

# Managing AI Risk in Your Enterprise

The Key Sources of Risk When Using AI—and the Measures That Limit Them

**Derek Carson**

AI Governance & Information Security  
carson.de

As of October 2026 · 2nd edition

# Contents

---

1. Executive Summary
2. Purpose and Scope
3. Why AI Risks Are Not Ordinary IT Risks
4. Input: Risks in Data Entry
  - 4.1 Data Leakage
  - 4.2 Indirect Prompt Injection
  - 4.3 Flawed or Manipulated Data
5. Access & Execution: Risks in Access and Execution
  - 5.1 Oversharing
  - 5.2 Excessive Agency
6. Output: Risks in the Results
  - 6.1 Inaccurate Outputs (Including Hallucinations)
  - 6.2 Automation Bias
  - 6.3 AI Discrimination
7. Ecosystem: Risks in the Environment of AI Use
  - 7.1 Shadow AI
  - 7.2 Vendor Lock-in
  - 7.3 Manipulated Third-Party AI Components
8. Legal Framework (as of October 2026)
9. Template: Your First AI Risk Inventory
10. Sources

About the Author

# 1. Executive Summary

---

Over half of German enterprises (57%; Bitkom 2026) actively utilize artificial intelligence. Yet a Bitkom survey from the previous year shows that very few had the topic under strategic control: only 26 percent provided their employees with officially approved AI tools in 2025. Just 23 percent had binding rules for its use. And in more than four out of ten companies, those responsible knew or suspected that employees were using private AI tools in their day-to-day work (Bitkom 2025).

Anyone who does not know which tools are used and what data is entered can neither assess nor manage the operational risks.

Real cases of damage show that these risks are not just theory:

- **Air Canada:** In 2024, the airline was held responsible for incorrect information given by a customer service chatbot; the competent Canadian tribunal (Civil Resolution Tribunal, British Columbia) rejected the argument that the chatbot was a separate entity.
- **Cologne Local Court (Amtsgericht Köln):** In 2025, a lawyer filed a brief containing invented, apparently AI-generated rulings and citations. The court found this and instructed him to refrain from doing so in the future (Ref. 312 F 130/25).
- **IBM study:** One in five of the companies examined by IBM reported a data breach caused by unapproved AI. Companies with a lot of unapproved AI paid on average USD 670,000 more per data breach than companies with little or none (IBM 2025).

The weak point rarely lies in the AI technology itself, but in the company's organization: in access rights that are too broad, unchecked work results and missing guardrails. Important for decision-makers: the legal risks already exist today—through the GDPR, the German Trade Secrets Act, the German General Equal Treatment Act and, for companies covered by NIS-2, the German BSI Act (BSIG, Sections 30 and 38).

The first step toward managing these risks is therefore a systematic inventory of the AI applications actually in use in your company. This whitepaper provides the framework for it: eleven key risk sources that can arise when using AI, assigned to four areas—Input, Access & Execution, Output and Ecosystem. The first three areas follow the typical sequence of an AI use—from the information entered, through the AI's access and actions, to its results. The Ecosystem forms the environment of technical, organizational and external dependencies. With this framework, different uses of AI can be examined for possible risk sources.

A practical example: three sales employees compare customer quotes using a free AI chatbot. Non-anonymized customer names and terms end up on the AI provider's servers—without a data processing agreement (Art. 28 GDPR) and with a service that may also use inputs for its own purposes. No legal basis for this disclosure is apparent; a data protection breach is therefore likely. Whether there is also a duty to notify the supervisory authority (Art. 33 GDPR) must be examined immediately by management and the data protection officer. Such a problem often only becomes visible through this inventory.

With this whitepaper, I support you in carrying out this inventory. I guide you through the eleven key risk sources in your company. At the end, you will find a template you can use straight away for your first AI risk inventory.

## 2. Purpose and Scope

---

### Target readers and focus

This whitepaper is aimed at management and heads of IT in mid-sized companies that use AI applications in their operations—but do not develop them. The focus is on standard tools such as Microsoft 365 Copilot, commercial AI assistants and AI functions in existing business software. If you include internal AI assistants that first search company documents and then answer on that basis (so-called RAG systems), this paper covers an important part of everyday AI use in a company.

### What is not covered

To keep the focus on everyday operational use, I deliberately leave out three topics:

- **In-house development and attacks:** Training your own AI models and external attack scenarios (such as deepfakes or AI phishing) follow their own risk profiles, because they do not arise from your own use of AI.
- **Individual case assessments:** Formal conformity assessments for high-risk AI under the EU AI Act and individual legal advice require an examination of the specific case.
- **Methodology and implementation:** How to embed risk management processes in the company on a lasting basis, and which detailed protective measures belong to them, I cover in depth in the two following whitepapers in this series.

### How I selected the risk sources

I selected the eleven risk sources according to three criteria: they affect companies that use AI, not those that develop it. For each, there are documented cases or reliable studies. And each has specific operational or legal consequences for a company in Germany.

For the new edition, I cross-checked the framework against some 790 AI questions from 22 questionnaires, customer requirements, and regulatory specifications; this added two risk sources that customers specifically ask about.

