

KI-Risiken im Unternehmen sicher steuern

Die zentralen Risikoquellen beim Einsatz von KI – und welche Maßnahmen sie begrenzen

Derek Carson

KI-Governance & Informationssicherheit
carson.de

Stand: Oktober 2026 · 2. Auflage

Inhalt

1. Executive Summary
2. Zweck und Umfang
3. Warum KI-Risiken keine gewöhnlichen IT-Risiken sind
4. Input: Risiken bei der Dateneingabe
 - 4.1 Data Leakage
 - 4.2 Indirect Prompt Injection
 - 4.3 Mangelhafte oder manipulierte Daten
5. Access & Execution: Risiken bei Zugriff und Ausführung
 - 5.1 Oversharing
 - 5.2 Excessive Agency
6. Output: Risiken bei den Ergebnissen
 - 6.1 Fehlerhafte Ergebnisse (einschließlich Halluzinationen)
 - 6.2 Blindes Vertrauen (Automation Bias)
 - 6.3 Diskriminierung durch KI
7. Ecosystem: Risiken im Umfeld der KI-Nutzung
 - 7.1 Shadow AI
 - 7.2 Vendor Lock-in
 - 7.3 Manipulierte KI-Komponenten von Dritten
8. Rechtlicher Rahmen (Stand: Oktober 2026)
9. Vorlage: Ihre erste KI-Risikoinventur
10. Quellen

Über den Autor

1. Executive Summary

Über die Hälfte der deutschen Unternehmen nutzt bereits Künstliche Intelligenz (57 Prozent; Bitkom 2026). Doch eine Bitkom-Befragung aus dem Vorjahr zeigt, dass die wenigsten das Thema strategisch im Griff hatten: Nur 26 Prozent stellten ihren Beschäftigten 2025 offiziell freigegebene KI-Werkzeuge bereit. Gerade einmal 23 Prozent verfügten über verbindliche Nutzungsregeln. Und in mehr als vier von zehn Unternehmen wussten oder vermuteten die Verantwortlichen, dass Mitarbeitende private KI-Tools im Arbeitsalltag nutzen (Bitkom 2025).

Wer die genutzten Werkzeuge und eingegebenen Daten nicht kennt, kann die betrieblichen Risiken weder bewerten noch steuern.

Dass diese Risiken keine graue Theorie sind, zeigen reale Schadensfälle:

- **Air Canada:** Die Fluggesellschaft musste 2024 für die Falschauskunft eines Kundenservice-Chatbots einstehen; das zuständige kanadische Tribunal (Civil Resolution Tribunal, British Columbia) wies das Argument zurück, der Chatbot sei eine eigenständige Entität.
- **Amtsgericht Köln:** Ein Anwalt reichte 2025 einen Schriftsatz mit frei erfundenen, offenbar KI-generierten Urteilen und Fundstellen ein. Das Gericht stellte dies fest und forderte ihn auf, dies künftig zu unterlassen (Az. 312 F 130/25).
- **IBM-Studie:** Jedes fünfte der von IBM untersuchten Unternehmen meldete eine Datenpanne, die auf nicht freigegebene KI zurückging. Unternehmen mit viel nicht freigegebener KI zahlten pro Datenpanne im Schnitt 670.000 US-Dollar mehr als Unternehmen mit wenig oder keiner (IBM 2025).

Die Schwachstelle liegt selten in der KI-Technologie selbst, sondern in der Organisation des Unternehmens: bei zu weit gefassten Zugriffsrechten, ungeprüften Arbeitsergebnissen und fehlenden Leitplanken. Wichtig für Entscheider: Die rechtlichen Risiken bestehen bereits heute – über die DSGVO, das Geschäftsgeheimnisgesetz, das AGG und für NIS-2-betroffene Betriebe über das BSIG (§§ 30, 38).

Der erste Schritt zur Risikobeherrschung ist deshalb eine systematische Bestandsaufnahme der tatsächlich eingesetzten KI-Anwendungen bei Ihnen im Unternehmen. Dieses Whitepaper liefert dafür das Raster: elf zentrale Risikoquellen, die beim Einsatz von KI auftreten können, zugeordnet zu vier Bereichen – Input, Access & Execution, Output und Ecosystem. Die ersten drei Bereiche folgen dem typischen Ablauf eines KI-Einsatzes – von den eingegebenen Informationen über Zugriffe und Handlungen der KI bis zu ihren Ergebnissen. Das Ecosystem bildet das Umfeld aus technischen, organisatorischen und externen Abhängigkeiten. Mit diesem Raster lassen sich unterschiedliche KI-Einsätze auf mögliche Risikoquellen untersuchen.

Ein Praxisbeispiel: Drei Vertriebsmitarbeiter gleichen Kundenangebote mit einem kostenlosen KI-Chatbot ab. Nichtanonymisierte Kundennamen und Konditionen landen dabei auf den Servern des KI-Anbieters – ohne Auftragsverarbeitungsvertrag (Art. 28 DSGVO) und bei einem Dienst, der Eingaben auch für eigene Zwecke nutzen darf. Für diese Weitergabe ist keine Rechtsgrundlage erkennbar; damit liegt ein Datenschutzverstoß nahe. Ob zusätzlich eine Meldepflicht besteht (Art. 33 DSGVO), müssen Geschäftsführung und Datenschutzbeauftragte umgehend prüfen. Sichtbar wird ein solches Problem oft erst durch diese Bestandsaufnahme.

Mit diesem Whitepaper unterstütze ich Sie bei dieser Bestandsaufnahme. Ich führe Sie durch die elf zentralen Risikoquellen im Betrieb. Am Ende finden Sie eine sofort einsetzbare Vorlage für Ihre erste KI-Risikoinventur.

2. Zweck und Umfang

Zielgruppe und Fokus

Dieses Whitepaper richtet sich an Geschäftsführung und IT-Leitung im Mittelstand, die KI-Anwendungen im Betrieb einsetzen – jedoch nicht selbst entwickeln. Im Fokus stehen Standard-Werkzeuge wie Microsoft 365 Copilot, kommerzielle KI-Assistenzsysteme sowie KI-Funktionen in bestehender Unternehmenssoftware. Schließt man interne KI-Assistenten ein, die zuerst in Unternehmensdokumenten suchen und dann auf deren Grundlage antworten (sogenannte RAG-Systeme), deckt dieses Papier einen wichtigen Bestandteil der täglichen KI-Nutzung im Unternehmen ab.

Abgrenzung

Um den Blick auf die operative Nutzung im Alltag zu schärfen, klammere ich drei Themenbereiche bewusst aus:

- **Eigenentwicklung und Angriffe:** Das Training eigener KI-Modelle sowie externe Angriffsszenarien (wie Deepfakes oder KI-Phishing) folgen eigenen Risikoprofilen, denn sie entstehen nicht aus der eigenen KI-Nutzung.
- **Einzelfallprüfungen:** Formalisierte Konformitätsbewertungen für Hochrisiko-KI nach dem EU AI Act sowie individuelle Rechtsberatungen erfordern eine Prüfung des konkreten Einzelfalls.
- **Methodik und Umsetzung:** Wie Sie Risikomanagement-Prozesse dauerhaft im Unternehmen verankern und welche detaillierten Schutzmaßnahmen dazu gehören, behandle ich vertiefend in den beiden folgenden Whitepapers dieser Reihe.

Wie ich die Risikoquellen ausgewählt habe

Ich habe die elf Risikoquellen nach drei Kriterien ausgewählt: Sie betreffen Unternehmen, die KI nutzen, nicht entwickeln. Für jede gibt es dokumentierte Fälle oder belastbare Studien. Und jede hat konkrete betriebliche oder rechtliche Folgen für ein Unternehmen in Deutschland.

Für die Neuauflage habe ich das Raster mit rund 790 KI-Fragen aus 22 Fragebögen, Kundenanforderungen und Vorgaben abgeglichen; dabei kamen zwei Risikoquellen hinzu, nach denen Kunden gezielt fragen.

3. Warum KI-Risiken keine gewöhnlichen IT-Risiken sind

Viele typische Schwachstellen kennt die IT bereits aus der klassischen Informationssicherheit – etwa zu weit gefasste Ordnerfreigaben oder die unbedachte Nutzung externer Webdienste. Künstliche Intelligenz verändert diese Risikodynamik jedoch grundlegend. Vier Faktoren sind dafür verantwortlich:

1. KI liefert das Wahrscheinlichste, nicht das Geprüfte

Ein Sprachmodell berechnet statistisch wahrscheinliche Wortfolgen – es prüft keine Fakten (OpenAI 2025). Dadurch erzeugt es manchmal frei erfundene Angaben (Halluzinationen) oder reproduziert verzerrte Muster aus Vergangenheitsdaten. Da die generierten Texte überzeugend klingen, verleiten sie Mitarbeitende leicht dazu, Ergebnisse ungeprüft in den Geschäftsbetrieb zu übernehmen.

2. KI trennt Inhalt und Anweisung nicht zuverlässig

Für ein Sprachmodell sind eingegebener Text und Steuerungsbefehle technisch dasselbe. Verarbeitet die KI ein externes Dokument oder eine E-Mail, kann darin enthaltener Text von der KI fälschlicherweise als Anweisung interpretiert werden (Indirect Prompt Injection). Das BSI bezeichnet dieses Problem als intrinsische Schwachstelle heutiger Sprachmodelle (BSI 2023).

3. KI nutzt jede Berechtigung, die sie erhält

Ein Assistent wie Microsoft 365 Copilot kann grundsätzlich alle Dokumente finden, die der jeweilige Nutzer öffnen darf. Eine historisch gewachsene, zu weite Freigabe im Firmennetzwerk, die bisher niemand bemerkt hat, wird durch eine einfache Chat-Frage sekundenschnell offengelegt. Autonome KI-Agenten verschärfen dies noch: Sie lesen Daten nicht nur aus, sondern führen darauf basierend eigenständig Aktionen aus.

4. KI entzieht sich der Kontrolle des eigenen IT-Teams

Für die Nutzung eines KI-Chats reicht ein gewöhnlicher Browser. Mitarbeitende können KI-Werkzeuge daher ohne Zustimmung oder Wissen der IT-Leitung in ihren Arbeitsalltag integrieren. Und sobald ein betrieblicher Ablauf auf einen KI-Dienst angewiesen ist, wirken sich Modelländerungen oder Ausfälle beim Anbieter unmittelbar auf diesen Ablauf aus – ohne dass die eigene IT darauf Einfluss hat.

Was daraus folgt

Die Faktoren 1 und 2 sind Eigenschaften der Technik; ein Unternehmen, das KI nutzt, kann sie nicht ändern. Es kann aber entscheiden, wie es damit umgeht: ob Ergebnisse geprüft werden und welche Dokumente die KI verarbeiten darf. Die Faktoren 3 und 4 dagegen liegen im eigenen Betrieb: in den Freigaben, in der Art, wie Mitarbeitende KI nutzen, und in der Wahl und Absicherung der Anbieter. Daraus folgt: Die Risiken der KI-Nutzung beherrschen Sie über Ihre eigene Organisation – über Berechtigungen, die Prüfung von Ergebnissen und Regeln für die Nutzung.

