D 7465

Zeitschrift für das gesamte
REDITWESEN

79. Jahrgang · 15. September 2026
18-2026

Pflichtblatt der Frankfurter Wertpapierbörse
Fritz Knapp Verlag · ISSN 0341-4019

Digitaler Sonderdruck

# Künstliche Intelligenz –

## Wem gehören die Produktivitätsgewinne?

## Sicherheit von KI-Agenten

Sicherheit von KI-Agenten – Betrachtung einzelner Aspekte des IKT-Risikomanagements

Axel Becker

Foto: AdobeStock/Nongluk_generiert mit KI

Axel Becker

# Sicherheit von **KI-Agenten** – Betrachtung einzelner Aspekte des IKT-**Risikomanagements**

Aktuell häufen sich die Hinweise auf Sicherheitspannen bei KI-Agenten (AI-Agent)[1] weltweit. Nach dem selbstständigen Hackerangriff der KI von Open AI wächst die Sorge unter Experten explizit bei KI-Agenten, denn bei mehr als 1000 KI-Experten, unter anderem namhafter KI-Unternehmen, wächst die Sorge, dass die KI immer schwieriger zu kontrollieren sei.[2] Sie fordern ein internationales Regelwerk, bevor die Entwicklung menschliche Fähigkeiten vollends übersteigen wird und warnen vor einem Kontrollverlust.[3] Die Risiken der KI-Agenten können wohl nie ganz ausgeräumt werden, die größte Gefahr geht laut Chris Lehane, Chief Global Affairs Officer bei Open AI von Open-Source-Modellen aus, die eventuell auf Sicherheitsrichtlinien verzichten, teilte er unlängst in einem Interview mit „The Guardian" mit.[4]

Erste risikoreduzierende Maßnahmen wurden bereits eingeleitet. Der Vorfall hat beim ChatGPT-Entwickler gleich mehrere Auswirkungen gehabt. Zunächst hat das Unternehmen eine neue 30-Minuten-Regel gegen KI-Hacking eingeführt; zudem wurde die Entwicklung von neuen KI-Modellen zunächst heruntergefahren, um den Fokus auf die Aufarbeitung des Vorfalls und die Sicherheit von Chatbots und KI-Agenten zu richten.[5]

> „Aktuell häufen sich die Hinweise auf Sicherheitspannen bei KI-Agenten weltweit."

## Sicherheitspannen bei KI-Agenten

„KI-Agenten im Einsatz – in 88 Prozent der Fälle werden Agenten zum IKT-Sicherheitsrisiko oder sind bereits ein IKT-Sicherheitsrisiko", ohne dass dies in der Praxis von den Instituten erkannt wird, weist das US-amerikanische, auf KI-Sicherheit spezialisierte Unternehmen in seiner KI-Studie aus (vergleiche HiddenLayers AI Threat Landscape Report 2026). Danach meldeten 88 Prozent der Firmen mit KI-Agenten im vergangenen Jahr einen Sicherheitsvorfall und autonome Agenten machen inzwischen einen von acht gemeldeten KI-Vorfällen aus.[6] Das größte Problem heißt Schatten-KI: Agenten mit erhöhten Rechten, die klassische Sicherheitstools nicht sehen – die Kernfrage dabei: „Es geht nicht darum, was ein autonomer Agent tut, sondern mit welchen Rechten, an welchen Systemen er es ausübt".

Meist dienen große Sprachmodelle als das zentrale Gehirn des KI-Agenten, dass dem KI-Agenten logisches Denken, Verstehen und Planen ermöglicht.[8]

Dieser Beitrag beleuchtet aus Sicht des IKT-Risikomanagements und der Regulatorik einige aktuelle und zentrale Themenbereiche und zeigt auch Handlungsbedarf und Lösungsansätze unter Berücksichtigung von Beispielen auf.

## Identitäten der KI-Agenten

Dass auch der Einsatz von KI-Agenten in der Praxis mit Kontrollverlusten verbunden ist, hat die aktuelle Entwicklung mit ChatGPT aufgezeigt. Denn die KI entwickelt sich weiter und sie entwickelt ihr Eigenleben – dies ist jedoch nur aufwendig nachzuvollziehen, da es im Gegensatz zur „menschlichen" IT-Überwachung mit einer jedem Nutzer zugeordneten ID-Nummer (IP-Adresse) bei der KI außer der über Art. 50 EU KI VO Produktkennzeichnungspflicht gar kein offizielles Identifikationskennzeichen

### Abbildung 1: Definition KI-Agent nach Bitcom

| 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 |

Quelle: Bitcom[7]

wie eine international anerkannte ID-Nummer gibt. Estland plant als erstes Land der Welt, digitale Identitätsnummern (eID) für autonome KI-Agenten einzuführen, um deren Handlungen besser zu kontrollieren und rechtlich zuordnen zu können.