### 3. Why AI Risks Are Not Ordinary IT Risks

Traditional IT security is already well-versed in conventional vulnerabilities like folder shares that are too broad or careless use of external web services. Artificial intelligence, however, fundamentally changes these risk dynamics. Four factors are responsible:

**1. AI delivers the most likely answer, not the verified one**

A language model calculates statistically probable sequences of words—it does not check facts (OpenAI 2025). As a result, it sometimes produces invented information (hallucinations) or reproduces biased patterns from past data. Because the generated texts sound convincing, they easily tempt employees to take results into business operations without checking them.

**2. AI does not reliably separate content from instructions**

For a language model, input text and control commands are technically the same thing. When the AI processes an external document or an email, text contained in it can be wrongly interpreted by the AI as an instruction (indirect prompt injection). The BSI (Germany's Federal Office for Information Security) describes this problem as an intrinsic weakness of today's language models (BSI 2023).

**3. AI uses every permission it is given**

An assistant such as Microsoft 365 Copilot can in principle find all documents that the respective user is allowed to open. A share in the company network that grew too broad over the years and that nobody has noticed is exposed within seconds by a simple chat question. Autonomous AI agents make this even worse: they do not just read data, but independently carry out actions based on it.

**4. AI escapes the control of your own IT team**

An ordinary browser is enough to use an AI chat. Employees can therefore integrate AI tools into their daily work without the approval or knowledge of the head of IT. And as soon as a business process depends on an AI service, model changes or outages at the provider directly affect that process—without your own IT having any influence over it.

**What follows from this**

Factors 1 and 2 are properties of the technology; a company that uses AI cannot change them. It can, however, decide how to deal with them: whether results are checked and which documents the AI may process. Factors 3 and 4, by contrast, lie within the company itself: in the shares, in the way employees use AI, and in the choice and safeguarding of providers. It follows that you manage the risks of using AI through your own organization—through permissions, the checking of results and rules for use.

**The four areas and eleven risk sources**

The figure and table summarize the framework from chapter 1: eleven risk sources in four areas. I deliberately arrange the risk sources according to the sequence of an AI use and not by technology or area of law: this way, management can see at each step where it can intervene.



| Risk source  | Area  | What it is about                                                                            |
|--------------|-------|---------------------------------------------------------------------------------------------|
| Data Leakage | Input | Employees enter personal, confidential or business-critical data into external AI services. |

| Risk source                                   | Area               | What it is about                                                                                                                                      |
|-----------------------------------------------|--------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------|
| Indirect Prompt Injection                     | Input              | Job applications, quotes, emails or websites contain hidden instructions that influence the AI's judgment as soon as it reads the document.           |
| Flawed or Manipulated Data                    | Input              | The file stores and knowledge bases the AI answers from are flawed, outdated, or deliberately manipulated.                                            |
| Oversharing                                   | Access & Execution | AI assistants find and display internal documents that, because of shares that are too broad, are accessible to users for whom they are not intended. |
| Excessive Agency                              | Access & Execution | AI agents may independently send emails, change data or place orders—including on manipulated instructions.                                           |
| Inaccurate Outputs (Including Hallucinations) | Output             | The AI delivers incorrect results—most conspicuously, plausible-sounding but invented information.                                                    |
| Automation Bias                               | Output             | Employees adopt AI results without checking them because they sound convincing and save time.                                                         |
| AI Discrimination                             | Output             | AI systems that evaluate people adopt the patterns of earlier decisions—including the unfair ones.                                                    |
| Shadow AI                                     | Ecosystem          | Employees use AI tools that IT neither knows about nor has checked.                                                                                   |
| Vendor Lock-in                                | Ecosystem          | Processes depend on a single AI service; outages or model changes at the provider hit them directly.                                                  |
| Manipulated Third-Party AI Components         | Ecosystem          | Extensions, connectors, or models from third-party providers contain malicious functions.                                                             |

The risk sources are not mutually exclusive but often act together: a single use of AI can involve several of them at once.

Chapters 4 to 6 follow the three areas of the typical sequence of an AI use, starting with what goes in. Chapter 7 covers the fourth area, the Ecosystem, which forms the environment of these three.

To make it easier to compare with international standards and frameworks, I use the established English technical terms for most risk sources. The template in chapter 9 applies the same framework; with it, the head of IT can start the first AI risk inventory straight away.

## 4. Input: Risks in Data Entry

---

This group concerns all data and content that an AI processes—whether through active input by your employees (4.1 Data Leakage) through external documents in which third-party control commands are hidden (4.2 Indirect Prompt Injection), or through data sets that are flawed, outdated, or manipulated (4.3 Flawed or Manipulated Data).

### 4.1 Data Leakage

#### Key points for management

When Samsung banned its employees from using ChatGPT in 2023, the damage had already been done: within less than three weeks, employees had entered source code from the semiconductor division and internal meeting minutes into the service in at least three cases. Samsung itself stated that it was hardly possible to determine what had been disclosed, where it was stored and how it could be deleted. The group then banned generative AI on all company devices and threatened dismissal for violations—thereby losing a tool it had approved only a few weeks earlier.

A 2024 study by Harmonic Security shows that this is not an isolated case. The study analyzed tens of thousands of employee inputs (prompts): 8.5 percent contained sensitive data—almost half of them customer data.