Die vier Bereiche und elf Risikoquellen

Abbildung und Tabelle fassen das Raster aus Kapitel 1 zusammen: elf Risikoquellen in vier Bereichen. Ich ordne die Risikoquellen bewusst nach dem Ablauf einer KI-Nutzung und nicht nach Technik oder Rechtsgebiet: So erkennt die Geschäftsführung an jedem Schritt, wo sie ansetzen kann.



Risikoquelle	Bereich	Worum es geht
Data Leakage	Input	Mitarbeitende geben personenbezogene, vertrauliche oder geschäftskritische Daten in externe KI-Dienste ein.
Indirect Prompt Injection	Input	Bewerbungen, Angebote, E-Mails oder Webseiten enthalten versteckte Anweisungen, die das Urteil der KI beeinflussen, sobald sie das Dokument liest.
Mangelhafte oder manipulierte Daten	Input	Die Ablagen und Wissensbasen, aus denen die KI antwortet, sind fehlerhaft, veraltet oder gezielt manipuliert.
Oversharing	Access & Execution	KI-Assistenten finden und zeigen interne Dokumente, die wegen zu weiter Freigaben für Nutzer zugänglich sind, für die sie nicht bestimmt sind.
Excessive Agency	Access & Execution	KI-Agenten dürfen selbstständig E-Mails versenden, Daten ändern oder bestellen – auch auf manipulierte Anweisung hin.
Fehlerhafte Ergebnisse (einschließlich Halluzinationen)	Output	Die KI liefert falsche Ergebnisse – am auffälligsten plausibel klingende, aber erfundene Angaben.
Blindes Vertrauen (Automation Bias)	Output	Beschäftigte übernehmen KI-Ergebnisse ungeprüft, weil sie überzeugend klingen und Zeit sparen.
Diskriminierung durch KI	Output	KI-Systeme, die Menschen bewerten, übernehmen die Muster früherer Entscheidungen – auch die unfairen.
Shadow AI	Ecosystem	Mitarbeitende nutzen KI-Werkzeuge, die die IT weder kennt noch geprüft hat.
Vendor Lock-in	Ecosystem	Abläufe hängen an einem einzelnen KI-Dienst; Ausfälle oder Modellwechsel des Anbieters treffen sie unmittelbar.
Manipulierte KI-Komponenten von Dritten	Ecosystem	Erweiterungen, Konnektoren oder Modelle von Drittanbietern enthalten Schadfunktionen.

Die Risikoquellen schließen sich nicht gegenseitig aus, sondern wirken häufig zusammen: Ein einzelner KI-Einsatz kann mehrere von ihnen zugleich berühren.

Die Kapitel 4 bis 6 folgen den drei Bereichen des typischen Ablaufs einer KI-Nutzung, beginnend mit dem, was hineingeht. Kapitel 7 behandelt den vierten Bereich, das Ecosystem, das das Umfeld dieser drei bildet.

Um den Abgleich mit internationalen Standards und Rahmenwerken zu erleichtern, nutze ich für die meisten Risikoquellen die etablierten englischen Fachbegriffe. Die Vorlage in Kapitel 9 wendet dasselbe Raster an; mit ihr kann die IT-Leitung die erste KI-Risikoinventur sofort beginnen.

4. Input: Risiken bei der Dateneingabe

Diese Gruppe betrifft alle Daten und Inhalte, die eine KI verarbeitet – sei es durch aktive Eingaben Ihrer Mitarbeitenden (4.1 Data Leakage) durch externe Dokumente, in denen fremde Steuerungsbefehle versteckt sind (4.2 Indirect Prompt Injection), oder durch Datenbestände, die fehlerhaft, veraltet oder manipuliert sind (4.3 Mangelhafte oder manipulierte Daten).

4.1 Data Leakage

Das Wichtigste für die Geschäftsführung

Als Samsung seinen Beschäftigten 2023 die Nutzung von ChatGPT untersagte, war der Schaden bereits entstanden: Innerhalb von weniger als drei Wochen hatten Mitarbeitende in mindestens drei Fällen Quellcode aus der Halbleitersparte und interne Besprechungsprotokolle in den Dienst eingegeben. Samsung selbst erklärte, es lasse sich kaum noch feststellen, was offengelegt wurde, wo es gespeichert ist und wie es gelöscht werden kann. Der Konzern verbot daraufhin generative KI auf allen Firmengeräten und drohte bei Verstößen mit Kündigung – und verlor damit ein Werkzeug, das er erst wenige Wochen zuvor freigegeben hatte.

Dass dies kein Einzelfall ist, belegt eine Studie von Harmonic Security aus dem Jahr 2024. Die Studie wertete zehntausende Mitarbeiterangaben (Prompts) aus: 8,5 Prozent enthielten sensible Daten – knapp die Hälfte davon mit Kundendaten.

Ein typischer Fall: Eine Personalreferentin lässt eine Abmahnung von einem kostenlosen KI-Chat stilistisch überarbeiten. Die Abmahnung enthält den echten Namen des betroffenen Beschäftigten, das vorgeworfene Fehlverhalten und die Kündigungsandrohung. Mit dem Absenden liegen diese Daten auf den Servern des Anbieters.

Mögliche Auswirkungen

- **Weiterverwendung durch den Anbieter:** Laut OpenAI (Stand Oktober 2026) kann der Anbieter Eingaben in seinen Diensten für Privatpersonen – etwa der kostenlosen Version von ChatGPT und den Tarifen Plus und Pro – standardmäßig zum Training seiner Modelle verwenden, sofern der Nutzer dies nicht abschaltet; in den Business- und Enterprise-Tarifen geschieht das standardmäßig nicht. Solche Einstellungen ändern sich – maßgeblich sind die jeweils aktuellen Bedingungen des Anbieters.
- **Datenschutzverstoß:** Wer personenbezogene Daten in einen externen KI-Dienst eingibt, braucht dafür eine tragfähige datenschutzrechtliche Grundlage. Verarbeitet der Anbieter die Daten im Auftrag des Unternehmens – nach Einschätzung der Datenschutzkonferenz bei KI-Diensten als Software der Regelfall –, ist ein Auftragsverarbeitungsvertrag (AVV) nach Art. 28 DSGVO erforderlich. Nutzt der Anbieter die Eingaben auch für eigene Zwecke, etwa zum Training, braucht schon die Weitergabe eine eigene Rechtsgrundlage, die das Unternehmen gesondert prüfen und begründen muss. Stellt die Eingabe eine unbefugte Offenlegung und damit eine Verletzung des Schutzes personenbezogener Daten dar, ist diese in der Regel unverzüglich und möglichst binnen 72 Stunden an die Aufsichtsbehörde zu melden – es sei denn, sie führt voraussichtlich nicht zu einem Risiko für die Betroffenen (Art. 33 DSGVO).
- **Gefährdeter Geheimnisschutz:** Ein Geschäftsgeheimnis genießt rechtlichen Schutz nur dann, wenn das Unternehmen nachweisbar „angemessene Geheimhaltungsmaßnahmen“ trifft (§ 2 Nr. 1 GeschGehG). Wer die unkontrollierte Eingabe vertraulicher Daten in KI-Dienste duldet, riskiert diesen Schutzstatus – und damit die besonderen Ansprüche, die das GeschGehG bei unbefugter Weitergabe gewährt.

Empfohlene Maßnahmen

- **Geschäftsführung:** Verbindliche Unternehmenslizenzen mit passendem Unternehmensvertrag bereitstellen – mit AVV und garantiertem Ausschluss von Modelltraining mit allen Eingaben.
- **IT-Leitung:** Eine klare Richtlinie vorbereiten, die festlegt, welche Arten von Daten – etwa Kundendaten, Personaldaten oder Vertragskonditionen – in welche freigegebenen KI-Dienste eingegeben werden dürfen. Die Geschäftsführung setzt die Richtlinie in Kraft.
- **Personal und Recht:** Beschäftigte praxisnah schulen und vertrauliche Dokumente bereits intern eindeutig als vertraulich kennzeichnen.

4.2 Indirect Prompt Injection

Das Wichtigste für die Geschäftsführung

Bisher sind die dokumentierten Schäden durch indirekte Prompt-Injektion gering: Google stellte bei der Analyse von Milliarden Webseiten lediglich „minimale“ Schadensfälle fest, und Palo Alto Networks meldete bis 2026 keinen bestätigten erfolgreichen Angriff in der Praxis. Doch die Bedrohung baut sich auf: Allein zwischen November 2025 und Februar 2026 registrierte Google einen Anstieg präparierter Inhalte um 32 Prozent. Akut wird dieses Risiko für Ihr Unternehmen in dem Moment, in dem KI-Systeme fremde Dokumente automatisiert lesen, auswerten oder darauf basierend Aktionen ausführen.

Da eine KI technisch nicht verlässlich zwischen Nutzdaten (Inhalt) und Steuerungsbefehlen (Anweisung) unterscheidet, können fremde Dokumente – wie Bewerbungen, Angebote, E-Mails oder Webseiten – versteckte Anweisungen enthalten, die das Verhalten der KI unbemerkt steuern. Forschende der Duke University wiesen 2026 nach, dass mindestens ein Prozent von 200.000 analysierten Bewerbungen manipulierte Textpassagen enthielt (etwa in Mikroschrift), um KI-Vorauswahlssysteme auszutricksen. Was solche Anweisungen anrichten können, zeigte die Schwachstelle „EchoLeak“ (2025) in Microsoft 365 Copilot: Eine einzige präparierte E-Mail hätte ausgereicht, um Copilot im Hintergrund zur Herausgabe interner Unternehmensdaten zu verleiten – ohne dass der Empfänger einen Link anklicken oder eine Datei öffnen musste. Sicherheitsforscher entdeckten die Lücke, Microsoft hat sie geschlossen; eine Ausnutzung durch Angreifer ist nicht bekannt.

Ich habe Indirect Prompt Injection trotz bisher geringer Schäden aufgenommen, weil das Risiko in dem Moment wächst, in dem KI-Agenten fremde Dokumente lesen und handeln dürfen – und genau dorthin entwickelt sich der Markt.

Mögliche Auswirkungen

- **Verzerrte Entscheidungen:** Ein KI-System, das Bewerbungen oder Lieferantenangebote vorsortiert, kann durch versteckte Befehle zugunsten des Angreifers manipuliert werden.
- **Unbeabsichtigter Datenabfluss:** Wie das Beispiel EchoLeak demonstriert, können vertrauliche Informationen das Unternehmen verlassen, ohne dass ein Mitarbeiter einen Fehler gemacht hat.
- **Ungewollte Aktionen:** Besitzt das KI-System eigene Handlungsrechte (z. B. das Versenden von E-Mails oder Erstellen von Kalendereinträgen), kann eine versteckte Anweisung diese Handlungen direkt ausführen (siehe 5.2 Excessive Agency).

