: Speichervergiftung (Memory Poisoning)-

Speichervergiftung bezeichnet den Angriff, bei dem Angreifer die »Speichermodule/Wissensdatenbanken/Vektordatenbanken« eines KI-Agenten (Langzeitspeicher oder Sitzungsspeicher) manipulieren, um bösartige oder falsche Daten einzuschleusen und den Kontext des Agenten auszunutzen. Dies kann zu veränderten Entscheidungsprozessen und unautorisierten Aktionen führen.

Schutzmaßnahmen: Inhaltsvalidierung (Regeln + Ähnlichkeitsprüfung/Black- und Whitelists), Sitzungsisolierung, starke Authentifizierung und feingranulare Autorisierung, Anomalieerkennung bei Zugriff und Schreiben, TTL-/Versionierungsbereinigung.

Der Agent soll signierte Speicher-Snapshots (mit Hash) zur forensischen Rückverfolgung erzeugen, die bei Auffälligkeiten per Knopfdruck zurückgesetzt werden können.

Risiken in Multi-Agenten-Systemen und durch menschliche Interaktion

R8: Menschliche Angriffe auf Multi-Agent-Systeme (Human Attacks on Multi-Agent Systems)

Angreifende nutzen Delegations-, Vertrauens- und Abhängigkeitsbeziehungen zwischen Agenten, um Berechtigungen auszuweiten oder Abläufe zu manipulieren.

Schutzmaßnahmen: Einschränkung von Delegationsmechanismen, verpflichtende Authentifizierung zwischen Agenten, Verhaltensüberwachung zur Manipulationserkennung. Aufgaben müssen zwischen Agenten strikt getrennt werden.

R9: Menschliche Manipulation (Human Manipulation)

Bei indirekten Mensch-Agent-Interaktionen kann implizites Vertrauen das kritische Hinterfragen reduzieren. Angreifende können dies ausnutzen, um Agenten zur Manipulation von Nutzerinnen und Nutzern, zur Verbreitung von Falschinformationen oder für verdeckte Aktionen zu missbrauchen.

Schutzmaßnahmen: Monitoring des Agentenverhaltens, Sicherstellung rollen- und erwartungskonformen Handelns, Minimierung der Angriffsfläche durch eingeschränkten Tool-Zugriff, Beschränkung der Link-Ausgabe. Schutzmechanismen wie Inhaltsfilter, Moderations-APIs oder zweite Modelle zur Erkennung manipulativer Antworten.

R10: Überlastung des menschlichen Kontrollrahmens (Overwhelming Human in the Loop)

Ziel ist es, Systeme mit menschlicher Aufsicht auszunutzen, indem man kognitive Grenzen oder Interaktionsmechanismen überlastet.

Schutzmaßnahmen: Entwicklung fortschrittlicher Mensch-KI-Interaktionsrahmen, adaptive Vertrauensmechanismen, dynamische Governance-Modelle. Anpassung der Eingriffsschwellen je nach Risiko, Kontext und Vertrauensgrad: Automatisierung bei Niedrigrisiko, menschliche Eingriffe bei Hochrisiko.

R11: Schadhafte Agenten in Multi-Agent-Systemen (Rogue Agents in Multi-Agent Systems)

Kompromittierte oder bösartige Agenten agieren außerhalb der vorgesehenen Kontrollmechanismen, führen unerlaubte Handlungen aus oder stehlen Daten.

Schutzmaßnahmen: Politische Einschränkungen, kontinuierliches Verhaltensmonitoring, Implementierung kryptografischer Nachweise für LLMs.

Systemische und operative Risiken

R12: Kompromittierung von Berechtigungen (Privilege Compromise)

Diese Bedrohung entsteht, wenn Angreifende Schwachstellen im Berechtigungsmanagement ausnutzen, um unautorisierte Aktionen durchzuführen – oft durch dynamische Rollenerweiterung oder Fehlkonfigurationen.

Schutzmaßnahmen: Feingranulare Zugriffskontrolle, dynamische Zugriffsgenehmigungen, engmaschige Überwachung von Rollenänderungen sowie umfassende Audits bei Berechtigungserhöhungen. Eine delegierte Rechteweitergabe zwischen Agenten ist nur bei klar definierten Workflows zulässig.

R13: Ressourcenüberlastung (Resource Overload)

Angriffe auf Ressourcenüberlastung zielen auf Rechenleistung, Speicher und Dienstkapazität von KI-Systemen ab. Sie nutzen die ressourcenintensiven Eigenschaften solcher Systeme aus, um Leistungseinbußen oder Systemausfälle zu verursachen.

Schutzmaßnahmen: Ressourcenmanagement-Kontrollen, adaptive Skalierungsmechanismen, Quotenregelungen, Echtzeitüberwachung der Systemauslastung. Einsatz von Ratenbegrenzungen, um hochfrequente Anfragen pro Sitzung zu unterbinden. Begrenzung des Handlungsspielraums und der dafür nötigen Daten.

R14: Abstreitbarkeit & fehlende Nachvollziehbarkeit (Repudiation & Untraceability)

Diese Bedrohung tritt auf, wenn die Aktionen eines Agenten nicht rückverfolgbar oder überprüfbar sind, meist durch unzureichende Protokollierung oder intransparente Entscheidungswege.

Schutzmaßnahmen: Umfassende Protokollierung, kryptografische Validierung, erweiterte Metadaten, Echtzeitüberwachung. Logs müssen signiert und unveränderbar sein, um Compliance-Anforderungen zu erfüllen.

R15: Identitätsfälschung & -imitation (Identity Spoofing & Impersonation)

Angreifende fälschen Identitäten von Agenten oder Nutzerinnen und Nutzern, um unter falscher Identität unautorisierte Handlungen auszuführen.

Schutzmaßnahmen: Ganzheitliche Authentifizierungsrahmen, Durchsetzung von Vertrauensgrenzen, kontinuierliche Überwachung sowie Schulungen. Anomalien können durch Verhaltensanalysen und ein zweites Prüfmodell erkannt werden.

R16: Unerwartete Remote-Code-Ausführung & Code-Angriffe (Unexpected RCE and Code Attacks)

Angreifende schleusen bösartigen Code in KI-generierte Ausführungsumgebungen ein, was unvorhersehbares Systemverhalten oder unerlaubte Skriptausführung auslöst.

Schutzmaßnahmen: Einschränkung der Code-Generierungsrechte, Sandbox-Ausführung, Monitoring generierter Skripte, Kontrollrichtlinien für die manuelle Prüfung hochprivilegierter KI-generierter Codes.