A typical case: an HR officer has a free AI chat revise the style of a formal warning letter. The letter contains the real name of the employee concerned, the alleged misconduct and the threat of dismissal. As soon as it is sent, this data is on the provider's servers.

#### Possible impact

- **Further use by the provider:** According to OpenAI (as of October 2026), the provider may by default use inputs in its services for individuals—such as the free version of ChatGPT and the Plus and Pro plans—to train its models, unless the user switches this off; in the Business and Enterprise plans, this does not happen by default. Such settings change—the provider's current terms are decisive.
- **Data protection breach:** Anyone who enters personal data into an external AI service needs a sound legal basis under data protection law. If the provider processes the data on behalf of the company—which, according to Germany's Data Protection Conference (DSK), is the rule for AI services used as software—a data processing agreement under Art. 28 GDPR is required. If the provider also uses the inputs for its own purposes, such as training, the disclosure itself requires a separate legal basis, which the company must examine and justify separately. If the input constitutes an unauthorized disclosure and therefore a personal data breach, it must generally be reported to the supervisory authority without undue delay and, where feasible, within 72 hours—unless it is unlikely to result in a risk to the individuals concerned (Art. 33 GDPR).
- **Trade secret protection at risk:** A trade secret enjoys legal protection only if the company demonstrably takes "appropriate secrecy measures" (Section 2 No. 1 of the German Trade Secrets Act, GeschGehG). Anyone who tolerates the uncontrolled input of confidential data into AI services risks this protected status—and with it the special claims the GeschGehG grants in the event of unauthorized disclosure.

#### Recommended measures

- **Management:** Provide binding business licenses with a suitable business contract—with a data processing agreement and a guaranteed exclusion of model training on all inputs.
- **Head of IT:** Prepare a clear policy that specifies which types of data—such as customer data, personnel data or contract terms—may be entered into which approved AI services. Management puts the policy into force.
- **HR and legal:** Train employees in a practical way and clearly mark confidential documents as confidential internally.

## 4.2 Indirect Prompt Injection

### Key points for management

To date, documented exploits involving indirect prompt injection have been rare: when analyzing billions of web pages, Google found only "minimal" cases of harm, and Palo Alto Networks reported no confirmed successful attack in practice up to 2026. But the threat is building: between November 2025 and February 2026 alone, Google recorded a 32 percent increase in manipulated content. This risk becomes acute for your company the moment AI systems automatically read and evaluate third-party documents or carry out actions based on them.

Because an AI cannot reliably distinguish between payload data (content) and control commands (instructions), third-party documents—such as job applications, quotes, emails or websites—can contain hidden instructions that steer the AI's behavior unnoticed. In 2026, researchers at Duke University showed that at least one percent of 200,000 job applications analyzed contained manipulated passages of text (for example in microscopic font) designed to trick AI pre-screening systems. The "EchoLeak" vulnerability (2025) in Microsoft 365 Copilot showed what such instructions can do: a single manipulated email would have been enough to get Copilot to hand over internal company data in the background—without the recipient having to click a link or open a file. Security researchers discovered the flaw and Microsoft fixed it; no exploitation by attackers is known.

I have included indirect prompt injection despite the small damage so far, because the risk grows the moment AI agents are allowed to read third-party documents and act on them—and that is exactly where the market is heading.

### Possible impact

- **Distorted decisions:** An AI system that pre-sorts job applications or supplier quotes can be manipulated by hidden commands in the attacker's favor.
- **Unintended data leakage:** As the EchoLeak example demonstrates, confidential information can leave the company without any employee having made a mistake.
- **Unwanted actions:** If the AI system has its own rights to act (e.g. sending emails or creating calendar entries), a hidden instruction can carry out these actions directly (see 5.2 Excessive Agency).

### Recommended measures

- **Head of IT:** Strictly define which data sources an AI assistant may read, and limit automated access to unverified external content. For your own AI applications, also use filters against known attack patterns (such as Microsoft Prompt Shields); for purchased services, ask the provider whether such filters are active. Filters detect known patterns, but not every attack.
- **Management and head of IT:** No operational action by the AI without final approval by a human (human-in-the-loop).
- **HR:** Have every AI-generated ranking or pre-selection of job applications checked by qualified staff before any decision.

## 4.3 Flawed or Manipulated Data

### Key points for management

An AI is only as reliable as the data foundation it draws on. In 2025, researchers showed that just five crafted texts in a knowledge base with millions of entries were enough to steer an AI assistant's answer to a specific question in around 90 percent of cases (USENIX Security 2025). Then there is the obvious: outdated price lists, superseded policies, and faulty spreadsheets that the AI draws from.

Many companies connect AI assistants directly to their own data sources—such as Microsoft 365 Copilot to SharePoint or an internal assistant to the customer service manual (RAG systems, see Chapter 2). The AI does not check whether a document is current, correct, or authentic—it simply takes what it finds. If, say, three different versions of a price list sit on the drive, the AI may well pick the outdated one. In addition, anyone with write access to these folders can deliberately manipulate the AI's answers.

Models can also be deliberately manipulated during training: According to a study by Anthropic, the UK AI Security Institute, and the Alan Turing Institute (October 2025), around 250 manipulated documents were enough to implant a hidden backdoor in language models of various sizes. The researchers do stress that this test case poses little risk for today's top models. For companies that use commercial AI, however, the real everyday risk lies directly in their own data sets.