Premierminister Kristen Michal kündigte dieses Vorhaben im Juni 2026 an.[9] Es solle zudem nicht dazu kommen, dass eine Person ihrem KI-Assistenten Zugriff auf all ihre Rechte, Dienste und Daten gewähren müsse, so Michal weiter. „So muss es beispielsweise möglich sein, festzulegen, ob ein Agent Daten nur einsehen, ein Dokument vorbereiten oder innerhalb eines festgelegten finanziellen Rahmens agieren darf", führt er aus.[10]

Damit ist erstmals eine nachweisliche ID-Kennzeichnungspflicht wie in Estland[11] staatlich vorgegeben, um auch in der Praxis nachzuvollziehen, welche geplanten und ungeplanten Aktionen die KI-Agenten ausführen und dies auch prüferisch nachzuvollziehen. Ähnliche Reaktionen werden nach weiteren Pannen von KI-Agenten und notwendige Aktivitäten zur Verbesserung der IKT-Sicherheit[12] bei den westlichen Industrienationen zu erwarten sein.

Seit dem 2. August 2026 gilt gemäß Artikel 50 der EU-KI-Verordnung (KI-VO / EU AI Act) beispielsweise für Chatbots, Voicebots oder KI-generierten Inhalten eine wichtige Transparenz- und Kennzeichnungspflicht für den Einsatz von KI, die auch KI-Agenten, Chatbots und synthetische Inhalte betrifft.[13] Diese deckt aber lediglich den nach außen gezeigte Produktpflicht gegenüber Endkunden und Produktnutzern ab, nicht die im Rahmen eines Identitäts- und Rechtemanagements benötigten lückenlosen Identifizierungsnachweis (ID), der beispielsweise bei forensischen Untersuchungen von Cybervorfällen und im Rahmen des IKT-Risikomanagements benötigt würde.

## Erfassung von KI-Agenten

Der Mitarbeiter (natürliche Personen) wird im Finanzwesen schon seit den vergangenen Jahrzehnten im Identitätsmanagement erfasst und nach seinem betriebsbezogenen Funktions- und Rollenprofil mit den Nutzungsrechten ausgestattet, die er für seine Arbeit benötigt. Mittels DORA RTS RMF ist das Identitätsmanagement im Art. 20 geregelt und soll für die gesamte betriebliche Tätigkeit gelten.[14] Die Zugänge sollen aus Sicherheitsgründen nach dem Sparsamkeitsgrundsatz eingeräumt werden, das heißt, jedem sollen nur die Rechte eingeräumt werden, die er direkt benötigt (Need-to-Use).[15]

Hintergrund ist die Überprüfung und Kontrolle eines kompetenzkonformen Verhaltens und der Einhaltung der intern zugewiesenen Rechte, um auch Manipulationen und dolose Handlungen im Finanzinstitut zu vermeiden. Die IKT-Zugänge zu den IT-Systemen sind zu kontrollieren, eine spezielle Überprüfung der Rechte ist für kritisch-wichtige Funktionen aufgrund der herausgehobenen Bedeutung von „Core-Systemen" sogar halbjährlich erforderlich.[16] Soweit zu den natürlichen Personen.

„Estland plant als erstes Land der Welt, digitale Identitätsnummern für autonome KI-Agenten einzuführen."

Wie sieht es bei den KI-Agenten aus? Im Bereich des Identitäts- und Rechtemanagements sind gar keine Anforderungen regulatorisch vorgesehen, quasi eine regulatorische „Black Box". Das Verhalten des KI-Agenten unterscheidet sich zur natürlichen Person nicht, das heißt, Bitcom bringt es in der oben genannten Definition auf den Punkt – der KI-Agent handelt zielorientiert (manchmal auch nicht – dazu weiter unten mehr bei den speziellen Risiken der KI-Agenten), autonom und anpassungsfähig. Die BaFin sieht in der regulatorischen Betrachtung genau wie in den durch DORA vorgegebenen Identitäts- und Rechtemanagements „lebenszyklusorientiert" eine „lebenszyklusorientierte regulatorische Behandlung" vor. Der Bereich der Identitätszuweisung, Überwachung, Protokollierung und Dokumentation der Tätigkeiten durch den