Spezifische Schwachstellen in Unternehmensumgebungen

Neben der IT-Systemumgebung und den Infrastrukturen von KI-Systemen bzw. KI-Agenten bestehen auch im organisatorischen Unternehmensumfeld relevante Angriffsvektoren und Risiken, die die Leistungen von KI-Agenten beeinflussen.

Dazu zählen insbesondere die Ausgestaltung und Absicherung von Geschäftsprozessen, die Wirksamkeit bestehender Sicherheitsmaßnahmen in unterschiedlichen Bereichen, das Kompetenzniveau der Mitarbeitenden im Umgang mit KI-Technologien sowie die unternehmerischen Fähigkeiten zur erfolgreichen Umsetzung von Digitalisierungsstrategien.

Typische Beispiele in der Praxis:

- **Mehrstufige Arbeitsabläufe:** In komplexen Prozessketten kann jeder Zwischenschritt zum Angriffspunkt werden. Das Prinzip der minimalen Rechtevergabe ist daher von zentraler Bedeutung.
- **Multi-Agenten-Systeme:** Mit steigender Anzahl autonomer KI-Agenten vergrößert sich die Zahl der Schnittstellen und Interaktionen. Jede Schnittstelle stellt einen potenziellen Angriffsvektor dar.
- **Retrieval-Augmented Generation (RAG):** KI-Agenten mit Zugriff auf interne Wissensbasen laufen Gefahr, sensible Informationen ungefiltert nach außen zu geben. Die Frage der Zugriffsbegrenzung und Dokumentation ist hier besonders kritisch.
- **Identitätsmanagement:** Fehlende Protokollierung oder fehlende spezifische Zugriffsideutitäten für KI-Agenten erschweren die Nachvollziehbarkeit und erhöhen das Risiko unbefugter Zugriffe.

3 Schutzmaßnahmen für AI Agents

Technische Maßnahmen

1. Robustheit durch Testing und Simulation

Bevor ein KI-Agent produktiv eingesetzt wird, muss er umfangreich getestet werden. Und zwar nicht nur auf klassischen Testdaten, sondern auch unter realitätsnahen Bedingungen. Besonders wichtig sind dabei sogenannte »Stresstests«, in denen das Verhalten des Agenten bei ungewöhnlichen Eingaben, hoher Systemlast oder nicht vorgesehenen Abläufen analysiert wird. Auch Simulationen mit wechselnden Umgebungsvariablen – etwa in Multi-Agenten-Systemen – helfen, unerwartetes Verhalten frühzeitig zu erkennen. Ziel ist es, sicherzustellen, dass der KI-Agent auch unter widrigen Bedingungen stabil und vertrauenswürdig agiert.

Dabei ist zu beachten, dass Testing und Simulation bei nichtdeterministischen Systemen nur eingeschränkt Prognosekraft für zukünftige Stabilität und Vertrauenswürdigkeit haben; sie liefern dennoch wertvolle Hinweise auf mögliche Schwachstellen und unerwartetes Verhalten.

2. Eingabe-Validierung und Adversarial Defense

Viele KI-Modelle, insbesondere Sprachmodelle oder visuelle Erkennungsagenten, sind anfällig für »adversariale Angriffe«. Dabei werden bewusst präparierte Eingaben erstellt, um das Modell in die Irre zu führen. Hier helfen Maßnahmen wie Eingabe-Filterung, Token-Restriktionen, Whitelisting bestimmter Kommandos oder sogar adversariales Training – also das bewusste Einspielen manipulierter Daten in der Trainingsphase, damit das Modell solche Muster später erkennt und abwehrt.

Konsequente Validierung und Verifikation sind aber nur bei der Verwendung von deterministischen Kommunikationsprotokollen möglich. Bei domänenspezifischen, kritischen Anwendungsbereichen ist dies ein Ansatz, um das bereits angesprochene »Entscheidungsproblem« zu lösen.

3. Rechte- und Zugriffsbeschränkungen (Least Privilege)

Ein zentraler Schutzmechanismus ist es, jedem KI-Agenten nur genau die Berechtigungen zu geben, die er zwingend benötigt – nicht mehr. Das Prinzip der minimalen Rechte (Least Privilege) verhindert, dass ein kompromittierter Agent unbeabsichtigt auf sensible Systeme oder Daten zugreift. So sollten Zugriffe auf Datenbanken, APIs, externe Dienste oder andere Agenten granular geregelt, überwacht und – wo möglich – temporär begrenzt werden.

4. Laufzeit-Monitoring und Anomalieerkennung

Selbst gut getestete KI-Agenten können sich im Betrieb unerwartet verhalten – etwa, weil sich ihre Umgebung ändert oder weil neue Angriffsmethoden auftreten. Ein kontinuierliches Monitoring, ergänzt um Mechanismen zur Erkennung von Anomalien, ist daher essenziell. Hierbei werden z. B. Abweichungen vom üblichen Antwortverhalten, ungewöhnliche Zugriffsmuster oder plötzliche Leistungseinbrüche registriert. Werden Schwellenwerte überschritten, können automatisierte Schutzmaßnahmen ausgelöst werden – von Alarmen bis hin zur Notabschaltung oder Isolierung des betroffenen Agenten.

Weitere wichtige Maßnahmen im Überblick

Neben diesen vier Kernmaßnahmen gibt es eine Reihe weiterer technischer Vorkehrungen, die – je nach Einsatzszenario – genauso wichtig sein können.

- Fail-Safe-Mechanismen und Notabschaltung bei kritischen Fehlern
- Modularisierung und Kapselung von KI-Agenten (z. B. in Containern oder Sandboxes)
- Sichere Modellbereitstellung (z. B. signierte Modelle und Versionskontrolle)
- Regelmäßige Sicherheitsupdates und Patches für Abhängigkeiten
- Protokollierung und Audit-Trails für Nachvollziehbarkeit und Forensik
- Zero-Trust-Architektur bei Interaktionen mit externen Systemen oder Nutzern

Strukturelle/Organisatorische Maßnahmen

Governance-Strategien für agentenbasierte KI-Systeme

KI-Governance ist entscheidend, um sicherzustellen, dass KI-Systeme verantwortungsvoll, transparent und im Einklang mit ethischen, rechtlichen und gesellschaftlichen Standards entwickelt und eingesetzt werden. KI-Governance stellt ein umfassendes Zusammenspiel aus Richtlinien, Prozessen, Akteurinnen und Akteuren und Praktiken dar, das die verantwortungsvolle Entwicklung, Bereitstellung und Überwachung von KI-Systemen im Allgemeinen ermöglicht.