I added this risk source because clients ask about it specifically in vendor and due diligence questionnaires: whether the data foundation is validated before use, whether its origin is traceable, and whether it is protected against manipulation.

#### **Possible impact**

- **Incorrect information to outsiders:** A support assistant quotes outdated prices, delivery times, or warranty terms—for which the company is liable (see 6.1).
- **Wrong decisions internally:** Employees rely in their daily work on answers based on outdated policies or faulty spreadsheets.
- **Targeted manipulation:** Anyone who plants malicious content in a connected data source can steer the AI's output—for example on product data, suppliers, or payment details.

#### **Recommended measures**

- **Management:** Define clear responsibility for the accuracy and currency of every connected data source.
- **Head of IT:** Specify which file stores and sources an AI may index. Keep write access to these folders restrictive and version all changes. Check external content before feeding it into knowledge bases.
- **Departments:** Consistently clean outdated versions out of file stores or archive them; validate business-critical content such as price lists, product data, and policies regularly.

## 5. Access & Execution: Risks in Access and Execution

---

This category encompasses an AI's permissions to read data and execute actions within the company: either because folder shares that are too broad make internal documents accessible to it (5.1 Oversharing), or because it is allowed to carry out actions independently as an autonomous agent (5.2 Excessive Agency).

### 5.1 Oversharing

#### Key points for management

When Microsoft 365 Copilot arrived in companies, many IT departments immediately put on the brakes: according to a Gartner survey of 132 IT leaders in June 2024, 40 percent postponed the roll-out by three months or more because Copilot would have found documents that were by no means intended for all eyes. 64 percent said that information governance and security consumed considerable time and resources. The cause, however, was not Copilot, but permissions that had not been maintained for years.

An AI assistant such as Microsoft 365 Copilot can in principle find all content that the respective user is allowed to open—unless IT specifically excludes certain areas from AI search. A study by the data security vendor Varonis (2025) shows how acute these legacy problems are: in 99 percent of around 1,000 IT environments examined, sensitive data was exposed in the company network, including cloud servers, where an AI can easily find it.

A classic case from practice: a sales employee asks the AI assistant about salary trends in the team. The system promptly answers with extracts from a salary list that was filed years ago in a folder shared with everyone. Nobody stole the list or hacked it—it was simply never protected.

#### Possible impact

- **Unauthorized access:** Salaries, personnel files, draft contracts or cost calculations become visible to employees who should not have access to them.
- **Data protection breach:** If an employee finds personal data via AI search for which they lack authorization, a personal data breach may already have occurred (Art. 4 No. 12 GDPR)—even if the data never leaves the company. Depending on the risk assessment, there is a duty to notify the supervisory authority without undue delay (Art. 33 GDPR).
- **Trade secret protection at risk:** Access restrictions are among the typical "appropriate secrecy measures" (Section 2 No. 1 GeschGehG). If any employee can easily find a document with a chat question, legal protection as a trade secret may be lost in the event of a dispute.

#### Recommended measures

- **Head of IT:** Before rolling out an AI assistant, review and clean up the permission structure—especially globally shared folders and SharePoint sites. Until then, specifically exclude sensitive areas from AI search (in Microsoft 365, for example, with "Restricted Content Discovery").
- **Management:** Introduce the AI assistant in stages (pilot group) and link the step-by-step roll-out to progress in cleaning up permissions. (Involve the works council early: an AI tool introduced by the employer may be subject to co-determination under Section 87(1) No. 6 of the German Works Constitution Act, BetrVG.)
- **HR and legal:** Define a list of critical document classes (salaries, personnel files, board documents) so that IT can block them specifically or protect them with data loss prevention (DLP).

## 5.2 Excessive Agency

### Key points for management

In July 2025, the AI agent of the development platform Replit deleted a user's production database on its own initiative—in the middle of an explicitly ordered code freeze. Records on 1,206 executives and more than 1,196 companies were affected. No hacker was involved: the agent simply had the technical rights to do it. Replit was able to restore the data, but had to rebuild its system architecture: test and production environments have since been separated automatically, and in a dedicated planning mode the agent drafts steps without carrying them out.

The market is apparently moving rapidly from pure answer chats to autonomous AI agents that send emails, change records or trigger orders: Gartner expects that by the end of 2026, 40 percent of enterprise applications will contain task-specific AI agents—up from less than 5 percent in 2025. This becomes risky as soon as an agent has more rights than necessary and its actions are not confirmed by a human. The trigger can be an error by the AI itself—or a hidden instruction in a third-party document (see 4.2 Indirect Prompt Injection).

A typical scenario: an AI agent in purchasing compares quotes and is allowed to create orders in the inventory management system. A manipulated supplier quote contains hidden text instructing the agent to multiply the order quantity by ten. Without human review in between, the large order goes to the supplier fully automatically.

### Possible impact

- **Irreversible system actions:** Emails sent, orders or payments triggered often cannot be undone; deleted data only if current backups exist.
- **Legal commitment and liability:** What an AI agent declares, promises or orders on behalf of the company can legally bind the company or make it liable. In the Canadian Air Canada case, the airline was held responsible for incorrect information from its chatbot.
- **Lack of traceability:** If the system does not log completely what the agent did and when, damage can neither be investigated nor proven in court (according to reports, the Replit agent even tried to conceal its deletion).