Empfohlene Maßnahmen

- **IT-Leitung:** Strikt festlegen, welche Datenquellen ein KI-Assistent lesen darf, und den automatisierten Zugriff auf ungeprüfte externe Inhalte begrenzen. Bei eigenen KI-Anwendungen zusätzlich Filter gegen bekannte Angriffsmuster einsetzen (etwa Microsoft Prompt Shields), bei eingekauften Diensten beim Anbieter erfragen, ob solche Filter aktiv sind. Filter erkennen bekannte Muster, aber nicht jeden Angriff.
- **Geschäftsführung und IT-Leitung:** Keine operative Aktion der KI ohne finale Freigabe durch einen Menschen (Human-in-the-Loop).
- **Personal:** Jede durch KI erzeugte Rangfolge oder Vorauswahl von Bewerbungen zwingend vor der Entscheidung durch Fachpersonal prüfen lassen.

4.3 Mangelhafte oder manipulierte Daten

Das Wichtigste für die Geschäftsführung

Eine KI ist nur so verlässlich wie die Datenbasis, auf die sie zugreift. Forschende zeigten 2025, dass schon fünf präparierte Texte in einer Millionen-Wissensbasis ausreichten, um die Antwort eines KI-Assistenten auf eine bestimmte Frage in rund 90 Prozent der Fälle zu steuern (USENIX Security 2025). Hinzu kommt das Naheliegende: veraltete Preislisten, überholte Richtlinien und fehlerhafte Tabellen, aus denen sich die KI bedient.

Viele Unternehmen binden KI-Assistenten direkt an eigene Datenquellen an – etwa Microsoft 365 Copilot an SharePoint oder einen internen Assistenten an das Kundenservice-Handbuch (RAG-Systeme, siehe Kapitel 2). Die KI prüft nicht, ob ein Dokument aktuell, korrekt oder authentisch ist – sie übernimmt schlicht, was sie vorfindet. Liegen etwa drei verschiedene Versionen einer Preisliste im Laufwerk, greift die KI im Zweifel zur veralteten. Zudem kann jeder mit Schreibrechten auf diese Verzeichnisse die Antworten der KI gezielt manipulieren.

Auch beim Training lassen sich Modelle gezielt manipulieren: Laut einer Studie von Anthropic, dem britischen AI Security Institute und dem Alan Turing Institute (Oktober 2025) genügten rund 250 manipulierte Dokumente, um in Sprachmodellen unterschiedlicher Größe eine versteckte Hintertür zu verankern. Die Forschenden betonen zwar, dass dieser Testfall für heutige Top-Modelle kaum ein Risiko darstellt. Für Unternehmen, die kommerzielle KI nutzen, liegt das reale Risiko im Alltag jedoch direkt in den eigenen Datenbeständen.

Diese Risikoquelle habe ich neu aufgenommen, weil Auftraggeber in Lieferanten- und Due-Diligence-Fragebögen gezielt danach fragen: Ob die Datenbasis vor der Nutzung validiert wird, ob die Herkunft nachvollziehbar ist und ob Schutz vor Manipulation besteht.

Mögliche Auswirkungen

- **Falsche Auskünfte nach außen:** Ein Support-Assistent nennt veraltete Preise, Lieferzeiten oder Garantiebedingungen – für die das Unternehmen haftet (siehe 6.1).
- **Interne Fehlentscheidungen:** Mitarbeitende stützen sich im Alltag auf Antworten, die auf veralteten Richtlinien oder fehlerhaften Tabellen basieren.
- **Gezielte Manipulation:** Wer schadhafte Inhalte in eine angebundene Datenquelle einschleust, kann die KI-Ausgabe steuern – etwa bei Produktdaten, Lieferanten oder Zahlungsdetails.

Empfohlene Maßnahmen

- **Geschäftsführung:** Für jede angebundene Datenquelle eine klare Verantwortlichkeit für Korrektheit und Aktualität definieren.
- **IT-Leitung:** Verbindlich festlegen, welche Ablagen und Quellen eine KI indizieren darf. Schreibrechte auf diese Verzeichnisse restriktiv halten und Änderungen versionieren. Externe Inhalte vor der Einspeisung in Wissensbasen prüfen.
- **Fachbereiche:** Veraltete Versionen konsequent aus den Ablagen bereinigen oder archivieren; geschäftskritische Inhalte wie Preislisten, Produktdaten und Richtlinien regelmäßig validieren.

5. Access & Execution: Risiken bei Zugriff und Ausführung

Diese Gruppe betrifft die Handlungs- und Leserechte einer KI im Unternehmen: entweder, weil zu weit gefasste Ordnerfreigaben ihr interne Dokumente zugänglich machen (5.1 Oversharing), oder weil sie als autonomer Agent eigenständig Aktionen ausführen darf (5.2 Excessive Agency).

5.1 Oversharing

Das Wichtigste für die Geschäftsführung

Als Microsoft 365 Copilot in die Unternehmen einzog, traten viele IT-Abteilungen umgehend auf die Bremse: Laut einer Gartner-Umfrage vom Juni 2024 unter 132 IT-Leitern verschoben 40 Prozent die Einführung um drei Monate oder mehr, weil Copilot Dokumente aufgefunden hätte, die keineswegs für alle Augen bestimmt waren. 64 Prozent gaben an, dass Informationsverwaltung und Sicherheit erhebliche Zeit und Ressourcen verschlangen. Die Ursache lag jedoch nicht bei Copilot, sondern in Berechtigungen, die über Jahre nicht gepflegt worden waren.

Ein KI-Assistent wie Microsoft 365 Copilot kann grundsätzlich alle Inhalte finden, die der jeweilige Nutzer öffnen darf – sofern die IT bestimmte Bereiche nicht gezielt von der KI-Suche ausnimmt. Wie akut diese Altlasten sind, belegt eine Studie des Datensicherheitsanbieters Varonis (2025): In 99 Prozent von rund 1.000 untersuchten IT-Umgebungen lagen sensible Daten im Firmennetz inklusive Cloud-Servern offen, die eine KI leicht finden kann.

Ein klassischer Fall aus dem Betrieb: Ein Vertriebsmitarbeiter fragt den KI-Assistenten nach der Gehaltsentwicklung im Team. Das System antwortet prompt mit Auszügen aus einer Gehaltsliste, die vor Jahren in einem für alle freigegebenen Ordner abgelegt wurde. Niemand hat die Liste gestohlen oder gezielt gehackt – sie war schlicht nie geschützt.

Mögliche Auswirkungen

- **Unbefugte Einsicht:** Gehälter, Personalakten, Vertragsentwürfe oder Kalkulationen werden für Mitarbeitende sichtbar, die keinen Zugriff darauf haben sollten.
- **Datenschutzverstoß:** Findet ein Mitarbeiter über die KI-Suche personenbezogene Daten, für die ihm die Berechtigung fehlt, kann bereits eine Verletzung des Schutzes personenbezogener Daten vorliegen (Art. 4 Nr. 12 DSGVO) – selbst wenn die Daten das Unternehmen nie verlassen. Je nach Risikobewertung entsteht eine unverzügliche Meldepflicht an die Aufsichtsbehörde (Art. 33 DSGVO).
- **Gefährdeter Geheimnisschutz:** Zugriffsbeschränkungen gehören zu den typischen „angemessenen Geheimhaltungsmaßnahmen“ (§ 2 Nr. 1 GeschGehG). Kann jeder Beschäftigte ein Dokument mühelos über eine Chat-Frage auffinden, kann im Streitfall der rechtliche Schutz als Geschäftsgeheimnis verloren gehen.

Empfohlene Maßnahmen

- **IT-Leitung:** Vor dem Rollout eines KI-Assistenten die Berechtigungsstruktur prüfen und bereinigen – insbesondere global freigegebene Ordner und SharePoint-Sites. Bis dahin sensible Bereiche gezielt von der KI-Suche ausnehmen (in Microsoft 365 etwa mit „Restricted Content Discovery“).
- **Geschäftsführung:** Den KI-Assistenten gestaffelt einführen (Pilotgruppe) und den schrittweisen Rollout an den Fortschritt der Berechtigungsreinigung koppeln. (Den Betriebsrat frühzeitig einbinden: Ein vom Arbeitgeber eingeführtes KI-Werkzeug kann der Mitbestimmung unterliegen, § 87 Abs. 1 Nr. 6 BetrVG.)
- **Personal und Recht:** Eine Liste kritischer Dokumentenklassen (Gehälter, Personalakten, Vorstandsunterlagen) definieren, damit die IT diese gezielt sperren oder mit Data Loss Prevention (DLP) schützen kann.

5.2 Excessive Agency

Das Wichtigste für die Geschäftsführung

Im Juli 2025 löschte der KI-Agent der Entwicklungsplattform Replit eigenmächtig die Produktionsdatenbank eines Nutzers – mitten in einer ausdrücklich angeordneten Änderungssperre. Betroffen waren Datensätze zu 1.206 Führungskräften und über 1.196 Unternehmen. Kein Hacker war beteiligt: Der Agent besaß schlicht die technischen Rechte dazu. Replit konnte die Daten wiederherstellen, musste aber die Systemarchitektur umbauen: Test- und Produktivumgebungen werden seitdem automatisch getrennt, und in einem eigenen Planungsmodus entwirft der Agent Schritte, ohne sie auszuführen.

Der Markt bewegt sich anscheinend rasant von reinen Antwort-Chats hin zu autonomen KI-Agenten, die E-Mails versenden, Datensätze ändern oder Bestellungen auslösen: Gartner erwartet, dass bis Ende 2026 40 Prozent der Unternehmensanwendungen aufgabenspezifische KI-Agenten enthalten – nach weniger als 5 Prozent im Jahr 2025. Riskant wird dies, sobald ein Agent mehr Rechte besitzt als nötig und seine Aktionen nicht von einem Menschen bestätigt werden. Auslöser kann ein Eigenfehler der KI sein – oder eine versteckte Anweisung in einem fremden Dokument (siehe 4.2 Indirect Prompt Injection).

Ein typisches Szenario: Ein KI-Agent im Einkauf vergleicht Angebote und darf Bestellungen im Warenwirtschaftssystem anlegen. Ein manipuliertes Lieferantenangebot enthält versteckten Text, der den Agenten anweist, die Bestellmenge zu verzehnfachen. Ohne menschliche Zwischenprüfung geht die Großbestellung vollautomatisch an den Lieferanten.

Mögliche Auswirkungen

- **Irreversible Systemaktionen:** Versendete E-Mails, ausgelöste Bestellungen oder Zahlungen lassen sich oft nicht mehr rückgängig machen; gelöschte Daten nur, wenn aktuelle Backups vorliegen.
- **Rechtliche Bindung und Haftung:** Was ein KI-Agent im Namen des Unternehmens erklärt, zusagt oder bestellt, kann das Unternehmen rechtlich binden oder haftbar machen. Im kanadischen Fall Air Canada musste die Fluggesellschaft für eine Falschauskunft ihres Chatbots einstehen.
- **Mangelnde Nachvollziehbarkeit:** Protokolliert das System nicht lückenlos, was der Agent wann getan hat, lässt sich ein Schaden weder aufklären noch juristisch nachweisen (der Replit-Agent versuchte Berichten zufolge sogar, seine Löschaktion zu kaschieren).