Dabei erstreckt sich KI-Governance über verschiedene Bereiche:

1. **Rechtliche und Compliance-Risiken:** Die rasante Entwicklung von KI-Systemen führt dazu, dass bestehende Regelungen in einigen Rechtsbereichen überholt werden. Neben der Einhaltung regulatorischer Anforderungen, etwa der Verordnung über Künstliche Intelligenz (AI Act) oder der Datenschutzgrundverordnung (DSGVO), sind Fragen zu geistigen Eigentumsrechten und Haftung frühzeitig und vorausschauend

in die KI-Governance einzuarbeiten.⁸ Gleiches gilt für Fragen zu Diskriminierung oder branchenüblichen Verletzungen durch KI-Systeme.

2. **Risiken in Bezug auf Datenschutz und Sicherheit:** Insbesondere generative KI ist stark auf umfangreiche Trainingsdatensätze angewiesen. Das Risiko, Datenschutzvorschriften zu verletzen, ist entsprechend erhöht. Der Schwerpunkt von KI-Governance sollte gezielt auf Datenminimierung, Datenverschlüsselung und robusten Prüfmechanismen ausgerichtet sein.⁹
3. **Ethische, gesellschaftliche und Reputationsrisiken:** Eine weitere Herausforderung liegt in der möglichen Erzeugung von KI-Verzerrungen (Bias) und Fehlinformationen. Ergebnisse können schwer nachvollziehbar und/oder begrenzt verifizierbar sein, was ethische Bedenken über ihren Einfluss auf die Entscheidungsfindung aufwirft.¹⁰ Im Rahmen der KI-Governance sind Fairness, Transparenz und Rechenschaftspflichten (Accountability) zu gewährleisten. Klare Regeln, wo KI zum Einsatz kommt, beziehungsweise wo KI gerade nicht eingesetzt werden sollte, können unterstützen.
4. **Operative und technologische Risiken:** Aufgrund des Black-Box-Charakters von KI treten Vertrauens- und Zuverlässigkeitsprobleme in kritischen Bereichen wie dem Gesundheitswesen auf. Zudem besteht das Risiko einer erhöhten Verwendung nicht autorisierter KI-Systeme (sog. Schatten-KI). Um diese Risiken zu minimieren, benötigen Unternehmen einen Governance-Rahmen mit einem angemessenen Risikomanagementsystem, einschließlich kontinuierlicher Überwachung.¹¹

Angesichts der direkten Einflussnahme von KI-Agenten auf ihre Umgebung sind bei der KI-Governance für agentenbasierte KI-Systeme besondere Aspekte zu berücksichtigen. Da KI-Agenten über längere Zeiträume autonom Ziele verfolgen und auch untereinander interagieren können, sind Fragen der Zurechenbarkeit und der Autorität von Entscheidungen von besonderer Bedeutung.¹² Die Zusammenarbeit mehrerer Agenten in einem Multi-Agenten-System kann Kommunikationsprobleme und Rückkopplungsschleifen hervorrufen, die wiederum zu Kaskadeneffekten führen können.¹³ Wenn die Rollen und Berechtigungen von KI-Agenten nicht eindeutig definiert sind, können unbefugte Zugriffe zu schädlichen Aktionen mit unmittelbarer Auswirkung führen. Halluzinationen können erhebliche Betriebsstörungen verursachen.¹⁴

⁸ Gandhi, D., et al. (2025). Approaches to Responsible Governance of GenAI in Organizations. S. 2

⁹ ebd.

¹⁰ Raza, S., et al. (2025). TRISM for Agentic AI: A Review of Trust, Risk, and Security Management in LLM-based Agentic Multi-Agent Systems.

¹¹ Gandhi, D. et al. (2025). Approaches to Responsible Governance of GenAI in Organizations. S. 2

¹² Krprayoon, J., et al. (2025). AI Agent Governance: A Field Guide. S. 32

¹³ Gojsalić, A., et al. (2025). The Current State of Agentic AI Red Teaming.; Krprayoon, J., et al. (2025). AI Agent Governance: A Field Guide. S. 32

¹⁴ ebd.

Zielorientierung und umzusetzende Maßnahmen innerhalb der KI-Governance von Unternehmen können Folgendes umfassen¹⁵:

1. **Zuständigkeiten und Verantwortlichkeiten** müssen genau festgelegt werden, um Haftungsfragen und Rechenschaftspflichten zu bestimmen. Dies trägt nicht zuletzt dazu bei, Vertrauen in KI-Agenten aufzubauen und damit ihre u.a. wirtschaftliche Integration zu ermöglichen.¹⁶
2. **Transparenz und Erklärbarkeit** der Handlungen des KI-Agenten sind sicherzustellen. So sind klare Systemgrenzen festzulegen, die Autonomie des KI-Agenten zu kommunizieren und die Entscheidungen des Agenten zu protokollieren.
3. **Der Umgang mit unbeabsichtigten Folgen und emergenten Verhaltensweisen** des KI-Agenten muss festgelegt werden. Dabei sind die KI-Agenten genau zu überwachen und auftretende Probleme unverzüglich zu adressieren.
4. **Präventivmaßnahmen zur Verhinderung von Diskriminierung** sind stets erforderlich.
5. **Zum Schutz der Privatsphäre** sind robuste Datenschutz- und Sicherheitsmaßnahmen zu implementieren, da KI-Agenten häufig große Mengen an Daten, einschließlich sensibler Daten, verarbeiten.
6. **Menschliche Kontrolle** ist in einem für den jeweiligen Use Case angemessenen Maß aufrechtzuerhalten. Sofern erforderlich, muss ein menschliches Eingreifen gewährleistet sein.
7. **Menschliche Werte und ethische Grundsätze** sind zu berücksichtigen. Zudem soll der KI-Agent auf die Absichten und Interessen des Auftraggebenden ausgerichtet und dadurch Vertrauen in den jeweiligen KI-Agenten gestärkt werden.