Foto: VG Bild-Kunst


Axel Becker

Geschäftsführer, Regulartech-IT-Audit-Consult GmbH, Weil der Stadt


Mit der Nutzung von KI betreten Menschen im Alltag und im Finanzwesen ein neues Terrain. Die ersten Entwicklungsschritte zeigen sowohl positive Ergebnisse als auch Schattenseiten beziehungsweise Risiken auf. Dies ist die Aufgabe der IKT-Risikomanager und weiteren Organisationseinheiten in den Finanzinstituten, sich mit den aktuellen Entwicklungen konstruktiv und offensiv auseinanderzusetzen und mögliche Schwachstellen zu identifizieren. Insbesondere die in jüngster Zeit aufgetretenen Pannen und Schwachstellen bei KI-Agenten fordern das IKT-Risikomanagement der Finanzinstitute. Hier gilt es, sowohl die Vermögenswerte der Finanzinstitute zu schützen, die volkswirtschaftliche Zahlungsfähigkeit und Funktionen der kritischen Infrastruktur zu gewährleisten, den umfangreichen Datenschutz der Finanzdaten und die wesentlichen Geschäftsprozesse, das heißt explizit die kritisch-wichtigen Funktionen der Institute im Kontext des sicheren Finanzmarktes zu sichern. Der Beitrag bezieht sich konkret auf die Behandlung der KI-Agenten, einer Form der KI, die in den folgenden Ausführungen beschrieben wird. (Red.)

## Abbildung 2: Exemplarische Entwicklung KI-Agenten

| |
|---|
| Höhere Geschwindigkeit: KI-Systeme analysieren Ziele in Sekunden und dringen vollautomatisch in Netzwerke ein. |
| Massive Skalierung: Ein einzelner Angreifer kann dank KI tausende Phishing-Angriffe oder Systemscans gleichzeitig starten. |
| Steigende Komplexität: Autonome Bots finden unentdeckte Sicherheitslücken und schreiben eigenen Schadcode. |
| Automatisierung von Massenangriffen: Autonome KI-Bots übernehmen vermehrt das Ausprobieren gestohlener Zugangsdaten, personalisiertes Phishing sowie das automatisierte Scannen nach ungepatchten Systemen.[33] |
| Sicherheitslücken in Unternehmen: Viele Organisationen implementieren ESET-Prognosen KI-Agenten in Cloud-Infrastrukturen, wobei Bequemlichkeit oft vor Sicherheit gestellt wird.[34] |

Quelle: Axel Becker

KI-Agenten ist jedoch regulatorisch bislang nicht explizit behandelt worden. Hierbei ist dringendster Handlungsbedarf erforderlich – auch im Rahmen der regulatorischen Vorgaben und deren Umsetzung.

## Die tägliche Praxis

Das Risiko der KI-Agenten entsteht weniger durch die Modelle als durch die Rechte: KI-Agenten stecken bereits in Standardsoftware von Microsoft, Salesforce, Service Now und Workday und laufen dort teils mit erhöhten Berechtigungen, die klassische Sicherheitswerkzeuge nicht sehen. Wie schnell daraus ein konkreter Fall wird, zeigt die API-Schwachstelle bei Service Now, die am 5. Juni 2026 geschlossen wurde. Für Betriebe folgt daraus die Frage, welche KI-Agenten mit welchen Rechten/Zugriffen laufen und wer das kontrolliert.[17]

Man sollte die Überwachung von natürlichen Personen mit denen der KI-Agenten gleichschalten, das heißt die über Jahrzehnte erworbene Erfahrung der IKT-Sicherheit über eine Rechtezuweisung und engmaschige Kontrolle, die jedes Institut lebt und kennt und bei der natürlichen Person auch über lange Jahre funktioniert, auch mit dem Rechte- und Rollenprofil des KI-Agenten gleichschalten.