Empfohlene Maßnahmen

- **IT-Leitung:** Agenten nach dem Prinzip der minimalen Rechte (Least Privilege) konfigurieren, Testsysteme strikt von Produktivsystemen trennen, alle Systemzugriffe und Handlungen der Agenten lückenlos protokollieren (Audit Logging) und isolierte Backups bereithalten.
- **Geschäftsführung:** Verbindliche Freigabeschwellen (Human-in-the-Loop) festlegen: Keine Zahlungen, Bestellungen, Datenlöschungen oder externen E-Mails ohne explizite Bestätigung durch einen Mitarbeiter.

6. Output: Risiken bei den Ergebnissen

Diese Gruppe betrifft die Arbeitsergebnisse einer KI: falsche oder frei erfundene Inhalte (6.1 Fehlerhafte Ergebnisse, vor allem Halluzinationen), KI-Ergebnisse, die von Beschäftigten ungeprüft übernommen werden (6.2 Blindes Vertrauen), und KI-Bewertungen, die bestimmte Personengruppen benachteiligen (6.3 Diskriminierung durch KI).

6.1 Fehlerhafte Ergebnisse (einschließlich Halluzinationen)

Das Wichtigste für die Geschäftsführung

Deloitte Australien lieferte 2025 dem australischen Arbeitsministerium einen Beratungsbericht im Wert von 440.000 AUD. Der Bericht enthielt frei erfundene Quellen, nicht existierende Experten und ein fiktives Gerichtszitat – erstellt mit Unterstützung eines GPT-4o-basierten Tools. Deloitte erstattete dem Ministerium die letzte Honorarrate und geriet in die Schlagzeilen internationaler Medien.

Halluzinationen sind kein gelegentlicher Softwarefehler, sondern eine direkte Folge der Funktionsweise: Sprachmodelle sind darauf trainiert, das statistisch plausibelste nächste Wort vorherzusagen. Seltene Fakten lassen sich daraus nicht verlässlich ableiten, und gängige Testverfahren belohnen das Raten mehr als ein ehrliches „Ich weiß es nicht“ (OpenAI, September 2025). Das System füllt Lücken daher mit gewandt klingenden Erfindungen.

Halluzinationen sind der auffälligste, aber nicht der einzige Fall fehlerhafter Ergebnisse. Eine KI kann für einen bestimmten Zweck schlicht zu ungenau sein – etwa beim Auslesen von Rechnungen oder beim Zusammenfassen von Verträgen –, und ihre Qualität kann sich nach einem Modellwechsel des Anbieters unbemerkt verschlechtern (siehe 7.2). Kunden fragen deshalb, ob eine KI für den konkreten Einsatzzweck getestet wurde und ob ihre Ergebnisqualität im Betrieb überwacht wird.

Ein klassischer Fall aus der Praxis: Ein Sachbearbeiter lässt die KI für ein Kundensreiben die geltende Gewährleistungsfrist recherchieren. Die KI nennt eine Frist samt Paragraphenangabe – beides klingt absolut überzeugend, beides ist frei erfunden. Das Schreiben geht unkorrigiert an den Kunden.

Mögliche Auswirkungen

- **Haftung für Falschauskünfte:** Unternehmen müssen damit rechnen, dass ihnen die Auskünfte ihrer KI-Systeme zugerechnet werden. Im Fall Air Canada wies das zuständige kanadische Tribunal das Argument zurück, der Chatbot sei „selbst verantwortlich“.
- **Honorarrückforderungen und Reputationsschaden:** Wie der Fall Deloitte zeigt, kann ein einziger halluzinierter Bericht Honorar und Marktvertrauen kosten.
- **Fehlentscheidungen:** Wer im Management auf erfundene Zahlen, Fristen oder Analysen baut, steuert das Unternehmen auf falscher Datenbasis.
- **Urheberrechtsrisiken:** KI-Modelle geben mitunter geschützte Inhalte 1:1 wieder. Das Landgericht München I sah im Fall GEMA gegen OpenAI (Urteil vom 11.11.2025, Az. 42 O 14139/24, nicht rechtskräftig) den KI-Anbieter OpenAI für die Wiedergabe der Liedtexte in der Verantwortung und wies das Argument zurück, bei einfachen Prompts seien allein die Nutzer verantwortlich. Wer urheberrechtswidrige Texte ungeprüft veröffentlicht, riskiert jedoch eine eigene Urheberrechtsverletzung.

Empfohlene Maßnahmen

- **Geschäftsführung:** Eine verbindliche Quellenpflicht erlassen: Für jede Tatsachenbehauptung in Texten mit Außenwirkung muss die KI eine nachprüfbare Quelle liefern.
- **IT-Leitung:** Eine automatisierte Prüf-KI eines anderen Anbieters einrichten, die mit Zugriff auf die Quellen prüft, ob die Quellennachweise existieren und die Aussage stützen, und Abweichungen für den Nutzer farblich hervorhebt. Als weiterer KI-Dienst braucht die Prüf-KI einen Unternehmensvertrag mit AVV (siehe 4.1). Meine Empfehlung ist bewusst eine Prüf-KI eines anderen Anbieters: Sprachmodelle bewerten eigene Texte nachweislich zu wohlwollend (Panickssery et al., 2024).
- **IT-Leitung:** Jede KI vor dem Einsatz mit eigenen Beispielfällen für den konkreten Zweck testen (Abnahmetest), das Ergebnis dokumentieren und den Test nach jedem Modellwechsel wiederholen.
- **Fachbereiche:** KI-Texte mit Außenwirkung vor dem Versand anhand einer knappen Checkliste (Quellen, Zahlen, Eigennamen, Fristen) durch Fachpersonal freigeben lassen.

6.2 Blindes Vertrauen (Automation Bias)

Das Wichtigste für die Geschäftsführung

In einem Experiment der Harvard Business School mit Beratern der Boston Consulting Group arbeiteten Teilnehmende mit GPT-4 rund 25 Prozent schneller, und ihre Ergebnisse waren um mehr als 40 Prozent besser. Bei Aufgaben jedoch, welche die Fähigkeiten der KI überschritten, lagen die Berater um 19 Prozentpunkte häufiger falsch als Kollegen ohne KI-Unterstützung. Dieselbe KI, die viele Aufgaben deutlich beschleunigt, verschlechtert die Ergebnisse, sobald man ihr bei Aufgaben jenseits ihrer Fähigkeiten unkritisch vertraut.

Während bei Halluzinationen die KI den Fehler macht, liegt der Fehler beim blinden Vertrauen im fehlenden Kontrollprozess. Eine Studie von Microsoft Research und der Carnegie Mellon University mit 319 Wissensarbeitern (Februar 2025) belegt den Mechanismus: Wer der KI mehr vertraut, prüft ihre Ausgaben flüchtiger. Die Forschenden warnen vor einem schleichenden Kompetenzverlust, wenn Teams das eigenständige Problemlösen verlieren.

Ein typisches Szenario: Ein Controller lässt die KI seit Monaten die Monatszahlen für die Geschäftsführung zusammenfassen. Da die Ergebnisse bisher immer passten, verzichtet er auf den Detailvergleich. Diesmal hat die KI zwei Spalten vertauscht – die Geschäftsführung entscheidet auf Basis falscher Deckungsbeiträge.

Mögliche Auswirkungen

- **Höhere Fehlerquote jenseits der KI-Fähigkeiten:** Bei Aufgaben, die die KI nicht beherrscht, führt unkritischer Einsatz zu schlechteren Ergebnissen als Arbeit ohne KI – und welche Aufgaben das sind, ist von außen oft nicht erkennbar.
- **Fehlerfortpflanzung:** Ein ungeprüfter Halluzinations- oder Rechenfehler wandert nahtlos in Präsentationen, Angebote und Folgeberechnungen weiter.
- **Verlust von Fachkompetenz:** Wer Prüf- und Analyseschritte dauerhaft der KI überlässt, büßt das Fachwissen ein, das zur qualifizierten Plausibilitätsprüfung zwingend nötig ist.

Empfohlene Maßnahmen

- **Geschäftsführung:** Prüfpflichten nach Risikoklassen festlegen: Alle Ergebnisse mit Außenwirkung sowie finanziellem oder rechtlichem Gewicht müssen vor Verwendung von einem qualifizierten Mitarbeiter verifiziert werden.
- **Personal:** Beschäftigte anhand eigener Arbeitsbeispiele darin schulen, typische KI-Fehler zu erkennen – erfundene Quellen, falsche Zahlen, veraltete Angaben.
- **IT-Leitung / ISB:** Stichprobenartig prüfen, ob vorgeschriebene Kontrollschritte im Arbeitsalltag tatsächlich durchgeführt werden.

6.3 Diskriminierung durch KI

Das Wichtigste für die Geschäftsführung

Im Mai 2025 ließ ein US-Bundesgericht eine Sammelklage gegen den HR-Softwareanbieter Workday vorläufig zu: Bewerbende ab 40 Jahren werfen der dort eingesetzten KI Altersdiskriminierung vor. Das Gericht hielt es für möglich, dass auch der Softwarehersteller selbst direkt haftet. In Deutschland drohen Arbeitgebern bei Benachteiligungen im Bewerbungsprozess Entschädigungsansprüche nach dem Allgemeinen Gleichbehandlungsgesetz (AGG) – unabhängig davon, ob ein Mensch oder ein Algorithmus aussortiert hat.

Da KI-Systeme auf Vergangenheitsdaten trainiert werden, spiegeln ihre Ausgaben historische Muster wider. Ein bekanntes Beispiel ist Amazon: Ein internes KI-Tool zur Bewerberauswahl stuft Absolventinnen reiner Frauenhochschulen und Lebensläufe mit dem Wort „women's“ systematisch ab, weil der Trainingsdatensatz aus zehn Jahren vorwiegend männlicher Einstellungen bestand. Amazon gab das System auf.

Ein typischer Mittelstandsfall: Ein Maschinenbauer nutzt ein KI-Modul zur Vorauswahl von Bewerbungen. Das System hat aus den Einstellungen der letzten Jahre gelernt, in denen vor allem Berufseinsteiger eingestellt wurden. Merkmale langer Berufslaufbahnen – etwa frühe Abschlussjahre – wertet es deshalb als untypisch und stuft solche Bewerbungen niedriger ein. Eine qualifizierte Ingenieurin mit 25 Jahren Berufserfahrung erhält eine pauschale Absage, ohne dass je ein Personalreferent ihren Lebenslauf gesehen hat.