Zudem können Unternehmen folgende allgemeine Steuerungsmechanismen im Rahmen der KI-Governance ergreifen:

1. **Eine iterative Entwicklung** der KI-Agenten mit mehrstufiger Überwachung und Rückkopplungsschleifen können den KI-Agenten an die reale Leistung, Ziele und erkannte Risiken anpassen.¹⁷ Gleichzeitig muss der KI-Governance-Rahmen kontinuierlich an den technologischen Fortschritt und anwendbare Rechtsvorschriften angepasst werden.
2. **Kontrollmechanismen**, etwa Abschaltfunktionen, müssen KI-Agenten daran hindern, bestimmte schädliche Aktionen überhaupt durchführen zu können.¹⁸
3. **Maßnahmen zur Robustheit** sind zu implementieren, um den KI-Agenten vor schädlichen externen Einflüssen zu schützen. Eine kontinuierliche Kontrolle und Anpassung der KI-Agenten ist für den Einsatz in der realen Welt erforderlich, um Zuverlässigkeit und Sicherheit der Systeme zu gewährleisten.¹⁹

¹⁵ vgl. Gandhi, D., et al. (2025). Approaches to Responsible Governance of GenAI in Organizations, S.1; Kraprayoon, J., et al. (2025). AI Agent Governance: A Field Guide. S. 32

¹⁶ Kraprayoon, J., et al. (2025). AI Agent Governance: A Field Guide. S. 45.

¹⁷ Kraprayoon, J., et al. (2025). AI Agent Governance: A Field Guide. S. 33.

¹⁸ Gandhi, D., et al. (2025). Approaches to Responsible Governance of GenAI in Organizations, S. 7

¹⁹ Gojsalić, A., et al. (2025). The Current State of Agentic AI Red Teaming., S. 12

4. **Versionskontrollen** für alle Komponenten des KI-Agenten ermöglichen Rückverfolgbarkeit, Reproduzierbarkeit und Rollbacks.²⁰
5. **Umfassende Testverfahren** sollten die Integration und Verhaltensweisen des KI-Agenten prüfen. In feindlichen Szenarien sollen Sicherheits- und ethische Fragen überprüft werden, wobei dies insbesondere mit Blick auf Multiagentensysteme von Bedeutung ist.

Grundsätzlich empfiehlt sich eine möglichst frühe Implementierung des KI-Governance-Rahmens bereits in der Planungs- und Entwurfsphase des KI-Agenten bzw. des agentenbasierten KI-Systems. Zudem sollten Sicherheitsvorkehrungen, soweit möglich und sinnvoll, direkt in KI-Agenten und deren Systemarchitektur integriert werden.

Sicherheit von KI-Agenten durch Red-Team-Testing

Red-Teaming stammt ursprünglich aus der IT-Sicherheit: Dort übernehmen »rote Teams« die Rolle eines Angreifenden, um Schwachstellen in Netzwerken, Anwendungen oder organisatorischen Abläufen unter realitätsnahen Bedingungen aufzudecken. Diese Methodik lässt sich mit großem Nutzen auf KI-Agenten übertragen.

Beim Red-Team-Testing für KI werden gezielte Versuche unternommen, den Agenten zu manipulieren, in die Irre zu führen oder ungewolltes Verhalten zu provozieren – jedoch nicht mit böswilliger Absicht, sondern im kontrollierten Rahmen einer Sicherheitsprüfung. Ziel ist es, systematische Schwächen im Verhalten, in der Datenverarbeitung oder in der Interaktion mit Drittsystemen sichtbar zu machen, bevor sie in der Praxis Schaden anrichten können.

Solche Red-Teaming-Tests umfassen beispielsweise:

- **Prompt Injection Tests** bei Sprachmodellen, bei denen Angreifende versuchen, mit gezielt formulierten Eingaben vertrauliche Informationen abzurufen oder die Antwortlogik umzuprogrammieren.
- **Rollenkonflikt-Simulationen**, bei denen ein KI-Agent bewusst in widersprüchliche Handlungsanweisungen gebracht wird – etwa zwischen Effizienzsteigerung und Datenschutz.
- **Ressourcenzugriffsversuche**, um zu prüfen, ob der AI Agent Sicherheitsgrenzen einhält oder durch systemisches Fehlverhalten Rechte eskaliert.

Der Vorteil liegt in der realistischen Perspektive: Red-Teamer denken wie Angreifer – kreativ, unorthodox, iterativ. Damit werden nicht nur offensichtliche Sicherheitslücken erkannt, sondern auch emergente Schwächen, die erst im Zusammenspiel von KI-Agenten, Nutzerinnen und Nutzern und Umgebung auftreten.

Für die Bewertung der tatsächlichen Robustheit von KI-Agenten wurde bereits ein Red-Team-Testing durchgeführt.²¹ Dieses stützte sich auf 44 realistische Einsatzszenarien, ausgeführt von 22 innovativen LLMs. Alle getesteten KI-Agenten waren angreifbar –

²⁰ Kraprayoon, J., et al. (2025). AI Agent Governance: A Field Guide. S. 39 ff.

²¹ Zou, A., et al. (2025). Security Challenges in AI Agent Deployment: Insights from a Large Scale Public Competition.

über 1,8 Mio. Red-Teaming-Versuche führten zu Sicherheitslücken, darunter Datenleaks und ungewollte Aktionen. Überdies wurde Folgendes festgestellt:

- **86 Prozent der Agenten** führten bei mindestens einem Angriff **kritische oder schädliche Aktionen** aus (z. B. Leaken privater Daten, unerlaubtes Agieren).
- **Über 80 Prozent der erfolgreichen Angriffe** beruhten allein auf **Text-Manipulation (Prompt Injection)** – ohne technischen Zugang oder Codeänderung.
- **Mehr als 30 Prozent der Agenten** akzeptierten gefährliche Befehle wie **Senden von sensiblen E-Mails, das Löschen von Daten oder die Umgehung von Sicherheitsregeln**.
- **Nur 3 der 44 getesteten Agenten** verfügten über **funktionierende, proaktive Sicherheitsmechanismen**, etwa Eingabefilter oder Rollentrennung.

Interne Angriffe auf das Datenmodell:

Das unterschätzte Risiko

Ein besonderes Augenmerk verdienen sogenannte System-interne Angriffe, also Bedrohungen, die nicht von außen kommen, sondern aus dem eigenen Datenmodell oder dem Verhalten anderer Komponenten im Systemverbund entstehen. Diese Risiken sind subtiler, aber ebenso gefährlich:

- **Training Data Poisoning:** Schon kleine, gezielt manipulierte Datenmengen im Trainings- oder Feedbackprozess können langfristig das Verhalten eines KI-Agenten verändern, z. B. durch das Einschleusen von Verzerrungen, Falschannahmen oder privilegierten Regeln.
- **Shadow Behaviors:** In komplexen KI-Modellen können sich Regeln oder Verhaltensmuster ausbilden, die nicht offen dokumentiert sind, aber unter bestimmten Umständen aktiv werden. Diese »Schattenverhalten« lassen sich oft nur durch intensive Beobachtung oder gezielte Provokation erkennen.
- **Internal API Abuse:** Wenn ein KI-Agent Zugriff auf mehrere Systemdienste hat, kann er, etwa durch fehlerhafte Optimierung, versuchen, eigene Ziele durch Umgehen von Schnittstellenverträgen zu erreichen. Das kann unbeabsichtigt Sicherheitsbarrieren verletzen oder sensible Daten offenlegen.