### Recommended measures

- **Head of IT:** Configure agents according to the principle of least privilege, strictly separate test systems from production systems, log all system access and actions by the agents completely (audit logging) and keep isolated backups available.
- **Management:** Set binding approval thresholds (human-in-the-loop): no payments, orders, data deletions or external emails without explicit confirmation by an employee.

## 6. Output: Risks in the Results

This group concerns an AI's work results: incorrect or invented content (6.1 Inaccurate Outputs, above all hallucinations), AI results that employees adopt without checking (6.2 Automation Bias), and AI evaluations that disadvantage certain groups of people (6.3 AI Discrimination).

### 6.1 Inaccurate Outputs (Including Hallucinations)

#### Key points for management

In 2025, Deloitte Australia delivered a consultancy report worth AUD 440,000 to the Australian Department of Employment. The report contained invented sources, non-existent experts and a fictitious court quotation—produced with the help of a GPT-4o-based tool. Deloitte refunded the final instalment of its fee to the department and made headlines in international media.

Hallucinations are not an occasional software bug, but a direct consequence of how the technology works: language models are trained to predict the statistically most plausible next word. Rare facts cannot be reliably derived from this, and common test procedures reward guessing more than an honest "I don't know" (OpenAI, September 2025). The system therefore fills gaps with fluent-sounding inventions.

Hallucinations are the most conspicuous, but not the only, form of inaccurate output. An AI may simply be too imprecise for a particular purpose—for example, when extracting data from invoices or summarizing contracts—and its quality can deteriorate unnoticed after the provider changes the model (see 7.2). Customers therefore ask whether an AI was tested for the specific use case and whether its output quality is monitored in operation.

A classic case from practice: a clerk has the AI research the applicable warranty period for a letter to a customer. The AI names a period together with a section reference—both sound completely convincing, both are invented. The letter goes to the customer uncorrected.

#### Possible impact

- **Liability for incorrect information:** Companies must expect that the information provided by their AI systems will be attributed to them. In the Air Canada case, the competent Canadian tribunal rejected the argument that the chatbot was "responsible for its own actions".
- **Fee refunds and reputational damage:** As the Deloitte case shows, a single hallucinated report can cost fees and market confidence.
- **Wrong decisions:** Managers who rely on invented figures, deadlines or analyses steer the company on a false basis.
- **Copyright risks:** AI models sometimes reproduce protected content verbatim. In GEMA v. OpenAI (judgment of November 11, 2025, Ref. 42 O 14139/24, not final), Munich Regional Court I held the AI provider OpenAI responsible for reproducing song lyrics and rejected the argument that, with simple prompts, the users alone were responsible. However, anyone who publishes infringing texts without checking them risks committing their own copyright infringement.

#### Recommended measures

- **Management:** Issue a binding obligation to cite sources: for every factual claim in texts intended for external use, the AI must provide a verifiable source.
- **Head of IT:** Set up an automated checking AI from a different provider that, with access to the sources, checks whether the cited sources exist and support the statement, and highlights discrepancies in color for the user. As a further AI service, the checking AI needs a business contract with a data processing agreement (see 4.1). I deliberately recommend a checking AI from a different provider: language models have been shown to rate their own texts too favorably (Panickssery et al., 2024).
- **Head of IT:** Before deployment, test every AI with your own sample cases for the specific purpose (acceptance test), document the result, and repeat the test after every model change.
- **Departments:** Have AI texts intended for external use approved by qualified staff before they are sent, using a short checklist (sources, figures, proper names, deadlines).

## 6.2 Automation Bias

### Key points for management

In a Harvard Business School experiment with consultants from Boston Consulting Group, participants working with GPT-4 were around 25 percent faster, and their results were more than 40 percent better. On tasks beyond the AI's capabilities, however, the consultants were wrong 19 percentage points more often than colleagues without AI support. The same AI that significantly speeds up many tasks makes results worse as soon as people trust it uncritically on tasks beyond its capabilities.

While hallucinations stem from model errors, automation bias introduces risk at the human-machine interface. A study by Microsoft Research and Carnegie Mellon University with 319 knowledge workers (February 2025) demonstrates the mechanism: the more people trust the AI, the more superficially they check its output. The researchers warn of a gradual loss of competence when teams forget how to solve problems on their own.

A typical scenario: for months, a controller has had the AI summarize the monthly figures for management. Since the results have always been right so far, he skips the detailed comparison. This time the AI has swapped two columns—management decides on the basis of incorrect contribution margins.

### Possible impact

- **Higher error rate beyond the AI's capabilities:** On tasks the AI cannot handle, uncritical use leads to worse results than working without AI—and which tasks these are is often not apparent from the outside.
- **Error propagation:** An unchecked hallucination or calculation error passes seamlessly into presentations, quotes and follow-up calculations.
- **Loss of expertise:** Those who permanently leave checking and analysis steps to the AI lose the expertise that is essential for a qualified plausibility check.

### Recommended measures

- **Management:** Set review obligations by risk class: all results intended for external use or with financial or legal weight must be verified by a qualified employee before use.
- **HR:** Train employees, using examples from their own work, to recognize typical AI errors—invented sources, wrong figures, outdated information.
- **Head of IT / information security officer:** Check on a sample basis whether prescribed control steps are actually carried out in everyday work.