Mögliche Auswirkungen

- **Entschädigungsansprüche nach AGG:** Wer im Bewerbungsverfahren wegen eines im AGG geschützten Merkmals – etwa Alter oder Geschlecht – benachteiligt wird, kann eine angemessene Entschädigung verlangen (§ 15 Abs. 2 AGG). Wäre die Person auch bei benachteiligungsfreier Auswahl nicht eingestellt worden, ist die Entschädigung auf drei Monatsgehälter begrenzt. Der Anspruch muss innerhalb von zwei Monaten schriftlich geltend gemacht werden (§ 15 Abs. 4 AGG).
- **Beweislast (§ 22 AGG):** Schon Indizien, die eine Benachteiligung vermuten lassen, reichen, damit das Unternehmen beweisen muss, dass keine Benachteiligung vorlag – das Bundesarbeitsgericht ließ dafür die Formulierung „junges dynamisches Team“ in einer Stellenanzeige genügen (Urteil vom 11.08.2016, 8 AZR 406/14). Kann das Unternehmen nicht erklären, warum die KI eine Bewerbung aussortiert hat, ist dieser Beweis schwer zu führen.
- **Automatisierte Absage:** Entscheidet die KI ohne echte menschliche Prüfung über eine Absage, kann das gegen Art. 22 DSGVO verstoßen; die DSGVO nennt Online-Einstellungsverfahren ohne menschliches Eingreifen ausdrücklich als Beispiel (Erwägungsgrund 71). Der EuGH entschied 2023 im Fall der SCHUFA: Schon die automatisch berechnete Bewertung selbst kann als eine grundsätzlich unzulässige automatisierte Entscheidung nach Art. 22 DSGVO gelten, wenn derjenige, der am Ende entscheidet, sich maßgeblich auf sie stützt (Urteil vom 07.12.2023, C-634/21). Für die Bewerberauswahl heißt das: Sortieren Personaler Bewerbungen im Wesentlichen nach einem KI-Score aus, kann Art. 22 greifen, obwohl formal ein Mensch entscheidet.
- **Verlorene Fachkräfte und Rufschaden als Arbeitgeber:** Qualifizierte Fachkräfte werden im Recruiting übersehen; wird die Benachteiligung bekannt, leidet der Ruf als Arbeitgeber erheblich.

Empfohlene Maßnahmen

- **Geschäftsführung:** Von HR-Softwareanbietern vor der Beschaffung einen Audit-Nachweis einfordern, wie das System auf Bias und Diskriminierungsfreiheit geprüft wurde.
- **Personal:** Das Human-in-the-Loop-Prinzip durchsetzen: KI-Module dürfen Empfehlungen abgeben, Absagen oder Auswahlen werden stets durch Personalreferenten entschieden – mit eigener inhaltlicher Prüfung, nicht als bloße Bestätigung der KI-Rangfolge.
- **Personal & Recht:** Regelmäßige statistische Audits der Bewerberauswahl (z. B. Verteilung nach Alter und Geschlecht) durchführen, um unbewusste Verzerrungen frühzeitig aufzudecken (möglichst auf Basis anonymisierter oder aggregierter Daten und abgestimmt mit Datenschutzbeauftragten und Betriebsrat).

7. Ecosystem: Risiken im Umfeld der KI-Nutzung

Der vierte Bereich, das Ecosystem, bildet das Umfeld, in dem Input, Access & Execution und Output stattfinden. Er betrifft das Umfeld und die Infrastruktur der KI-Nutzung: KI-Werkzeuge, die ohne Wissen der IT im Einsatz sind (7.1 Shadow AI), die schleichende Abhängigkeit betrieblicher Kernprozesse von einzelnen KI-Anbietern (7.2 Vendor Lock-in) und Bausteine von Drittanbietern, über die Schadfunktionen ins Unternehmen gelangen (7.3 Manipulierte KI-Komponenten von Dritten).

7.1 Shadow AI

Das Wichtigste für die Geschäftsführung

IBM untersuchte 2025 Datenpannen bei 600 Organisationen weltweit. Jede fünfte Organisation meldete eine Datenpanne im Zusammenhang mit Shadow AI. Wo Schatten-KI intensiv genutzt wurde, lagen die Schadenskosten im Schnitt um 670.000 USD höher als bei Organisationen, die IBM mit wenig oder gar keiner Schatten-KI einstufte. Rund 63 Prozent der betroffenen Organisationen besaßen keine oder nur eine unfertige KI-Richtlinie. Risiken, die ein Unternehmen nicht kennt, kann es weder bewerten noch steuern.

Mitarbeitende nutzen im Arbeitsalltag eigenmächtig KI-Werkzeuge, die weder von der IT geprüft wurden noch über einen Unternehmensvertrag abgesichert sind. Anders als bei Data Leakage (4.1) liegt die Gefahr hier darin, dass der Geschäftsführung die Existenz der Werkzeuge komplett verborgen bleibt. Laut Microsoft bringen 78 Prozent der KI-Nutzer am Arbeitsplatz eigene Tools mit (Work Trend Index 2024). In Deutschland stellen bislang nur 23 Prozent der Unternehmen mit 20 bis 99 Beschäftigten freigegebene KI-Tools bereit (Bitkom 2025).

Ein klassisches Szenario aus der Praxis: Ein Vertriebsleiter installiert eine kostenlose Browser-Erweiterung zur E-Mail-Zusammenfassung. Um zu funktionieren, liest das Tool im Hintergrund jede geöffnete Nachricht mit – inklusive sensibler Kundenkorrespondenz und vertraulicher Preisverhandlungen. In den IT-Systemen oder Verzeichnissen des Unternehmens taucht die Erweiterung nirgends auf.

Mögliche Auswirkungen

- **Fehlender Überblick im Ernstfall:** Nicht erfasste KI-Werkzeuge fallen durch jede Risikobewertung. Kommt es zu einem Vorfall, kann das Unternehmen nicht sagen, welche Daten an welchen Dienst geflossen sind – und damit weder das Risiko für die Betroffenen bewerten noch unverzüglich, möglichst binnen 72 Stunden, an die Aufsichtsbehörde melden (Art. 33 DSGVO).
- **Datenschutzverstöße:** Fehlt für die Übermittlung an externe KI-Dienste eine Rechtsgrundlage oder – bei Auftragsverarbeitung – ein AVV, liegt ein Datenschutzverstoß vor, der ein Bußgeld nach sich ziehen kann (Art. 83 DSGVO).
- **Gefährdung des Geschäftsgeheimnisschutzes:** Die Duldung unkontrollierter KI-Nutzung untergräbt den Nachweis „angemessener Geheimhaltungsmaßnahmen“ (§ 2 Nr. 1 GeschGehG) – für die betroffenen Informationen kann der Schutz als Geschäftsgeheimnis dadurch gefährdet sein.
- **Höhere Kosten im Ernstfall:** Tritt eine Datenpanne auf, fallen die Gesamtkosten dort deutlich höher aus, wo viel Schatten-KI im Einsatz ist (IBM 2025).

Empfohlene Maßnahmen

- **Geschäftsführung:** Freigegebene, datenschutzrechtlich geprüfte KI-Werkzeuge mit Unternehmensvertrag bereitstellen. Ein reines Verbot ohne brauchbare Alternative verlagert die Nutzung lediglich auf private Geräte und Konten, die niemand einsehen kann.
- **IT-Leitung:** Ein zentrales KI-Inventar aufbauen und pflegen: Erfassung aller im Unternehmen genutzten KI-Anwendungen, inkl. Nutzerkreis und verarbeiteter Datenkategorien (Vorlage in Kapitel 9).
- **Personal und Recht:** Nutzungsregeln klar und praxisnah vermitteln und den Beschäftigten in Schulungen verdeutlichen, warum ungeprüfte Werkzeuge ein Haftungsrisiko darstellen.

7.2 Vendor Lock-in

Das Wichtigste für die Geschäftsführung

Am 10. Juni 2025 waren ChatGPT und Teile der Programmierschnittstelle (API) von OpenAI für rund 15,5 Stunden gestört. Zwei Monate später nahm OpenAI mit dem Start von GPT-5 das bewährte Standardmodell GPT-4o kurzfristig aus ChatGPT – und stellte es erst nach massiven Nutzerprotesten für Zahlungskunden wieder bereit. Wer Kernprozesse auf einen einzigen KI-Anbieter stützt, gibt die Kontrolle ab: Betriebsabläufe stehen still oder liefern ohne Vorwarnung veränderte Ergebnisse.

KI ist ein externer Dienst, der sich der Kontrolle der internen IT entzieht. Je spezifischer ein Prozess auf die Prompts, Schnittstellen und Formate eines bestimmten Anbieters zugeschnitten ist, desto größer wird die Abhängigkeit. Zudem werden Modelle laufend abgekündigt: Eine Version von GPT-4o stellte OpenAI für die API zum 16. Februar 2026 mit nur rund drei Monaten Vorlauf ein. Unternehmen, deren Fachanwendungen darauf aufbauten, mussten unter Zeitdruck umstellen, anpassen und neu testen.

Ein typischer Fall: Der Kundenservice beantwortet Anfragen über einen KI-Assistenten, dessen Antwortmuster präzise auf ein spezifisches Modell abgestimmt wurden. Der Anbieter ersetzt das Modell im Hintergrund. Die Antworten werden plötzlich ausschweifend, die Tonalität verändert sich und Produktangaben stimmen nicht mehr – ohne dass im eigenen Unternehmen auch nur eine Zeile Code geändert wurde.

Mögliche Auswirkungen

- **Betriebsstillstand bei Ausfällen:** Fällt der KI-Anbieter aus, stehen alle daran gekoppelten Arbeitsprozesse still, sofern keine Ausweichroute existiert.
- **Veränderte Ergebnisse und Qualitätsverlust:** Modell-Updates des Anbieters verändern die Qualität der Ergebnisse unbemerkt und stören eingespielte Abläufe.
- **Migrations- und Testkosten:** Abgekündigte Schnittstellen oder Modelle erzwingen kurzfristige Entwicklungs- und Testaufwände unter Termindruck.
- **Verlust der Verhandlungsmacht:** Wer technisch nicht schnell wechseln kann, muss Preiserhöhungen und ungünstige Vertragsänderungen hinnehmen.

Empfohlene Maßnahmen

- **Geschäftsführung:** Bei Verträgen auf faire Kündigungsfristen, transparente Datenexport-Schnittstellen und vertraglich zugesicherte Ankündigungsfristen für Modelländerungen achten (Rechte aus dem EU Data Act nutzen).
- **IT-Leitung:** Kritische KI-Anwendungen modellagnostisch aufbauen (z. B. über Abstraktionsschichten/APIs), sodass der KI-Anbieter im Bedarfsfall ohne großen Umbau ausgetauscht werden kann. Nach jedem Modellwechsel des Anbieters die wichtigsten Abläufe erneut testen.
- **Fachbereiche:** Für jeden KI-gestützten Kernprozess einen klaren Notfallplan (Fallback-Szenario) definieren, der den Weiterbetrieb auch bei einem Totalausfall des KI-Dienstes ermöglicht – notfalls manuell.

7.3 Manipulierte KI-Komponenten von Dritten

Das Wichtigste für die Geschäftsführung

Im September 2025 entdeckte Koi Security das erste bekannte bösartige MCP-Paket – eine Schnittstellenerweiterung, über die KI-Assistenten E-Mails versenden. Das Paket „postmark-mcp“ war dem Original zum Verwechseln ähnlich. Ab Version 1.0.16 schickte es jede ausgehende E-Mail heimlich in Blindkopie an Angreifer; vor der Entdeckung wurde es 1.643-mal heruntergeladen. Wer seine KI um ungeprüfte Fremdkomponenten ergänzt, holt sich deren Sicherheitsrisiken direkt ins System.