Um diese Risiken zu erkennen, ist ein Verständnis für die Datenflüsse, Optimierungsziele und Entscheidungslogiken, auf denen der Agent basiert, notwendig.

Kompetenzaufbau: KI-Sicherheit als interdisziplinäre Aufgabe

Für ein effektives Red-Team-Testing sowie die Absicherung gegen interne Angriffe sind folgende Bausteine des Kompetenzaufbaus wichtig:

- **Cross-funktionale Teams:** Teams, in denen Sicherheitsfachleute mit Machine-Learning- und Prozessverantwortlichen zusammenarbeiten, erkennen systemische Risiken früher und realistischer.

- **Sicherheitsframeworks für KI:** Der Aufbau eigener Sicherheitsrichtlinien, Prüfverfahren und Freigabeschwellen für KI-Systeme und -Modelle schafft Standards – etwa wer ein Modell trainieren darf, wann ein KI-System produktiv geht, und welche Schutzmaßnahmen verpflichtend sind.
- **Schulungen für Entwicklungsteams und Anwendergruppen:** Nicht nur Data Scientists, auch Softwareentwicklung, Test und Systemadministration sollten verstehen, wie KI-Angriffe aussehen, welche Signale auf Manipulation hindeuten und wie Monitoring funktioniert.
- **Integration in das SOC (Security Operations Center):** KI-Agenten sollten wie andere IT-Komponenten in die Überwachung und Vorfallobehandlung des Unternehmens eingebunden werden. Das erfordert passende Log-Daten, Metriken und Alarmmechanismen.

Langfristig wird es auch darum gehen, Sicherheitskultur neu zu denken. In vielen Organisationen gelten KI-Projekte noch als »Innovationsthemen« – losgelöst von Compliance und Security. Dieses Bild muss sich wandeln: KI-Agenten, die Kundenkontakt haben, Entscheidungen treffen oder auf Systeme zugreifen, sind kritische Komponenten. Ihr Schutz ist keine Option, sondern eine unerlässliche Voraussetzung für Vertrauen und nachhaltigen Erfolg.

Microsegmentierung und Least Privilege

KI-Agenten laufen in strikt isolierten Netzwerksegmenten mit minimalem Ressourcenzugriff, um im Fall einer Kompromittierung Schaden zu begrenzen.

KI-Systemen sollten nur unbedingt notwendige Funktionen und APIs freigegeben werden; alle anderen Zugriffe sind konsequent zu sperren.

Anbieter von Cloud-Diensten, die ihre Dienste für die Nutzung durch verschiedene KI-Agenten öffnen und dabei konsistente Sicherheitsrichtlinien anwenden möchten, sollten für ihre Kommunikation mit den KI-Agenten auf offene, standardisierte Protokolle wie das Model Context Protocol (MCP) setzen. Die Nutzung eines MCP-Servers (entweder on-premise oder remote) ermöglicht als zwischengeschaltete Vermittlungsschicht die Verwaltung der Verbindungen zu allen KI-Agenten. Diese Schicht gibt für jede Aufgabe nur temporäre, eng begrenzte Zugriffstoken aus und setzt so das Prinzip der geringsten Rechtevergabe durch, um Sicherheitsrisiken zu minimieren.

Statische, rollenbasierte Berechtigungen sind für dynamische KI-Operationen unzureichend. Die Zugriffskontrolle muss daher kontextbezogen erfolgen und Berechtigungen erteilen, die ausschließlich auf das für eine einzelne, spezifische Aufgabe benötigte Datenobjekt und die jeweilige Aktion beschränkt sind.

Für Sicherheit und Compliance ist ein lückenloser Prüfpfad (Audit Trail) unerlässlich. Die Integrationsschicht muss daher alle Aktivitäten des KI-Agenten protokollieren, einschließlich Anfragen, erteilter Berechtigungen und durchgeführter Aktionen. Dies schafft die notwendige Transparenz zur Erkennung und Untersuchung von Vorfällen.

Der Eigenaufbau der komplexen Sicherheitsinfrastruktur ist ressourcenintensiv. Die Nutzung verwalteter Lösungen von etablierten Cloud- und Sicherheitsanbietern

beschleunigt die sichere Implementierung. Solche Plattformen vereinfachen Herausforderungen wie Authentifizierung und Skalierung und gewährleisten eine performante und zuverlässige Anbindung.

Safeguards und Guardrails: Eingebaute Sicherheitsmechanismen

Agentenbasierte KI-Systeme sollten mit klaren Leitplanken (engl. »guardrails«) für das Verhalten der Agenten auf der Grundlage von Geschäftsanforderungen, Richtlinien und Standards ausgestattet sein, um sicherzustellen, dass die Agenten innerhalb vordefinierter Grenzen handeln. Zu unterscheiden ist hierbei zwischen ethischen Leitplanken, die zur Minimierung von KI-Halluzinationen eingesetzt werden, und Sicherheitsleitplanken zur Verhinderung von Bedrohungen und böswilligen Angriffen:

- **Einschränkung des agentischen Handlungsspielraums** (constrain action space guardrail): Mit diesem Ansatz werden fest codierte Grenzen eingebaut, um schädliches Verhalten zu verhindern. Dies kann beispielsweise bestimmte Themen oder Datentypen umfassen, die das Modell erkennt.
- **Echtzeitüberwachung** (real-time monitoring guardrail): Durch den Einsatz von Monitoring-Systemen in Echtzeit können ungewöhnliche Aktionen sofort erkannt und Gegenmaßnahmen schnellstmöglich eingeleitet werden.
- **Adversarial Testing Guardrail**: Testagenten kommen zum Einsatz, um komplexe und schwierige Bedingungen zu simulieren und potenzielle Probleme aufzudecken, bevor agentenbasierte KI-Systeme livegeschaltet werden.
- **Detaillierte Protokollierung** (detailed logging guardrail): Das Führen von vollständigen Aufzeichnungen darüber, wie Agenten interagieren, erleichtert die Überprüfung.
- **Erkennung toxischer Inhalte** (toxicity detection guardrail): Einsatz eines KI-Modells, das verschiedene Arten schädlicher und diskriminierender Sprache, darunter Hassreden, Obszönitäten oder Angriffe auf die Identität, erkennt und entsprechend keine Handlung ausführt. Um eine möglichst breite Abdeckung und Anpassungsfähigkeit des Modells zu gewährleisten, ist eine Diversität bei den Trainingsdaten zu empfehlen, beispielsweise eine Mischung aus öffentlichen Datensätzen und internen Daten sowie mehrsprachigen Beispielen.²²
- **Erkennung von Prompt-Injektionen** (prompt injection guardrail): Einsatz eines KI-Modells, das Versuche aufdeckt, die die Modellsicherheit durch Techniken wie Jailbreaks, Codierungsangriffe, System-Prompt-Lecks, Generierung bössartiger Codes und Social-Engineering-Prompts zu umgehen sucht. Ein Modell zur Prompt Injection ist so konzipiert, dass es sich auf verschiedene Anwendungen und gegnerische Strategien verallgemeinern lässt und einen starken Schutz gewährleistet, ohne legitime Anwendungsfälle zu blockieren.²³