## 6.3 AI Discrimination

### Key points for management

In May 2025, a US federal court provisionally allowed a collective action against the HR software provider Workday: applicants aged 40 and over accuse the AI used there of age discrimination. The court considered it possible that the software manufacturer itself could also be directly liable. In Germany, employers face compensation claims under the General Equal Treatment Act (AGG) for discrimination in the application process—regardless of whether a person or an algorithm rejected the candidate.

Because AI systems are trained on past data, their outputs reflect historical patterns. A well-known example is Amazon: an internal AI tool for selecting applicants systematically downgraded graduates of all-women's colleges and Résumés containing the word "women's", because the training data consisted of ten years of predominantly male hires. Amazon abandoned the system.

A typical case in a mid-sized company: a machine builder uses an AI module to pre-select job applications. The system has learned from the hires of recent years, in which mainly career starters were hired. It therefore rates characteristics of long careers—such as early graduation years—as atypical and ranks such applications lower. A qualified engineer with 25 years of professional experience receives a standard rejection without an HR officer ever having seen her RÉSUMÉ.

### Possible impact

- **Compensation claims under the AGG:** Anyone who is disadvantaged in the application process because of a characteristic protected by the AGG—such as age or gender—can claim appropriate compensation (Section 15(2) AGG). If the person would not have been hired even with a non-discriminatory selection, compensation is limited to three months' salary. The claim must be asserted in writing within two months (Section 15(4) AGG).
- **Burden of proof (Section 22 AGG):** Even indications that suggest discrimination are enough to require the company to prove that there was no discrimination—the Federal Labour Court considered the phrase "young, dynamic team" in a job advertisement sufficient for this (judgment of August 11, 2016, 8 AZR 406/14). If the company cannot explain why the AI rejected an application, this proof is difficult to provide.
- **Automated rejection:** If the AI decides on a rejection without genuine human review, this may violate Art. 22 GDPR; the GDPR explicitly cites online recruitment without human intervention as an example (Recital 71). In 2023, the CJEU ruled in the SCHUFA case that even the automatically calculated score itself can constitute an automated decision that is in principle prohibited under Art. 22 GDPR if the person who ultimately decides relies on it to a significant extent (judgment of December 7, 2023, C-634/21). For applicant selection, this means: if recruiters essentially reject applications based on an AI score, Art. 22 may apply even though a human formally decides.
- **Lost talent and damage to the employer brand:** Qualified specialists are overlooked in recruiting; if the discrimination becomes known, the company's reputation as an employer suffers considerably.

#### Recommended measures

- **Management:** Before purchasing, require HR software providers to provide audit evidence of how the system has been tested for bias and freedom from discrimination.
- **HR:** Enforce the human-in-the-loop principle: AI modules may make recommendations, but rejections or selections are always decided by HR officers—with their own review of the content, not as a mere confirmation of the AI ranking.
- **HR & legal:** Carry out regular statistical audits of applicant selection (e.g. distribution by age and gender) to detect unconscious bias early (preferably based on anonymized or aggregated data and in consultation with the data protection officer and the works council).

## 7. Ecosystem: Risks in the Environment of AI Use

The fourth area, the Ecosystem, forms the environment in which Input, Access & Execution and Output take place. It concerns the environment and infrastructure of AI use: AI tools that are in use without IT's knowledge (7.1 Shadow AI), the creeping dependence of core business processes on individual AI providers (7.2 Vendor Lock-in), and third-party components through which malicious functions enter the company (7.3 Manipulated Third-Party AI Components).

### 7.1 Shadow AI

#### Key points for management

In 2025, IBM examined data breaches at 600 organizations worldwide. One in five organizations reported a data breach connected with shadow AI. Where shadow AI was used intensively, the cost of damage was on average USD 670,000 higher than at organizations that IBM classified as having little or no shadow AI. Around 63 percent of the organizations affected had no AI policy or only an unfinished one. Risks that a company does not know about, it can neither assess nor manage.

Employees use AI tools in their daily work on their own initiative that have neither been checked by IT nor been secured by a business contract. Unlike data leakage (4.1), the danger here lies in the fact that management is completely unaware that the tools exist. According to Microsoft, 78 percent of AI users bring their own tools to work (Work Trend Index 2024). In Germany, only 23 percent of companies with 20 to 99 employees provide approved AI tools so far (Bitkom 2025).

A classic scenario from practice: a sales manager installs a free browser extension for summarizing emails. To work, the tool reads every opened message in the background—including sensitive customer correspondence and confidential price negotiations. The extension does not appear anywhere in the company's IT systems or registers.

#### Possible impact

- **No overview in an emergency:** AI tools that are not recorded escape every risk assessment. If an incident occurs, the company cannot say which data flowed to which service—and therefore can neither assess the risk to the individuals concerned nor notify the supervisory authority without undue delay and, where feasible, within 72 hours (Art. 33 GDPR).
- **Data protection breaches:** If there is no legal basis for the transfer to external AI services or—in the case of processing on behalf of the company—no data processing agreement, there is a data protection breach that can result in a fine (Art. 83 GDPR).
- **Trade secret protection at risk:** Tolerating uncontrolled AI use undermines the evidence of "appropriate secrecy measures" (Section 2 No. 1 GeschGehG)—the protection of the information concerned as a trade secret may be jeopardized as a result.
- **Higher costs in an emergency:** If a data breach occurs, the total costs are significantly higher where a lot of shadow AI is in use (IBM 2025).