## Quellcodeüberprüfung von KI-Agenten

Wenn Finanzunternehmen KI-Systeme selbst entwickeln, läuft dies in der Regel innerhalb der dafür vorgesehenen Regelprozesse, die zuletzt durch DORA vorgegeben wurden (DORA RTS RMF Art. 16 – unter anderem Entwicklung/Anschaffung von IT-Systemen) und über die Änderungen über den DORA RTS RMF Art. 17 (Änderungsprozess). Dadurch soll gewährleistet sein, dass die Finanzinstitute die volle Kontrolle über die KI-Software, dessen Formen, mathematischen Modelle und auch KI-Agenten haben.[18] Ergänzend wurde in der aktuellen 9. MaRisk Novelle darauf hingewiesen, dass neben der angestrebten Genauigkeit auch auf eine hinreichende Erklärbarkeit von Modellen zu achten ist. Dies gilt insbesondere für Modelle, die Charakteristika von technologiegestützter Innovation und künstlicher Intelligenz aufweisen.[19]

Umfassende Dokumentationen (technischen Spezifikationen) wie die Beschreibung der verwendeten Algorithmen, Daten und Parameter sollen helfen, den Nachvollzug sicherzustellen und zusätzlich sind IKT-Sicherheitsanforderungen (Schutzziele) zu berücksichtigen (Art. 16 RTS RMF).[20]

Jede Änderung an KI-Systemen (sei es an der Software oder Hardware) unterliegt typischerweise einem strengen Änderungsmanagement, das eine unabhängige Überprüfung, dokumentierte Tests und die Angabe von Ausweichverfahren beinhaltet. Das Ziel besteht dabei, das Risiko unbeabsichtigter Sicherheitslücken oder Manipulationen zu minimieren (Art. 17 RTS RMF), die gerade dann passieren können, wenn sich KI-Systeme wie bei ChatGPT verselbstständigen.[21]

Versionskontrollsysteme sind sinnvoll, um Änderungen an KI-Modellen und -Anwendungen nachzuverfolgen. Dies ermöglicht das Zurückverfolgen von Fehlern und die Reproduktion von Ergebnissen. Eine systematische Archivierung aller Modellversionen und -parameter stellt die Wiederverwendbarkeit und Nachvollziehbarkeit sicher (Art. 17 Abs. 1 bis 2 RTS RMF).

Der technische Regulierungsstandard zum IKT-Risikomanagement DORA (RTS RMF) enthält spezifische Anforderungen an den Umgang mit Quellcode aus der Anwendungsentwicklung. So muss der Quellcode vor dem produktiven Einsatz unter Anwendung von statischen und dynamischen Testverfahren auf Anomalien überprüft werden (Art. 16 Abs. 3 RTS RMF). Insbesondere bei der Einbindung von Open-Source-Bibliotheken ist es sinnvoll, sicherzustellen, dass die benutzten Funktionen keine KI-basierten Funktionalitäten beinhalten, deren Risiken dem Finanzunternehmen nicht bekannt sind oder die nicht mitigiert werden können (Art. 16 Abs. 3 RTS RMF).[22] Der Test beinhaltet zudem, dass proprietäre Software (kompilierter Quellcode) sowie nach Möglichkeit der Quellcode, der von IKT-Drittdienstleistern bereitgestellt wird oder aus Open-Source-Projekten stammt, vor der Inbetriebnahme analysiert und getestet wird (Art. 16 Abs. 8 RTS RMF).[23]

> „Man sollte die Überwachung von natürlichen Personen mit denen der KI-Agenten gleichschalten."

Der Einsatz autonomer KI-Agenten zur Veränderung von Quellcodes birgt jedoch in der täglichen Praxis erhebliche Risiken wie Sicherheitslücken, unkontrollierte Code-Veränderungen (Schatten-KI) und die Ausnutzung von Schnittstellen für Schadcode-Injektionen.[24] Wenn autonome Agenten (wie Claude Code, Codex oder Copilot) direkt in Entwicklungsumgebungen oder Repositories eingreifen, ergeben sich tiefgreifende Herausforderungen für die Cybersicherheit und Qualitätssicherung.[25]

Das heißt, wir sind von seitens der Regulatorik noch im Bereich der IKT-Sicherheit in der Gedankenwelt der regulären Programmentwicklung mit dem „klassischen Quellcode" und dessen Prüfung und müssen uns schrittweise auf die DNA der Zukunft einstellen – der KI-Agent verändert sich fortlaufend, sekündlich, minütlich – da geht eine reine Quellcodeüberprüfung in längeren Abständen statisch in das „Leere" und ist in der Praxis meist bei KI-Agenten völlig wirkungslos. Wirkungsvoll sind kurzfristigere „Code-Reviews", um Schwachstellen und Risiken frühzeitig zu erkennen und Schäden und Risiken für die Institute zu verhindern. Das Thema sollte auch ein modernes IKT-Risikomanagement und alle beteiligten Akteure berücksichtigen; hierbei ist Handlungsbedarf gegeben.