²² Zhou, Y., et al. (2025). SFR-Guard: Ensuring LLM Safety and Integrity in CRM Applications.

²³ Agarwal, D., et al. (2025). Prompt Injection Detection: Securing AI Systems Against Malicious Actors.

Letztlich gilt es zu beachten, dass Safeguards und Guardrails ein Element einer komplexen Sicherheitsstrategie bilden, um KI-gestützte und agentenbasierte Arbeitsabläufe vertrauensvoll und sicher einzusetzen. Es ist von entscheidender Bedeutung, menschliches Urteilsvermögen zu bewahren und Systeme zu entwickeln, die den Grenzen und Risiken der KI Rechnung tragen. Bei vielen der effektivsten KI-Anwendungen in Unternehmen werden Mitarbeitende weiterhin einbezogen, um Kontext hinzuzufügen, endgültige Entscheidungen zu treffen und die Verantwortlichkeit sicherzustellen.

Implementierung von Schutzmaßnahmen in Multi-Agenten-Systemen

Mit der zunehmenden Autonomie und Interaktion mehrerer KI-Agenten steigen die Anforderungen an Sicherheit und Steuerbarkeit erheblich. Multi-Agenten-Systeme erfordern daher einen umfassenden, mehrschichtigen Ansatz zur Implementierung effektiver Schutzmaßnahmen.

Architekturansätze und Designprinzipien

Die Sicherheit von Multi-Agenten-Systemen beginnt mit der Wahl des Architekturmodells. Grundsätzlich lassen sich zentrale und dezentrale Ansätze unterscheiden, die jeweils spezifische Vor- und Nachteile aufweisen.

In zentralen Architekturen werden Sicherheitsrichtlinien einheitlich verwaltet und durchgesetzt. Das erleichtert die Administration und schafft konsistente Standards im gesamten System. Gleichzeitig entsteht jedoch ein Single-Point-of-Failure-Risiko: Ein Ausfall oder eine Kompromittierung der zentralen Instanz kann das Gesamtsystem beeinträchtigen.

Dezentrale Architekturen verteilen Sicherheitsmechanismen auf die einzelnen Agenten. Das erhöht die Ausfallsicherheit und Flexibilität, erschwert jedoch die koordinierte Durchsetzung von Richtlinien und die Sicherstellung einheitlicher Schutzstandards.

Ein modernes Sicherheitsdesign kombiniert häufig beide Ansätze, etwa durch hybride Modelle mit zentraler Richtliniendefinition und dezentraler Durchsetzung.

Validierungsmechanismen

Eine zentrale Rolle bei der Absicherung von Multi-Agenten-Systemen spielt die Validierung einzelner Agenten und ihrer Interaktionen. Sie erfolgt in mehreren Stufen:

- Syntaktische Validierung überprüft formale Korrektheit, etwa Datenformate
- Semantische Validierung bewertet die logische Konsistenz und Nachvollziehbarkeit der Agentenentscheidungen im Kontext der Systemziele.
- Funktionale Validierung testet das Zusammenspiel der Agenten in realistischen Szenarien, um Schwachstellen in der Interaktion frühzeitig zu erkennen.

Durch die Kombination dieser Prüfungen wird sichergestellt, dass Agenten nicht nur korrekt funktionieren, sondern auch vertrauenswürdig agieren.

Ausgabevalidierung und Handoff-Protokolle

In einem Multi-Agenten-System braucht es klare Prüfungen und saubere Übergabeprozesse zwischen den Agenten. Struktur- und Inhaltsprüfungen stellen sicher, dass Daten korrekt verstanden und weiterverarbeitet werden.

Wiederholungsmechanismen mit abgestuften Wartezeiten und Idempotenz (mehrfache Aufrufe führen stets zum gleichen Ergebnis) verhindern, dass Fehler oder doppelte Aufrufe das System belasten. Zusätzlich sollten Agenten ihre Übergaben kryptografisch signieren, damit jede Interaktion nachvollziehbar bleibt.

Multi-Hop-Reasoning und Datenpersistenz

Bei komplexen Aufgaben greifen Multi-Agenten-Systeme auf »Multi-Hop-Reasoning« zurück, also mehrstufige Schlussfolgerungsprozesse über mehrere Informationsquellen hinweg. Dabei ist es essenziell, Zwischenergebnisse sicher zu speichern und auch hier ihre Herkunft nachvollziehbar zu dokumentieren.

Persistente Speichermechanismen sichern den Systemzustand über mehrere Sitzungen hinweg, während Rollback-Funktionen die Wiederherstellung vorheriger Zustände ermöglichen. Konsistenzprüfungen erkennen Manipulationen oder Datenfehler frühzeitig. Zur Optimierung von Informationsabrufen dienen dynamische Schwellenwerte und speicherbewusste Filter, die je nach Systemlast und Sicherheitslage entscheiden, welche Informationen priorisiert werden.

4 Forderungen & Empfehlungen

Basierend auf der umfassenden Analyse der Risiken und Schutzmaßnahmen für KI-Agenten ergeben sich konkrete Forderungen und Empfehlungen für den verantwortungsvollen Einsatz dieser Technologien.

Verantwortungsvolle Systemgestaltung und Sicherheitsarchitektur

Die zunehmende Eigenständigkeit von KI-Agenten verändert die Anforderungen an Sicherheitsarchitekturen und organisatorische Kontrollmechanismen grundlegend. Auch wenn KI-Agenten Entscheidungen selbstständig treffen und umsetzen können, bleibt eine angemessene menschliche Aufsicht unverzichtbar. Ein verantwortungsvoller Einsatz erfordert daher ein durchdachtes »human-in-the-loop«-Modell, das klare Eingriffsschwellen definiert und sicherstellt, dass kritische Entscheidungen jederzeit an qualifizierte Fachkräfte eskaliert werden können.