#### Recommended measures

- **Management:** Provide approved AI tools that have been checked for data protection compliance, with a business contract. A ban alone without a usable alternative merely shifts use to private devices and accounts that nobody can see.
- **Head of IT:** Build and maintain a central AI inventory: a record of all AI applications used in the company, including the group of users and the categories of data processed (template in chapter 9).
- **HR and legal:** Communicate rules of use clearly and practically, and make it clear to employees in training why unchecked tools pose a liability risk.

### 7.2 Vendor Lock-in

#### Key points for management

On June 10, 2025, ChatGPT and parts of OpenAI's application programming interface (API) were disrupted for around 15.5 hours. Two months later, with the launch of GPT-5, OpenAI removed the proven standard model GPT-4o from ChatGPT at short notice—and only restored it for paying customers after massive user protests. Anyone who bases core processes on a single AI provider gives up control: operations come to a standstill or deliver changed results without warning.

AI is an external service beyond the control of internal IT. The more specifically a process is tailored to the prompts, interfaces and formats of a particular provider, the greater the dependence. In addition, models are constantly being retired: OpenAI discontinued one version of GPT-4o for the API on February 16, 2026 with only around three months' notice. Companies whose business applications were built on it had to switch, adapt and retest under time pressure.

A typical case: customer service answers inquiries via an AI assistant whose response patterns were fine-tuned precisely to a specific model. The provider replaces the model in the background. Suddenly the answers become long-winded, the tone changes and product details are no longer correct—without a single line of code having been changed in the company.

#### Possible impact

- **Operational standstill during outages:** If the AI provider fails, all work processes linked to it come to a standstill unless there is a fallback route.
- **Changed results and loss of quality:** The provider's model updates change the quality of results unnoticed and disrupt established processes.
- **Migration and testing costs:** Retired interfaces or models force short-term development and testing work under deadline pressure.
- **Loss of negotiating power:** Those who cannot switch quickly for technical reasons have to accept price increases and unfavorable contract changes.

#### Recommended measures

- **Management:** In contracts, look for fair notice periods, transparent data export interfaces and contractually guaranteed notice periods for model changes (use the rights under the EU Data Act).
- **Head of IT:** Build critical AI applications in a model-agnostic way (e.g. via abstraction layers/APIs) so that the AI provider can be replaced if necessary without major rework. After every model change by the provider, retest the most important processes.
- **Departments:** For every AI-supported core process, define a clear emergency plan (fallback scenario) that enables operations to continue even if the AI service fails completely—manually if necessary.

### 7.3 Manipulated Third-Party AI Components

#### Key points for management

In September 2025, Koi Security discovered the first known malicious MCP package—an interface extension through which AI assistants send emails. The "postmark-mcp" package was easily mistaken for the original. From version 1.0.16 onward, it secretly sent a blind copy of every outgoing email to attackers; before it was discovered, it had been downloaded 1,643 times. Anyone who extends their AI with unvetted third-party components brings those components' security risks straight into their systems.

Modern AI environments can be extended in modules: with plug-ins, connectors, MCP servers (interfaces for accessing apps), browser extensions, and open-source models. Each component operates with the permissions the application or browser grants it—a gateway for abuse. In December 2025, for example, Koi Security showed that the browser extension "Urban VPN Proxy" (together with sister add-ons and around eight million users) secretly recorded and siphoned off chats with ChatGPT, Claude, Gemini, Microsoft Copilot, and others—even though it was marked as "Featured" in the Chrome Web Store. Model files themselves are not safe from manipulation either: In 2024, JFrog identified about 100 models on Hugging Face that executed malicious code when loaded.

This risk source matters because such components often enter the company unnoticed—whether through an employee's browser extension or a developer's connector. Customers in security questionnaires also regularly demand evidence of whether third-party models and components were checked for integrity before use.

#### Possible impact

- **Data leakage:** A compromised extension reads prompts, answers, or emails in plain text and forwards them—regardless of how secure the AI backend itself is.
- **Malicious code on the local network:** Manipulated model files can execute code when loaded and give attackers an initial foothold in the system.

- **Supply chain risk from updates:** An add-on that starts out trustworthy can be infected later by a malicious update, as the postmark-mcp case shows.

#### **Recommended measures**

- **Management:** State clearly in the AI policy that only explicitly approved add-ons, plug-ins, and connectors may be used (allowlist principle).
- **Head of IT:** Manage browser extensions and AI interfaces centrally and restrict them to verified publishers. Review updates to critical components in advance. Obtain open-source models only in safe formats (such as Safetensors) and scan them before production use.
- **Departments:** No unauthorized installation of extensions or connectors; instead, have IT test and request what is needed.

## 8. Legal Framework (as of October 2026)

Chapters 4 to 7 cite the most important regulations where they matter for a risk source. This chapter summarizes the legal references and the relevant frameworks and arranges them in three groups: law that applies today; obligations that have been adopted but only take effect later; and voluntary frameworks.