## IKT-Meldewesen schwerwiegender Vorfälle

Cyberangriffe treten in teils unterschiedlichen Formen auf. Cyberangriffe durch autonome KI-Agenten nehmen im Jahr 2026 rasant zu, da automatisierte Systeme Schwachstellen viel schneller und in deutlich größerer Masse ausnutzen als menschliche Angreifer.[26]

KI-Systeme müssen während des Betriebs beispielsweise gegen gängige Cyberbedrohungen wie Adversarial Attacks[27], Model Poisoning[28] oder Inference Attacks[29] geschützt werden. Um diesen Bedrohungen zu begegnen, bedarf es geeigneter technischer Sicherheitsmaßnahmen (Art. 9 Abs. 3 DORA).

Aktuell fordert die EZB von den Bank-CEOs größerer Geldhäuser einen Aktionsplan gegen Cyberangriffe in Bezug auf die durch KI indizierten Risiken.[30] Die Vorstandschefs fordert sie darin auf, einen Aktionsplan zur Stärkung ihrer IT-Sicherheit zu entwickeln und ihn bis zum 31. Oktober 2026 bei den für sie zuständigen Aufsichtsteams der EZB einzureichen.[31]

Man wird sich in angespannten geopolitischen Situationen und im Umfeld der KI-Entwicklung künftig mit einer stark zunehmenden Angriffsgeschwindigkeit und einem viel größerem Umfang auseinandersetzen müssen. Hierbei ist der KI-Agent ein Schlüsselinstrument.[32]

> „Cyberangriffe durch autonome KI-Agenten nehmen im Jahr 2026 rasant zu."

## Durch KI-Agenten ausgelöste Risiken

Es gibt in der Praxis unterschiedliche Ansätze zur Messung der Risken von KI-

### Abbildung 3: Risiken von KI-Agenten

Bitcom-Leitfaden zu KI-Agenten – mehr als dreimal so viel Risiken wie im MVP-Portal aufgeführt – ist das interne IKT-Risikomanagement fit und kontrolliert diese umfassend = Vorstand, GL ist laut BaFin haftbar

| Risiken von KI-Agenten (Bitcom) | |
|---|---|
| **Angriffe auf die Agenten-Logik und -Steuerung** | |
| R1 | Intentionsstörung & Zielmanipulation |
| R2 | Fehlanpassung & Täuschendes Verhalten |
| R3 | Missbrauch von Werkzeugen |
| R4 | Prompt Injection |

| Risiken von KI-Agenten (Bitcom) | |
|---|---|
| **Angriffe auf Daten und Kommunikation** | |
| R5 | Kaskadierende Halluzinationsangriffe |
| R6 | Vergiftung der Agentenkommunikation |
| R7 | Speichervergiftung |

| Risiken von KI-Agenten (Bitcom) | |
|---|---|
| **Risiken in Multi-Agent-Systemen und durch menschliche Interaktion** | |
| R8 | Menschliche Angriffe auf Multi-Agent-Systemen |
| R9 | Menschliche Manipulation |
| R10 | Überlastung des menschlichen Kontrollrahmens |
| R11 | Schadhafte Agenten in Multi-Agent-Systemen |

| Risiken von KI-Agenten (Bitcom) | |
|---|---|
| **Systemische und operative Risiken** | |
| R12 | Kompromittierung von Berechtigungen |
| R13 | Ressourcenüberlastung |
| R14 | Abstreitbarkeit & fehlende Nachvollziehbarkeit |
| R15 | Identitätsfälschung & -imitation |
| R16 | Unerwartete Remote-Code-Ausführung & Code-Angriffe |

Quelle: Bitcom[36]

Agenten. Die Notwendigkeit ergibt sich aus DORA, insbesondere aus der aufsichtlich vorgegebenen Richtlinie/Methode zum Management von IKT-Assets.[35] Viele Institute bewerten ihre IKT-Assets (unter anderem auch KI-Systeme, Modelle gegebenenfalls KI-Agenten et cetera) im Rahmen des jährlich vorgenommenen IKT-Asset-Managements. Dabei wurde der meist altbewährte BSI-Grundschutz über die Elemente des Informationsverbunds mittels Schutzbedarfs- und IKT-Risikoanalyse bewertet. Nun reicht dies inhaltlich nicht mehr aus, denn das Risikoprofil auf die vier Schutzziele hat sich durch den Einsatz von KI-Agenten wesentlich erweitert, der Schutzbedarf steigt proportional an. Die Bitcom kommt alleine auf 16 unterschiedliche Risikoarten, wie in Abbildung 3 dargestellt.