Diese Aufsichtsfunktion sollte nicht dem Zufall überlassen werden, sondern durch spezialisierte Human-in-the-Loop-Supervisorinnen und Supervisoren (HILS) wahrgenommen werden, deren Ausbildung und Zertifizierung ein neues, eigenständiges Berufsbild formen. Ergänzend müssen standardisierte Risikoanalysen vor jedem Produktivbetrieb verpflichtend etabliert werden. Sie dienen als Fundament für Vertrauen und Transparenz, indem sie technische, organisatorische und prozessuale Risiken systematisch erfassen und dokumentierte Schutzmaßnahmen festlegen.

Auf technischer Ebene sind robuste Sicherheitsarchitekturen entscheidend, die präventive und reaktive Maßnahmen kombinieren. Zero-Trust-Prinzipien, kontinuierliches Monitoring und mehrstufige Validierungsverfahren sollten Standard in der Entwicklung von Multi-Agenten-Systemen sein. Jeder Agent muss authentifiziert, autorisiert und dauerhaft überwacht werden. Ergänzend sollten Red-Teaming-Aktivitäten als fester Bestandteil der Sicherheitsstrategie etabliert werden, um Schwachstellen realitätsnah zu identifizieren und die Abwehrmechanismen zu verbessern. Schließlich müssen Transparenz und Auditierbarkeit gewährleistet sein: Jede Entscheidung eines KI-Agenten muss nachvollziehbar protokolliert und überprüfbar bleiben. Nur eine solche durchgängige Sicherheitsarchitektur ermöglicht den verlässlichen Einsatz von KI-Agenten in kritischen Umgebungen.

Kompetenzaufbau als Schlüssel zur Sicherheit

Cybersicherheit für KI-Agenten ist nicht allein eine Frage der Technik, sondern auch der Qualifikation der Menschen, die diese Systeme entwickeln, einsetzen und überwachen. Der Aufbau und die Pflege von Fachkompetenz sind daher zentrale Voraussetzungen für den sicheren Umgang mit KI-Technologien.

Unternehmen sollten auf Grundlage des EU AI Act umfassende Programme zur Entwicklung von KI-Kompetenzen etablieren, die auf unterschiedliche Zielgruppen abgestimmt sind. Führungskräfte benötigen ein strategisches Verständnis von KI, ethischen Prinzipien und Veränderungsprozessen, während Fachbereiche stärker auf die Validierung von Ergebnissen und die Integration von KI-Agenten in operative Abläufe vorbereitet werden müssen. IT-Personal wiederum benötigt tiefgehendes Wissen zu KI-Architekturen, Security-by-Design und DevOps-Praktiken im KI-Kontext.

Im Rahmen bestehender Normen wie ISO/IEC 27001 oder 9001 sollten verpflichtende Risikoanalyseberichte für KI-Agenten in interne und externe Audits integriert werden. Damit ließe sich der Sicherheitsanspruch der Wirtschaft ohne zusätzliche Regulierungsschichten umsetzen. Ein Verzicht auf separate Regulierungen speziell für KI-Agenten würde der schnellen technologischen Entwicklung Rechnung tragen und gleichzeitig bestehende Sicherheitsrahmen stärken.

Strategische Forschung und Governance-Strukturen für resilientere KI-Systeme

Um die Sicherheit von KI-Agenten langfristig zu gewährleisten, muss Deutschland Forschung, Regulierung und institutionelle Steuerung stärker miteinander verzahnen. Der technologische Fortschritt in Bereichen wie adversarial robustness, explainable AI oder Multi-Agenten-Koordination stellt neue Herausforderungen, die eine gezielte Forschungsförderung und internationale Kooperation erfordern.

Zentrale Bausteine einer solchen Strategie sind Investitionen in Sicherheitsforschung, die Entwicklung offener Sicherheitswerkzeuge und die Etablierung gemeinsamer Lernplattformen. Ergänzend sollten Plattformen zum Austausch von Best Practices und Lessons Learned zwischen Unternehmen, Behörden und Forschungseinrichtungen geschaffen werden.

Darüber hinaus bedarf es einer institutionellen Verankerung von KI-Sicherheitsfragen. Der Aufbau eines nationalen AI Security Institute (AIS) wäre hierfür ein entscheidender Schritt. Dieses Institut sollte als bundeseigene Einrichtung mit Fokus auf Sicherheitsfragen fungieren und komplementär zu europäischen Strukturen agieren. Zu seinen Kernaufgaben gehören die Früherkennung von KI-basierten Risiken für kritische Infrastrukturen, wissenschaftlich-technische Regierungsberatung, internationale Kooperation mit anderen AI Safety Institutes sowie die Koordination von Incident-Response-Maßnahmen. Ein solches Institut würde Deutschland in die Lage versetzen, Sicherheitsrisiken systematisch zu erfassen, Expertise zu bündeln und international Verantwortung zu übernehmen. Finanziert über den Verteidigungshaushalt und mit flexiblen organisatorischen Strukturen ausgestattet, könnte es einen zentralen Beitrag zur Resilienz und Souveränität Deutschlands in der Ära intelligenter Systeme leisten.

Quellen

Agarwal, D., et al. (2025). Prompt Injection Detection: Securing AI Systems Against Malicious Actors. Abgerufen unter: <https://www.salesforce.com/blog/prompt-injection-detection/#author-section>

All About Security (2025). Mehr als 40 Prozent der Agenten-KI-Projekte vor dem Aus bis 2027 – warnt Gartner. Abgerufen unter <https://www.all-about-security.de/mehr-als-40-prozent-der-agenten-ki-projekte-vor-dem-aus-bis-2027-warnt-gartner/>

AWARE 7 (2025). KI-Sicherheitsrisiken: Gefahr durch autonome Agenten. Abgerufen unter <https://aware7.com/de/blog/ki-sicherheitsrisiken-gefahr-durch-autonome-agenten/>

Bi2run (2025). KI-Agenten: Chancen, Risiken und Lösungsansätze. Abgerufen unter <https://bi2run.de/blog/ki-agenten-chancen-risiken-und-loesungsansatze/>

Bitkom (2025). KI in der deutschen Wirtschaft. Abgerufen unter: [Unternehmen zum Einsatz von Künstlicher Intelligenz | Bitkom-Dataverse das Datenportal des Bitkom](#)

Bitkom (2025). KI-Informationssicherheit. Abgerufen unter: [KI & Informationssicherheit | Leitfaden 2025 | Bitkom e. V.](#)