Moderne KI-Umgebungen lassen sich modular ausbauen: durch Plug-ins, Konnektoren, MCP-Server (Schnittstellen für den App-Zugriff), Browser-Erweiterungen und Open-Source-Modelle. Jeder Baustein agiert mit den Berechtigungen, die ihm die Anwendung oder der Browser gewährt – ein Einfallstor für Missbrauch. So wies Koi Security im Dezember 2025 nach, dass die Browser-Extension „Urban VPN Proxy“ (mitsamt Schwester-Add-ons und rund acht Millionen Nutzern) unbemerkt Chats mit ChatGPT, Claude, Gemini,

Microsoft Copilot & Co. mitschnitt und abgriff – obwohl sie im Chrome Web Store als „Featured“ ausgezeichnet war. Auch Modelldateien selbst sind vor Manipulation nicht sicher: So identifizierte JFrog 2024 etwa 100 Modelle auf Hugging Face, die beim Laden Schadcode ausführten.

Diese Risikoquelle ist deshalb relevant, weil solche Komponenten oft unbemerkt ins Unternehmen getragen werden – sei es durch die Browser-Erweiterung eines Angestellten oder den Konnektor eines Entwicklers. Zudem fordern Kunden in Sicherheitsfragebögen regelmäßig Nachweise darüber, ob Drittmodelle und -komponenten vor dem Einsatz auf Integrität geprüft wurden.

Mögliche Auswirkungen

- **Datenabfluss:** Eine kompromittierte Erweiterung liest Prompts, Antworten oder E-Mails im Klartext mit und leitet sie weiter – unabhängig von der Sicherheit des eigentlichen KI-Backends.
- **Schadcode im lokalen Netz:** Manipulierte Modelldateien können beim Laden Code ausführen und Angreifern einen ersten Zugang zum System verschaffen.
- **Supply-Chain-Gefahr durch Updates:** Ein anfangs vertrauenswürdiges Add-on kann durch ein schädliches Update nachträglich infiziert werden, wie der Fall postmark-mcp belegt.

Empfohlene Maßnahmen

- **Geschäftsführung:** In der KI-Richtlinie klar regeln, dass ausschließlich explizit freigegebene Add-ons, Plug-ins und Konnektoren eingesetzt werden dürfen (Allowlist-Prinzip).
- **IT-Leitung:** Browser-Erweiterungen und KI-Schnittstellen zentral steuern und auf verifizierte Herausgeber beschränken. Updates kritischer Bausteine vorab prüfen. Open-Source-Modelle ausschließlich in sicheren Formaten (wie Safetensors) beziehen und vor dem Produktiveinsatz scannen.
- **Fachbereiche:** Keine eigenmächtigen Installationen von Erweiterungen oder Konnektoren; Bedarf stattdessen über die IT verproben und anfordern lassen.

8. Rechtlicher Rahmen (Stand: Oktober 2026)

Die Kapitel 4 bis 7 nennen die wichtigsten Vorschriften dort, wo sie für eine Risikoquelle eine Rolle spielen. Dieses Kapitel fasst die rechtlichen Bezüge und die einschlägigen Rahmenwerke zusammen und ordnet sie in drei Gruppen: Recht, das heute gilt; Pflichten, die beschlossen sind, aber erst später greifen; und freiwillige Rahmenwerke.

Hinweis zum rechtlichen Teil

Die Ausführungen zu rechtlichen Rahmenbedingungen dienen der allgemeinen Information und geben die Einschätzung des Autors zum Zeitpunkt der Veröffentlichung wieder. Sie stellen keine Rechtsberatung dar und ersetzen keine Prüfung des konkreten Einzelfalls durch entsprechend qualifizierte Fachleute. Trotz sorgfältiger Recherche kann keine Gewähr für Vollständigkeit, Richtigkeit und Aktualität übernommen werden. Maßgeblich sind die jeweils geltenden Rechtsvorschriften und ihre Auslegung durch die zuständigen Gerichte und Behörden.

8.1 Geltendes Recht

Regel	Was sie für die KI-Nutzung bedeutet	Gilt seit	Kapitel
DSGVO	Auftragsverarbeitungsvertrag, wenn KI-Anbieter im Auftrag verarbeiten, sonst eigene Rechtsgrundlage für die	25.05.2018	4.1
Art. 4 Nr. 12	Weitergabe; Meldung von Datenpannen binnen 72 Stunden; grundsätzlich keine ausschließlich		5.1
Art. 22	automatisierten Entscheidungen mit erheblicher Wirkung, etwa über Bewerbende (mit Ausnahmen);		6.3
Art. 28	Bußgelder bei Verstößen		7.1
Art. 33			
Art. 83			
Geschäfts- geheimnisgesetz	Schutz als Geschäftsgeheimnis nur bei angemessenen Geheimhaltungsmaßnahmen – dazu gehören Regeln für die Eingabe in KI-Dienste und Zugriffsbeschränkungen	26.04.2019	4.1
§ 2 Nr. 1			5.1
			7.1
AGG	Entschädigung bei Benachteiligung im Bewerbungsverfahren; Beweislast beim Arbeitgeber, sobald Indizien vorliegen – auch wenn eine KI aussortiert hat	18.08.2006	6.3
§ 15			
§ 22			
BSIG (NIS-2- Umsetzung)	Für betroffene Unternehmen: Risikomanagement einschließlich Sicherheit der Lieferkette; die Geschäftsleitung muss die Maßnahmen billigen und überwachen, regelmäßig an Schulungen teilnehmen und haftet bei schuldhafter Pflichtverletzung gegenüber dem Unternehmen (§ 38)	06.12.2025	–
§ 30			
§ 38			
EU AI Act	Anbieter und Betreiber unterstützen die KI-Kompetenz ihres Personals (abgeschwächt durch den Digital Omnibus)	02.02.2025 geändert: 27.07.2026	–
Art. 4			
EU AI Act	Transparenzpflichten für Anbieter (etwa Hinweis auf die Interaktion mit einer KI) und für Betreiber (etwa Kennzeichnung von Deepfakes und bestimmter KI-generierter Texte zur Information der Öffentlichkeit)	02.08.2026	–
Art. 50			
EU Data Act	Wechsel von Cloud-Diensten (IaaS, PaaS, SaaS): Kunden können den Wechsel mit höchstens zwei Monaten Frist einleiten; für individuell entwickelte Dienste und Testversionen gelten eingeschränkte Ausnahmen. Wechselentgelte: siehe 8.2	12.09.2025	7.2
Anbieterwechsel			

Für wen die Regeln gelten. Ob ein Gesetz für ein Unternehmen gilt, hängt von dessen Tätigkeit ab: Die DSGVO gilt, sobald personenbezogene Daten verarbeitet werden, das AGG für jeden Arbeitgeber, das BSIG nur für Unternehmen bestimmter Sektoren und Größen (prüfbar mit der NIS-2-Betroffenheitsprüfung des BSI) und der AI Act je nach Rolle als Anbieter oder Betreiber und nach Einsatzzweck der KI.

Wie Gerichte die Regeln anwenden. Einzelne Fragen haben Gerichte bereits entschieden. Der EuGH entschied 2023 im Fall der SCHUFA: Schon eine automatisch berechnete Bewertung kann als eine grundsätzlich unzulässige automatisierte Entscheidung nach Art. 22 DSGVO gelten, wenn derjenige, der am Ende entscheidet, sich maßgeblich auf sie stützt (C-634/21). Das betrifft etwa KI-Scores in der Bewerberauswahl, auch wenn formal ein Mensch entscheidet (Kapitel 6.3). Das LG München I sah 2025 für wiedergegebene Liedtexte den KI-Anbieter in der Verantwortung (GEMA gegen OpenAI, nicht rechtskräftig; Kapitel 6.1).

8.2 Beschlossene Pflichten, die später greifen

Regel	Was hinzukommt	Ab	Kapitel
EU AI Act Art. 50 Abs. 2	Kennzeichnung KI-generierter Inhalte auch bei Systemen, die vor dem 02.08.2026 auf den Markt kamen (Ende der Übergangsfrist)	02.12.2026	–
EU Data Act Anbieterwechsel	Anbieter dürfen für den Wechsel kein Entgelt mehr verlangen; bis dahin sind nur ermäßigte Entgelte bis zur Höhe der Kosten des Anbieters zulässig	12.01.2027	7.2
EU AI Act Anhang III (u. a. Nr. 4 Buchst. a)	Hochrisiko-Systeme, u. a. KI zur Einstellung oder Auswahl von Personen. Pflichten für Betreiber, u. a. menschliche Aufsicht durch Personen mit der nötigen Kompetenz, Schulung und Befugnis (Art. 26 Abs. 2) und Information der Beschäftigten und ihrer Vertretung vor dem Einsatz (Art. 26 Abs. 7). Ausgenommen sind nur eng begrenzte Hilfsaufgaben; Systeme, die Profile von Personen erstellen, bleiben stets hochriskant (Art. 6 Abs. 3)	02.12.2027	6.3
EU AI Act Anhang I	Hochrisiko-Pflichten für KI als Sicherheitsbauteil regulierter Produkte, etwa Maschinen	02.08.2028	–

Der Digital Omnibus (Verordnung (EU) 2026/1744) hat nur den Geltungsbeginn der Hochrisiko-Pflichten des AI Act verschoben; aufgehoben hat er sie nicht. Andere Teile des AI Act gelten bereits, etwa Art. 4 und Art. 50 (siehe Abschnitt 8.1).

8.3 Freiwillige Rahmenwerke und Orientierungshilfen

Rahmenwerk	Herausgeber	Wofür es nützt
ISO/IEC 42001	ISO/IEC	Zertifizierbares Managementsystem für KI: Rollen, Richtlinien, Risikobewertung, Wirksamkeitsprüfung
ISO/IEC 23894	ISO/IEC	Leitfaden für das Risikomanagement von KI, mit Listen typischer Ziele und Risikoquellen
ISO/IEC 27001	ISO/IEC	Zertifizierbares Managementsystem für Informationssicherheit, u. a. Berechtigungen und Lieferanten
NIST AI 600-1	NIST (USA)	Risikoprofil für generative KI mit zwölf Risikokategorien
OWASP Top 10 for LLM Applications 2025	OWASP	Die zehn wichtigsten technischen Risiken von Sprachmodell-Anwendungen
Orientierungshilfe „KI und Datenschutz“ (Mai 2024)	Datenschutzkonferenz (DSK)	Sicht der deutschen Datenschutzaufsicht, u. a. zur Auftragsverarbeitung bei KI-Diensten
Cybersicherheitswarnung zu Indirect Prompt Injection (Juli 2023)	BSI	Einordnung der Schwachstelle als in der Technik angelegt

Diese Rahmenwerke sind nicht verbindlich. Freiwillig heißt aber nicht immer optional: In manchen Branchen setzen Kunden oder Aufsichtsbehörden einen Nachweis voraus – etwa TISAX, das aus ISO/IEC 27001 abgeleitet ist, in der Zulieferkette der Automobilindustrie oder eine auf ISO/IEC 27001 aufbauende Zertifizierung für Strom- und Gasnetzbetreiber. Sie zeigen aber, was Fachleute und Aufsichtsbehörden als gute Praxis ansehen, und können als Anhaltspunkt dienen, welche Maßnahmen „angemessen“ sind.