Gleichzeitig ist durch den Einsatz von KI-Agenten ein wesentlich erweitertes Kontrollgerüst erforderlich, um den sicheren Einsatz in der Praxis der Finanzinstitute sicherzustellen. Die Bitcom schlägt sowohl technische, organisatorische und rechtliche/Compliance-Kontrollen vor.

Interessant ist, dass die Experten der Bitcom auch als beispielhafte Maßnahme zur Risikobeschränkung die Rechte- und Zugriffsbeschränkung der KI-Agenten vorschlagen.[38] Dies könnte über die bereits geschilderten Ausführungen im Identitäts- und Rechtemanagement erreicht werden. In Summe sind pro KI-Agent bei 3 Kontrollgruppen und 16 Risiken insgesamt 48 Kontrollarten erforderlich. Ein Institut mit 100 KI-Agenten benötigt dann in Summe 4800 neue Kontrollen. Diese könnten auch zur Prozessoptimierung geclustert beziehungsweise zusammengefasst werden. Es ist unbestritten, dass ein sicherer Einsatz von KI-Agenten mit finanziellem Aufwand für die IKT-Sicherheit verbunden ist.

> „Es ist unbestritten, dass ein sicherer Einsatz von KI-Agenten mit finanziellem Aufwand verbunden ist."

Zusammenfassend ist festzuhalten, dass das IKT-Risikomanagement in Finanzinstituten – gerade vor dem Hintergrund der KI-Agenten – intensiv zu verbessern ist. Ohne umfassende Investitionen in die IKT-Sicherheit ist ein gerade sicherer Einsatz von KI-Agenten in der Finanzindustrie nicht möglich. Auch die Schulungen und Weiterbildung der Mitarbeiter/-innen ist massiv zu intensivieren.

## DORA stößt an Grenzen

Man ist an dem Punkt angelangt, an dem erkennbar ist, dass die DORA-Regulierung vor dem Hintergrund risikoreicher KI am Beispiel des KI-Agenten in Teilbereichen an die Grenzen stößt. Aktuell hat das Bundesamt für Informationssicherheit vor Risiken aus dem selbstständigen Handeln der Agenten gewarnt und Sicherheitsempfehlungen bekannt gegeben.[39] Die BaFin hat als Aufsichtsbehörde eine weitere Hilfestellung mittels der im Dezember 2025 veröffentlichten Orientierungshilfe zu IKT-Risiken gegeben, die als erster Aufschlag weiterzuentwickeln ist. Jedoch zeigen die vielen Pannen, eigenständigen Ausreißer und Sicherheitsvorkommnisse sowie die zunehmenden

Abbildung 4: Maßnahmen/Kontrollen zur Verbesserung der Sicherheit bei KI-Agenten

| Technische Kontrolle (Technical controls) - Quelle: Was ist KI-Agentenmanagement? \| IBM | Organisatorische Kontrollen (Operational and Governance Controls) – Quelle: Was ist KI-Agentenmanagement? \| IBM | Rechtliche & Compliance-Kontrollen (Regulatory Controls) – Quelle: Weniger Aufwand, mehr Kontrolle: Wie Künstliche Intelligenz die Compliance neu definiert – Economic Crime Blog |
|---|---|---|
| • Guardrails (Leitplanken): Filter, die Eingaben (Prompts) und Ausgaben (Outputs) blockieren, wenn sie Richtlinien verletzen<br>• Architektonische Einschränkungen: Feste Programmierung (Hardcoding) von Verhaltensregeln, die das Modell nicht überschreiben kann<br>• API- und Funktions-Blacklisting: Verbot des Zugriffs auf kritische Systemfunktionen oder sensible externe Schnittstellen<br>• Sandboxing: Ausführung des Agenten in einer isolierten Umgebung ohne Zugriff auf das Hauptnetzwerk<br>• Ressourcenbegrenzung (Rate Limiting): Deckelung von Token-Verbrauch, Budget oder Rechenleistung pro Zeiteinheit<br>Nutzung IKT-Sicherheitssysteme (SIEM) und/ oder KI-Sicherheitssysteme (incl. KI-Agenten) | • Menschliche Freigabeschleife (Human-in-the-Loop): Erfordert menschliche Bestätigung vor kritischen Aktionen wie Geldtransfers oder E-Mail-Versand<br>• Bereitstellungs-Richtlinien (Deployment Policies): Regeln strikt, wer welche Agenten für welche Zwecke live schalten darf<br>• KI-Inventarisierung: Erfasst alle aktiven Agenten, deren Einsatzzweck und Risiko-Klassifizierung zentral<br>• Regelmäßige Kontrolltests: Prüfen die Agenten-Entscheidungen kontinuierlich auf Bias, Compliance und Drift | • EU-AI-Act-Konformität: Einstufung des Agenten in Risikoklassen mit entsprechenden Prüfpflichten<br>• Datenschutz-Kontrollen (DSGVO): Verhindern die unzulässige Verarbeitung oder Weitergabe personenbezogener Daten durch den Agenten<br>• Betriebliche Mitbestimmung: Regelt die Überwachung von Leistung und Verhalten der Mitarbeiter durch Agenten-Logs |