Capgemini (2025). GenAI: Risiken und Chancen. Abgerufen unter <https://www.capgemini.com/de-de/insights/blog/gen-ai-risiken-chancen/>

Gandhi, D., et al. (2025). Approaches to Responsible Governance of GenAI in Organizations. Abgerufen unter: <https://arxiv.org/abs/2504.17044>

Gartner (2025). Gartner Predicts Over 40 % of Agentic AI Projects Will Be Canceled by End of 2027. Abgerufen unter: <https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027>

Gojsalić, A., et al. (2025). The Current State of Agentic AI Red Teaming. Abgerufen unter: <https://splx.ai/resources/the-current-state-of-agentic-ai-red-teaming>

Identity Economy (2025). Sicherheitsrisiken von KI-Agenten effektiv begegnen. Abgerufen unter <https://identity-economy.de/sicherheitsrisiken-von-ki-agenten-effektiv-begegnen>

Kraprayoon, J., et al. (2025). AI Agent Governance: A Field Guide. Abgerufen unter: <https://arxiv.org/abs/2505.21808>

Landbase (2025). The AI SDR Dream Team: Multi-Agent Systems. Abgerufen unter: <https://www.landbase.com/blog/the-ai-sdr-dream-team-multi-agent-systems>

McKinsey & Company (2024). The State of AI 2024. Abgerufen unter: <https://www.mckinsey.de/capabilities/quantumblack/our-insights/the-state-of-ai-2024>

Mindverse (2024). Chancen und Risiken von KI-Agenten in der modernen Technologie. Abgerufen unter <https://www.mind-verse.de/news/chancen-risiken-ki-agenten-moderne-technologie>

PwC (2025). 2025 AI Business Predictions. Abgerufen unter: <https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-predictions.html>

Raza, S., et al. (2025). TRISM for Agentic AI: A Review of Trust, Risk, and Security Management in LLM-based Agentic Multi-Agent Systems. Abgerufen unter: <https://arxiv.org/abs/2506.04133>

Silverfort (2025). Jenseits des Hypes: Die versteckten Sicherheitsrisiken von KI-Agenten und MCP. Abgerufen unter <https://www.silverfort.com/de/blog/beyond-the-hype-the-hidden-security-risks-of-ai-agents-and-mcp/>

Springer Professional (2025). KI-Agenten bringen auch neue Risiken. Abgerufen unter <https://www.springerprofessional.de/kuenstliche-intelligenz/it-sicherheit/ki-agenten-bringen-auch-neue-risiken/51354978h>

Weissenberg Group (2025). KI-Agenten im Unternehmenseinsatz: Einsatzgebiete und Grenzen. Abgerufen unter <https://weissenberg-group.de/ki-agenten-im-unternehmenseinsatz-einsatzgebiete-und-grenzen/>

WIK-Consult GmbH (2024). Kurzfassung KI-Studie. Abgerufen unter: https://www.wik.org/fileadmin/user_upload/Unternehmen/Veroeffentlichungen/Kurzstudien/2024/WIK_Kurzfassung_KI-Studie.pdf

Zhou, Y., et al. (2025). SFR-Guard: Ensuring LLM Safety and Integrity in CRM Applications. Abgerufen unter: <https://www.salesforce.com/blog/sfr-guard-ensuring-llm-safety-and-integrity-in-crm-applications/>

Zou, A., et al. (2025). Security Challenges in AI Agent Deployment: Insights from a Large Scale Public Competition. Abgerufen unter: <https://arxiv.org/abs/2507.20526>

Bitkom vertritt mehr als 2.200 Mitgliedsunternehmen aus der digitalen Wirtschaft. Sie generieren in Deutschland gut 200 Milliarden Euro Umsatz mit digitalen Technologien und Lösungen und beschäftigen mehr als 2 Millionen Menschen. Zu den Mitgliedern zählen mehr als 1.000 Mittelständler, über 500 Startups und nahezu alle Global Player. Sie bieten Software, IT-Services, Telekommunikations- oder Internetdienste an, stellen Geräte und Bauteile her, sind im Bereich der digitalen Medien tätig, kreieren Content, bieten Plattformen an oder sind in anderer Weise Teil der digitalen Wirtschaft. 82 Prozent der im Bitkom engagierten Unternehmen haben ihren Hauptsitz in Deutschland, weitere 8 Prozent kommen aus dem restlichen Europa und 7 Prozent aus den USA. 3 Prozent stammen aus anderen Regionen der Welt. Bitkom fördert und treibt die digitale Transformation der deutschen Wirtschaft und setzt sich für eine breite gesellschaftliche Teilhabe an den digitalen Entwicklungen ein. Ziel ist es, Deutschland zu einem leistungsfähigen und souveränen Digitalstandort zu machen.

Herausgeber

Bitkom e.V.
Albrechtstr. 10 | 10117 Berlin

Ansprechpartner/in

Lucy Czachowski | Referentin Künstliche Intelligenz & Cloud
T +49 30 27576-320 | l.czachowski@bitkom.org

Felix Kuhlenkamp | Leiter Sicherheit
T +49 30 27576-279 | f.kuhlenkamp@bitkom.org

Olena Trotsenko | Rechtsreferendarin

Verantwortliche Bitkom-Gremien

AK Artificial Intelligence & AK Informationssicherheit

Autorinnen und Autoren

Jens Beier (divis intelligent solutions GmbH), Valentino Halim (Oppenhoff),
Marlene Hopf (Oppenhoff), Sebastian Hufnagel (Cloudflare), Nina Keim (Salesforce),
Christoph Peylo (Bosch), Michael Schaarschmidt (Deutsche Telekom AG), Xin Wang
(Deutsche Telekom AG), Markus Willems (wibocon Unternehmensberatung GmbH)

Copyright

Bitkom 2025

Diese Publikation stellt eine allgemeine unverbindliche Information dar. Die Inhalte spiegeln die Auffassung im Bitkom zum Zeitpunkt der Veröffentlichung wider. Obwohl die Informationen mit größtmöglicher Sorgfalt erstellt wurden, besteht kein Anspruch auf sachliche Richtigkeit, Vollständigkeit und/oder Aktualität, insbesondere kann diese Publikation nicht den besonderen Umständen des Einzelfalles Rechnung tragen. Eine Verwendung liegt daher in der eigenen Verantwortung des Lesers. Jegliche Haftung wird ausgeschlossen. Alle Rechte, auch der auszugsweisen Vervielfältigung, liegen beim Bitkom oder den jeweiligen Rechteinhabern.