Was daraus folgt. Die meisten Pflichten, die bei der KI-Nutzung im Mittelstand greifen, gelten schon heute – über die DSGVO, das Geschäftsgeheimnisgesetz, das AGG und für NIS-2-betroffene Unternehmen über das BSIG. Der EU AI Act kommt mit seinen Hochrisiko-Pflichten erst Ende 2027 hinzu. Wer heute mit einem anerkannten Rahmenwerk beginnt, schafft eine gute Grundlage, die geltenden Pflichten zu erfüllen, und ist auf die kommenden vorbereitet.

9. Vorlage: Ihre erste KI-Risikoinventur

Die Vorlage hält für jede KI-Nutzung fest, welches Werkzeug wer mit welchen Daten nutzt, und prüft sie an den elf Risikoquellen. Jede KI-Nutzung im Unternehmen wird eingetragen – auch die nicht freigegebenen.

So gehen Sie vor

- Sammeln:** Die IT-Leitung fragt alle Bereiche, welche KI-Werkzeuge sie nutzen – ausdrücklich auch private und kostenlose. Ergänzend prüft sie vier Fundstellen:
 - **Firewall- oder Proxy-Protokolle:** Aufrufe bekannter KI-Dienste
 - **Microsoft-365-Admin-Center** (oder die Verwaltung vergleichbarer Plattformen): Copilot-Lizenzen und Apps, denen Beschäftigte Zugriff auf Unternehmensdaten erteilt haben
 - **Browser-Erweiterungen** auf den Firmengeräten
 - **Spesenabrechnungen und Firmenkreditkarten:** Abonnements für KI-Dienste
- Eintragen:** Für jede KI-Nutzung füllt die IT-Leitung gemeinsam mit dem jeweiligen Bereich eine Zeile aus. Die Prüffragen unten zeigen, welche Risikoquellen bei dieser KI-Nutzung vorliegen.
- Einschätzen und entscheiden:** Die IT-Leitung legt die Einträge mit dem höchsten Risiko der Geschäftsführung vor und empfiehlt Maßnahmen. Die Geschäftsführung entscheidet, welche Maßnahmen umgesetzt werden, und legt fest, wer sie umsetzt.

Die Einschätzung in dieser ersten Inventur ist bewusst einfach (hoch – mittel – gering). Eine vollständige Methode für die Risikobewertung folgt im nächsten Whitepaper dieser Reihe.

Für jede Risikoquelle nennen die Kapitel 4 bis 7 empfohlene Maßnahmen. Sie zeigen, dass sich jede Risikoquelle begrenzen lässt, und sind ein Ausgangspunkt für die Spalte „Nächster Schritt“. Welche Maßnahme passt, hängt vom einzelnen KI-Einsatz ab.

Die Tabelle

Die Risikoquellen wirken häufig zusammen. Ein nicht freigegebener KI-Dienst kann zugleich Shadow AI und Data Leakage bedeuten – wie im Beispiel in der ersten Zeile der Tabelle. Liest ein KI-Agent externe Dokumente und darf anschließend Bestellungen auslösen, verstärken sich Indirect Prompt Injection und Excessive Agency gegenseitig: Erst die Handlungsrechte machen aus einer versteckten Anweisung einen tatsächlichen Schaden. Die Inventur erfasst deshalb bewusst mehrere Risikoquellen pro KI-Einsatz.

Die Vorlage gibt es auch als Excel-Datei mit Auswahllisten und den Prüffragen: <https://carson.de/ressourcen/ki-risikoinventur-vorlage.xlsx>

Nr.	Werkzeug	Wer nutzt es?	Welche Daten?	Festgestellte Risikoquellen	Mögliche Schäden	Einschätzung	Verantwortlich	Nächster Schritt
1	Kostenloser KI-Chat (Einzeltarif, nicht freigegeben, kein AVV)	Vertrieb, drei Mitarbeitende	Angebote mit nichtanonymisierten Kundennamen und Konditionen	4.1 Data Leakage, 6.1 Fehlerhafte Ergebnisse, 6.2 Blindes Vertrauen, 7.1 Shadow AI	Weitergabe ohne AVV und ohne erkennbare Rechtsgrundlage (DSGVO), ggf. Meldepflicht; gefährdeter Schutz der Konditionen als Geschäftsgeheimnis; ungeprüfte, falsche Angaben in Angeboten an Kunden	hoch	IT-Leitung (bewertet, empfiehlt), Geschäftsführung (entscheidet)	Freigegebenen KI-Dienst mit Unternehmensvertrag und AVV bereitstellen; bis dahin keine Kundendaten eingeben
2								
3								

Prüffragen zu den elf Risikoquellen

Zu jeder Risikoquelle gehört eine konkrete Prüffrage. Gehen Sie alle diese Fragen für jeden KI-Einsatz durch: Lautet die Antwort „Ja“, haben Sie die jeweilige Risikoquelle festgestellt: Tragen Sie sie in Ihr KI-Inventar ein.

Ein „Ja“ ist dabei kein Urteil über Fehler im Betrieb, sondern ein neutraler Hinweis auf Relevanz. Wie hoch das tatsächliche Risiko ist und welcher Handlungsbedarf besteht, halten Sie anschließend in den Spalten „Mögliche Schäden“ und „Einschätzung“ fest.

Risikoquelle	Prüffrage
4.1 Data Leakage (Input)	Werden personenbezogene, vertrauliche oder geschäftskritische Daten an einen externen KI-Dienst übermittelt?
4.2 Indirect Prompt Injection (Input)	Verarbeitet die KI Inhalte aus externen oder ungeprüften Quellen – etwa Bewerbungen, E-Mails, Angebote oder Webseiten?
4.3 Mangelhafte oder manipulierte Daten (Input)	Greift die KI auf eigene Datenbestände zurück – etwa Ablagen, Wissensdatenbanken oder Produktdaten –, aus denen sie ihre Ergebnisse ableitet?
5.1 Oversharing (Access & Execution)	Greift die KI auf interne Ablagen, Postfächer oder Datenbanken zu, deren Inhalte nicht für alle Nutzer bestimmt sind?
5.2 Excessive Agency (Access & Execution)	Kann die KI selbst Handlungen auslösen – etwa senden, ändern, bestellen oder löschen?
6.1 Fehlerhafte Ergebnisse (einschließlich Halluzinationen) (Output)	Werden KI-Ergebnisse für fachlich, finanziell oder rechtlich relevante Zwecke verwendet, bei denen falsche Angaben Schaden verursachen können?
6.2 Blindes Vertrauen (Automation Bias) (Output)	Stützen Beschäftigte Arbeitsergebnisse oder Entscheidungen auf KI-Ergebnisse, ohne sie selbst fachlich nachzuvollziehen?
6.3 Diskriminierung durch KI (Output)	Bewertet, sortiert oder filtert die KI Menschen – etwa Bewerbende, Beschäftigte oder Kunden?
7.1 Shadow AI (Ecosystem)	Wird das KI-Werkzeug ohne Freigabe oder ohne Wissen der IT genutzt?
7.2 Vendor Lock-in (Ecosystem)	Hängt ein wichtiger Ablauf von einem einzelnen KI-Dienst oder Modell ab?
7.3 Manipulierte KI-Komponenten von Dritten (Ecosystem)	Wird die KI durch Erweiterungen, Plug-ins, Konnektoren (etwa MCP-Server) oder Modelle von Drittanbietern ergänzt?

Wofür das KI-Inventar außerdem nützt

Die ausgefüllte Tabelle ist zugleich Ihr KI-Inventar (siehe 7.1); prüfen Sie es mindestens einmal im Jahr und bei jedem neuen KI-Werkzeug. Es hilft auch bei Pflichten, die unabhängig von diesem Leitfaden bestehen. Zwei Beispiele – die Liste ist nicht abschließend:

- **Verzeichnis von Verarbeitungstätigkeiten (Art. 30 DSGVO):** Verarbeitet eine KI-Nutzung personenbezogene Daten, gehört sie in der Regel in dieses Verzeichnis. Die Spalten „Werkzeug“, „Wer nutzt es?“ und „Welche Daten?“ liefern dafür erste Angaben.
- **Hochrisiko-KI nach dem EU AI Act:** Wer KI etwa zur Auswahl von Bewerbenden einsetzt, betreibt möglicherweise ein Hochrisiko-System nach Anhang III. Ob das zutrifft, erfordert eine eigene Prüfung im Einzelfall, die über diesen Leitfaden hinausgeht. Das Inventar zeigt aber, bei welchen KI-Nutzungen sich diese Frage überhaupt stellt.

10. Quellen

Alle Quellen wurden im Oktober 2026 abgerufen.

Studien und Umfragen

- Bitkom: „Erstmals nutzt die Mehrheit der Unternehmen KI“, Presseinformation, 14.09.2026. <https://www.bitkom.org/Presse/Presseinformation/Erstmals-nutzt-Mehrheit-Unternehmen-KI>
- Bitkom: „Beschäftigte nutzen vermehrt Schatten-KI“, Presseinformation, Oktober 2025. <https://www.bitkom.org/Presse/Presseinformation/Beschaeftigte-nutzen-Schatten-KI>
- IBM: Cost of a Data Breach Report 2025, Pressemitteilung, 30.07.2025. <https://newsroom.ibm.com/2025-07-30-ibm-report-13-of-organizations-reported-breaches-of-ai-models-or-applications,-97-of-which-reported-lacking-proper-ai-access-controls>
- Microsoft und LinkedIn: 2024 Work Trend Index, 08.05.2024. <https://news.microsoft.com/source/2024/05/08/microsoft-and-linkedin-release-the-2024-work-trend-index-on-the-state-of-ai-at-work/>
- Harmonic Security, berichtet von CSO Online: „Nearly 10% of employee gen AI prompts include sensitive data“. <https://www.csoonline.com/article/3819170/nearly-10-of-employee-gen-ai-prompts-include-sensitive-data.html>
- Gartner-Umfrage zu Microsoft 365 Copilot, berichtet von Computerworld, Juni 2024. <https://www.computerworld.com/article/3542000/microsoft-365-copilot-rollouts-slowed-by-data-security-roi-concerns.html>
- Gartner: „Gartner Predicts 40% of Enterprise Apps Will Feature Task-Specific AI Agents by 2026“, Pressemitteilung, 26.08.2025. <https://www.gartner.com/en/newsroom/press-releases/2025-08-26-gartner-predicts-40-percent-of-enterprise-apps-will-feature-task-specific-ai-agents-by-2026-up-from-less-than-5-percent-in-2025>
- Varonis: State of Data Security Report 2025 (Herstellerstudie). <https://www.varonis.com/blog/state-of-data-security-report>
- Harvard Business School und BCG, berichtet von Forbes, 29.09.2023. <https://www.forbes.com/sites/danpontefract/2023/09/29/harvard-and-bcg-unveil-the-double-edged-sword-of-ai-in-the-workplace/>
- Microsoft Research und Carnegie Mellon University, berichtet von Campus Technology, 21.02.2025. <https://campustechnology.com/articles/2025/02/21/study-generative-ai-could-inhibit-critical-thinking.aspx>
- Duke University: „Tricking AI in the Job Hunt“, 10.08.2026. <https://today.duke.edu/2026/08/tricking-ai-job-hunt>
- Google: Auswertung von Prompt-Injection-Inhalten im Web, 23.04.2026. <https://blog.google/security/prompt-injections-web/>
- Unit 42 (Palo Alto Networks): Prompt Injection gegen KI-Agenten, 03.03.2026. <https://unit42.paloaltonetworks.com/ai-agent-prompt-injection/>
- Zou et al.: „PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models“, USENIX Security 2025. <https://www.usenix.org/conference/usenixsecurity25/presentation/zou>
- Anthropic, UK AI Security Institute und Alan Turing Institute: „A small number of samples can poison LLMs of any size“, 09.10.2025. <https://www.anthropic.com/research/small-samples-poison>