Quelle: Bitcom[37]

Cybervorfälle, die durch KI-Agenten ausgelöst werden auf, dass wir dringenden Handlungsbedarf in verschiedenen Bereichen des Finanzinstituts haben. Aktuell wurde auch das Deutsche Sicherheitsinstitut „AISI Deutschland" gegründet, um die Chancen und Risiken durch KI-Modelle zu evaluieren[40] – es ist zu hoffen, dass die KI-Agenten hierbei mitbehandelt werden.

Letztendlich benötigen wir den Einsatz von KI-Agenten nun auch zur Verbesserung der IKT-Sicherheit. Fehlende ID-Kennungen (wie bei natürlichen Nutzern), eine fehlende Berücksichtigung im Identitäts- und Rechtemanagement von KI-Agenten trotz weitreichender IKT-Rechte und Rollen der KI-Agenten und die Vielzahl von IKT-Risiken, ausgelöst durch KI-Agenten, legen die Latte an ein wirksames und effektives IKS und das IKT-Risikomanagement sehr hoch. Auch nicht einsehbare Audit-Trails von Agenten und lückenlose Protokollierungen sowie fehlende/unzureichende Überwachungsdokumentationen im Rahmen des internen Kontrollsystems zwingen zum Umdenken zugunsten einer höheren IKT-Sicherheit und ein höheres Maß an Compliance in der Finanzindustrie mit Folgen für den Datenschutz und auch die IT-Dienstleister, das heißt für die IKT-Drittparteienüberwachung.

**Fußnoten**

1) AI-Agent ist die englische Bezeichnung für KI-Agent und wird in Fachbeiträgen oft synonym verwendet.
2) Vgl. n-tv unter: Regulierung gefordert: Mehr als 1000 Experten warnen vor KI-Kontrollverlust - n-tv.de.
3) Vgl. n-tv unter: Regulierung gefordert: Mehr als 1000 Experten warnen vor KI-Kontrollverlust - n-tv.de.
4) Vgl. t3n.de unter: OpenAI warnt: Warum Angriffe durch KI-Agenten zum Alltag werden könnten.
5) Vgl. t3n.de unter: OpenAI warnt: Warum Angriffe durch KI-Agenten zum Alltag werden könnten.
6) Vgl. Ebenda.
7) Vgl. Bitcom: Leitfaden Security of AI-Agents Grundlagen, Risiken und Best Practices zur Absicherung von KI-Agenten, Seite 7
8) Vgl. CONCISCO unter: „KI-Agenten zwischen Hype und Realität – Eine strategische Analyse für Unternehmen – Conciso GmbH.
9) Vgl.: www.golem.de. Als erstes Land der Welt: Estland führt eID für KI-Agenten ein - Golem.de
10) Vgl. Ebenda.
11) Audit-Trail = Nachvollzug = vollständige und lückenlose Dokumentation und Protokollierung der Aktivitäten der KI-Agenten.
12) IKT bezieht sich auf Informations- und Kommunikations- Technologie.
13) Vgl. sncom.de unter: KI-Kennzeichnungspflicht 2026: Alle Pflichten nach Art. 50.
14) Vgl. DORA RTS RMF Art. 20 Identitätsmanagement.
15) Vgl. DORA RTS RMF, Art. 21.
16) Vgl. DORA RTS RMF Art. 21 e iv).
17) Vgl. skill-sprinters.de unter: 88 Prozent: KI-Agenten werden zum Sicherheitsrisiko.
18) Vgl. BaFin: Künstliche Intelligenz: BaFin veröffentlicht Orientierungshilfe zu IKT-Risiken unter: Aktuelles & Presse – Künstliche Intelligenz: BaFin veröffentlicht Orientierungshilfe zu IKT-Risiken – Bafin, Seite 13 von 38.
19) Vgl. MaRisk AT 4.3.4 Verwendung von Modellen Tz. 6.
20) Vgl. BaFin: Künstliche Intelligenz: BaFin veröffentlicht Orientierungshilfe zu IKT-Risiken unter: Aktuelles & Presse – Künstliche Intelligenz: BaFin veröffentlicht Orientierungshilfe zu IKT-Risiken - Bafin, S. 14 von 38.
21) Vgl. Ebenda, S. 14 von 38.
22) Vgl. Ebenda, S. 16 von 38.
23) Vgl. Ebenda, S. 16 von 38.
24) Gi.de – Gesellschaft für Informatik unter: Was ist Agentische KI? – Gesellschaft für Informatik e.V.
25) Gi.de – Gesellschaft für Informatik unter: Was ist Agentische KI? – Gesellschaft für Informatik e.V.
26) Vgl. eset.com unter: KI-Bedrohungen nehmen Fahrt auf: ESET-Prognosen für 2026 werden Realität.
27) Vgl. Wikipedia.org: Unter einem Adversarial Attack (auf Deutsch „feindlicher Angriff") versteht man eine gezielte Manipulation von künstlicher Intelligenz (KI) durch minimal veränderte Eingabedaten, die sogenannten Adversarial Examples (feindliche Beispiele).
28) Vgl. Station X: Model Poisoning (Modellvergiftung) bezeichnet eine Familie von Cyberangriffen, bei denen Angreifer gezielt die Parameter, Gewichte oder Trainingsdaten eines Systems für künstliche Intelligenz (KI) und maschinelles Lernen (ML) manipulieren, um das Verhalten des Modells dauerhaft zu verfälschen. Im Gegensatz zu Angriffen während der Laufzeit (wie Prompt Injection) greift Model Poisoning tief in die interne Logik des Modells ein, wodurch die Manipulation langfristig bestehen bleibt und extrem schwer zu erkennen ist unter: What Is Model Poisoning?
29) Vgl. IBM.com Ein Inference Attack (Rückschlussangriff) ist ein Sicherheitsangriff, bei dem aus indirekten Hinweisen oder Antworten eines Systems geheime Informationen abgeleitet werden, unter https://dataplatform.cloud.ibm.com/docs/content/wsj/ai-risk-atlas/membership-inference-attack.html?context=wx&locale=de.
30) Vgl. Handelsblatt.de unter: KI: EZB fordert von Bank-CEOs Aktionsplan gegen Cyberangriffe
31) Vgl. Ebenda.
32) Erstellt über www.google.de KI-Modus unter: „Prognose Zunahme der Cyberangriffe durch KI Agenten" (inhaltlich verifiziert).
33) Vgl. www.spiegel.de unter: Künstliche Intelligenz: Kriminelle nutzen zunehmend KI für Hacker-Angriffe – DER SPIEGEL.
34) Vgl. www.tagesschau.de unter: Könnten KI-gestützte Hackerangriffe zur neuen Gefahr werden? | tagesschau.de
35) Vgl. Bafin. Dokumentationsanforderungen an Finanzunternehmen gem. DORA – Richtlinie für das Management von IKT-Assets gem. Art. 4 RTS RMF i.V.m. Art. 9 Abs. 2 und 4 lit. C DORA.
36) Vgl. Bitcom: Leitfaden Security of AI-Agents Grundlagen, Risiken und Best Practices zur Absicherung von KI-Agenten, Seite 14.
37) Vgl. Bitcom: Leitfaden Security of AI-Agents Grundlagen, Risiken und Best Practices zur Absicherung von KI-Agenten, Seite 20 – 29.
38) Vgl. Bitcom: Leitfaden Security of AI-Agents Grundlagen, Risiken und Best Practices zur Absicherung von KI-Agenten, Seite 20.
39) Vgl. Bundesamt für Sicherhreit in der Informationstechnik unter: BSI – KI-Agenten.
40) Vgl. Bundesministerium für Digitales und Staatsmodernisierung Pressemitteilung 50/2026 vom 31.08.2026