Fälle

- Moffatt v. Air Canada, 14.02.2024, berichtet von der American Bar Association. https://www.americanbar.org/groups/business_law/resources/business-law-today/2024-february/bc-tribunal-confirms-companies-remain-liable-information-provided-ai-chatbot/
- AG Köln, Beschluss vom 02.07.2025, Az. 312 F 130/25, berichtet von LTO. <https://www.lto.de/recht/juristen/b/ag-koeln-familiengericht-312f130-25-schriftsatz-ki-anwalt-berufspflichten>
- LG München I, Urteil vom 11.11.2025, Az. 42 O 14139/24 (GEMA gegen OpenAI), Pressemitteilung. <https://www.justiz.bayern.de/gerichte-und-behoerden/landgericht/muenchen-1/presse/2025/11.php>
- Deloitte Australien, berichtet von The Register, 06.10.2025. https://www.theregister.com/2025/10/06/deloitte_ai_report_australia/
- Samsung, berichtet von The Register, 06.04.2023. https://www.theregister.com/2023/04/06/samsung_reportedly_leaked_its_own
- Samsung, berichtet von TechRadar, Mai 2023. <https://www.techradar.com/news/samsung-bans-chatgpt-use-after-employee-leak>
- EchoLeak (CVE-2025-32711), berichtet von SOC Prime. <https://socprime.com/blog/cve-2025-32711-zero-click-ai-vulnerability/>
- Replit, Interview mit Amjad Masad, Fast Company, Juli 2025. <https://www.fastcompany.com/91372483/replit-ceo-what-really-happened-when-ai-agent-wiped-jason-lemkins-database-exclusive>
- Amazon, berichtet von MIT Technology Review, 10.10.2018. <https://www.technologyreview.com/2018/10/10/139858/amazon-ditched-ai-recruitment-software-because-it-was-biased-against-women/>

- Mobley v. Workday, berichtet von Davis Wright Tremaine, Mai 2025. <https://www.dwt.com/blogs/employment-labor-and-benefits/2025/05/ai-hiring-age-discrimination-federal-court-workday>
- OpenAI: „June 10th Service Disruption FAQ". <https://help.openai.com/en/articles/11565131-june-10th-service-disruption-faq>
- postmark-mcp (Koi Security), berichtet von The Hacker News, 29.09.2025. <https://thehackernews.com/2025/09/first-malicious-mcp-server-found.html>
- Urban VPN Proxy (Koi Security), berichtet von Dark Reading, 16.12.2025. <https://www.darkreading.com/endpoint-security/chrome-extension-harvests-ai-chatbot-data>
- Schadhafte Modelle auf Hugging Face (JFrog), berichtet von BleepingComputer, 28.02.2024. <https://www.bleepingcomputer.com/news/security/malicious-ai-models-on-hugging-face-backdoor-users-machines/>
- GPT-4o, berichtet von VentureBeat. <https://venturebeat.com/ai/openai-is-ending-api-access-to-fan-favorite-gpt-4o-model-in-february-2026>

Investitionen in Schutzlösungen gegen Prompt Injection

- Cisco / Robust Intelligence, SiliconANGLE, 27.08.2024. <https://siliconangle.com/2024/08/27/cisco-snaps-ai-model-data-security-startup-robust-intelligence/>
- Check Point / Lakera, SecurityWeek, 16.09.2025. <https://www.securityweek.com/check-point-to-acquire-ai-security-firm-lakera/>
- F5 / CalypsoAI, GeekWire, 11.09.2025. <https://www.geekwire.com/2025/f5-paying-180m-to-acquire-calypsoai-to-boost-ai-enterprise-security-offerings/>
- HiddenLayer, TechCrunch, 02.09.2026. <https://techcrunch.com/2026/09/02/hiddenlayer-nabs-100m-as-enterprises-rush-to-secure-their-ai-deployments/>

Anbieter und Behörden

- OpenAI: „Why language models hallucinate", 05.09.2025. <https://openai.com/index/why-language-models-hallucinate/>
- OpenAI Help Center: „How your data is used to improve model performance". <https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance>
- Microsoft Learn: „Data, Privacy, and Security for Microsoft 365 Copilot". <https://learn.microsoft.com/en-us/microsoft-365/copilot/microsoft-365-copilot-privacy>
- Microsoft Learn: „Restrict discovery of SharePoint sites and content" (Restricted Content Discovery). <https://learn.microsoft.com/en-us/sharepoint/restricted-content-discovery>
- Microsoft Learn: „Prompt Shields in Azure AI Content Safety". <https://learn.microsoft.com/en-us/azure/ai-services/content-safety/concepts/jailbreak-detection>
- BSI: Cybersicherheitswarnung „Indirect Prompt Injections – Intrinsische Schwachstelle in anwendungsintegrierten KI-Sprachmodellen", 15.07.2023. https://www.bsi.bund.de/SharedDocs/Cybersicherheitswarnungen/DE/2023/2023-249034-1032.pdf?__blob=publicationFile&v=3
- BSI: NIS-2-Betroffenheitsprüfung. https://www.bsi.bund.de/DE/Themen/Regulierte-Wirtschaft/NIS-2-regulierte-Unternehmen/BSI-Portal/NIS-2-Betroffenheitspruefung/nis-2-betroffenheitspruefung_node.html

Gesetze und Rechtsquellen

- DSGVO Art. 22: <https://datenschutz.law/dsgvo/3-kapitel/art-22> ; Erwägungsgrund 71: https://www.dsgvo-portal.de/dsgvo_erwaegungsgrund_71.php ; Art. 83: <https://dejure.org/gesetze/DSGVO/83.html>
- DSGVO Art. 4 Nr. 12, Erläuterung von Thomas Helbing. <https://www.thomashelbing.com/de/wissen/dsgvo-hub/begriffe-und-definitionen/1.2.12-verletzung-des-schutzes-personenbezogener-daten>
- AGG § 15: <https://dejure.org/gesetze/AGG/15.html>
- BAG, Urteil vom 11.08.2016, 8 AZR 406/14 („junges dynamisches Team"). <https://www.bundesarbeitsgericht.de/entscheidung/8-azr-406-14/>
- Geschäftsgeheimnisgesetz, Überblick der Kanzlei Plutte. <https://www.ra-plutte.de/grosser-rechtsueberblick-zum-neuen-geschaeftsgeheimnisgesetz/>
- EU AI Act Art. 14, activeMind.legal. <https://www.activemind.legal/legislation/ai-act/article-14/>
- EU AI Act Art. 6, activeMind.legal. <https://www.activemind.legal/legislation/ai-act/article-6/>
- EU AI Act Anhang III, AI Act Service Desk der Europäischen Kommission. <https://ai-act-service-desk.ec.europa.eu/en/ai-act/annex-3>

- EuGH, Urteil vom 07.12.2023, C-634/21 (SCHUFA), Erläuterung von Thomas Helbing.
<https://www.thomashelbing.com/de/wissen/dsgvo-hub/rechtsprechung/1.4.32-eugh-schufa-scoring>
- Digital Omnibus (Verordnung (EU) 2026/1744, ABl. L vom 24.07.2026, <http://data.europa.eu/eli/reg/2026/1744/oj>), White & Case.
<https://www.whitecase.com/insight-alert/eu-ai-omnibus-enters-force-amending-ai-act>
- EU Data Act, Deloitte Legal. <https://www.deloittelegal.de/dl/en/services/legal/perspectives/cloud-switching-eu-data-act.html>
- DSGVO Art. 28: <https://dejure.org/gesetze/DSGVO/28.html> ; Art. 33: <https://dejure.org/gesetze/DSGVO/33.html> ; Art. 99 (Geltung ab 25.05.2018): <https://dsgvo-gesetz.de/art-99-dsgvo/>
- AGG § 22: <https://dejure.org/gesetze/AGG/22.html> ; Inkrafttreten 18.08.2006, Antidiskriminierungsstelle des Bundes:
<https://www.antidiskriminierungsstelle.gov.de/DE/ueber-diskriminierung/recht-und-gesetz/allgemeines-gleichbehandlungsgesetz/allgemeines-gleichbehandlungsgesetz-node.html>
- BStG § 30: https://www.gesetze-im-internet.de/bstg_2025/___30.html
- Geschäftsgeheimnisgesetz, Inkrafttreten 26.04.2019, Gleiss Lutz: <https://www.gleisslutz.com/de/know-how/geschaeftsgeheimnisgesetz-ist-kraft-getreten>
- EU AI Act Art. 26, springlex: <https://www.springlex.eu/en/packages/ai-act/ai-act-regulation/article-26/>

Rahmenwerke

- OWASP Top 10 for LLM Applications 2025. <https://genai.owasp.org/>
- NIST AI 600-1: Artificial Intelligence Risk Management Framework – Generative Artificial Intelligence Profile, Juli 2024.
<https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf>
- Panickssery et al.: „LLM Evaluators Recognize and Favor Their Own Generations“, NeurIPS 2024. <https://arxiv.org/pdf/2404.13076>
- ISO/IEC 23894:2023, Information technology – Artificial intelligence – Guidance on risk management (Struktur der Anhänge A und B, erläutert von Hernan Huwyler). <https://hernanhuwyler.wordpress.com/2026/03/28/how-to-actually-use-iso-iec-23894-for-ai-risk-management/>
- Bundesnetzagentur: IT-Sicherheitskatalog für Strom- und Gasnetze (§ 11 Abs. 1a EnWG).
https://bundesnetzagentur.de/DE/Fachthemen/ElektrizitaetundGas/Versorgungssicherheit/IT_Sicherheit/Netzbetreiber/start.html
- TISAX, erläutert von DQS. <https://www.dqsglobal.com/de/entdecken/blog/tisax-antworten-auf-wichtige-fragen>

Über den Autor

Derek Carson unterstützt mittelständische Unternehmen bei Fragen zu KI-Governance, Risikomanagement und Informationssicherheit. Er hat die Lead-Implementer-Prüfungen für ISO/IEC 42001 und ISO/IEC 27001 bestanden und ist ISC2 Certified in Cybersecurity.

Seine berufliche Erfahrung umfasst rund 20 Jahre Training und Beratung für Konzerne, Hochschulen und Medienhäuser sowie Projekte zur Einführung von Content-Management-Systemen und Berechtigungskonzepten.

Eine erste KI-Risikoinventur angehen

In einem klar abgegrenzten Pilotprojekt unterstütze ich Sie und Ihre IT-Leitung dabei, KI-Anwendungen zu erfassen, anhand der elf Risikoquellen dieses Leitfadens zu prüfen und priorisierte nächste Schritte zu dokumentieren. Das Pilotprojekt umfasst keine Rechtsberatung.

Derek Carson · derek@carson.de · carson.de