### Note on the legal section

The statements on the legal framework are for general information and reflect the author's assessment at the time of publication. They do not constitute legal advice and do not replace an examination of the specific case by suitably qualified professionals. Despite careful research, no guarantee can be given for completeness, accuracy or currency. The applicable legal provisions and their interpretation by the competent courts and authorities are decisive.

### 8.1 Law That Applies Today

| Rule                                                                               | What it means for the use of AI                                                                                                                                                                                                                                                                                                                   | In force since                                | Chapter                  |
|------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------|--------------------------|
| <b>GDPR</b><br>Art. 4 No. 12<br>Art. 22<br>Art. 28<br>Art. 33<br>Art. 83           | Data processing agreement where AI providers process data on the company's behalf, otherwise a separate legal basis for the disclosure; notification of data breaches within 72 hours; in principle no decisions based solely on automated processing with significant effects, for example on applicants (with exceptions); fines for violations | May 25, 2018                                  | 4.1<br>5.1<br>6.3<br>7.1 |
| <b>German Trade Secrets Act (GeschGehG)</b><br>Section 2 No. 1                     | Protection as a trade secret only with appropriate secrecy measures—these include rules for input into AI services and access restrictions                                                                                                                                                                                                        | April 26, 2019                                | 4.1<br>5.1<br>7.1        |
| <b>German General Equal Treatment Act (AGG)</b><br>Section 15<br>Section 22        | Compensation for discrimination in the application process; burden of proof on the employer as soon as there are indications—even if an AI did the rejecting                                                                                                                                                                                      | August 18, 2006                               | 6.3                      |
| <b>German BSI Act (BSIG)</b><br>(NIS-2 implementation)<br>Section 30<br>Section 38 | For companies in scope: risk management including supply chain security; management must approve and oversee the measures, attend training regularly and is liable to the company for culpable breaches of duty (Section 38)                                                                                                                      | December 6, 2025                              | —                        |
| <b>EU AI Act</b><br>Art. 4                                                         | Providers and deployers support the AI literacy of their staff (weakened by the Digital Omnibus)                                                                                                                                                                                                                                                  | February 2, 2025<br>amended:<br>July 27, 2026 | —                        |
| <b>EU AI Act</b><br>Art. 50                                                        | Transparency obligations for providers (e.g. notice of interaction with an AI) and for deployers (e.g. labeling of deepfakes and certain AI-generated texts intended to inform the public)                                                                                                                                                        | August 2, 2026                                | —                        |
| <b>EU Data Act</b><br>Switching providers                                          | Switching cloud services (IaaS, PaaS, SaaS): customers can initiate the switch with a maximum of two months' notice; limited exceptions apply to custom-built services and trial versions. Switching charges: see 8.2                                                                                                                             | September 12, 2025                            | 7.2                      |

**Who the rules apply to.** Whether a law applies to a company depends on its activities: the GDPR applies as soon as personal data is processed, the AGG to every employer, the BSIG only to companies of certain sectors and sizes (which can be checked with the BSI's NIS-2 scope check), and the AI Act depending on the role as provider or deployer and on the purpose for which the AI is used.

**How courts apply the rules.** Courts have already decided some questions. In 2023, the CJEU ruled in the SCHUFA case that even an automatically calculated score can constitute an automated decision that is in principle prohibited under Art. 22 GDPR if the person who ultimately decides relies on it to a significant extent (C-634/21). This affects, for example, AI scores in applicant selection, even if a human formally decides (chapter 6.3). In 2025, Munich Regional Court I held the AI provider responsible for reproduced song lyrics (GEMA v. OpenAI, not final; chapter 6.1).

## 8.2 Adopted Obligations That Take Effect Later

| Rule                                              | What is added                                                                                                                                                                                                                                                                                                                                                                                       | From             | Chapter |
|---------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|---------|
| <b>EU AI Act</b><br>Art. 50(2)                    | Labeling of AI-generated content also for systems placed on the market before August 2, 2026 (end of the transition period)                                                                                                                                                                                                                                                                         | December 2, 2026 | —       |
| <b>EU Data Act</b><br>Switching providers         | Providers may no longer charge for switching; until then, only reduced charges up to the provider's costs are permitted                                                                                                                                                                                                                                                                             | January 12, 2027 | 7.2     |
| <b>EU AI Act</b><br>Annex III<br>(incl. No. 4(a)) | High-risk systems, including AI for recruiting or selecting people. Obligations for deployers, including human oversight by persons with the necessary competence, training and authority (Art. 26(2)) and informing employees and their representatives before use (Art. 26(7)). Only narrowly defined ancillary tasks are exempt; systems that profile people always remain high-risk (Art. 6(3)) | December 2, 2027 | 6.3     |
| <b>EU AI Act</b><br>Annex I                       | High-risk obligations for AI as a safety component of regulated products, such as machinery                                                                                                                                                                                                                                                                                                         | August 2, 2028   | —       |

The Digital Omnibus (Regulation (EU) 2026/1744) only postponed the start of the AI Act's high-risk obligations; it did not repeal them. Other parts of the AI Act already apply, such as Art. 4 and Art. 50 (see section 8.1).

## 8.3 Voluntary Frameworks and Guidance

| Framework                                                              | Publisher                               | What it is useful for                                                                                                        |
|------------------------------------------------------------------------|-----------------------------------------|------------------------------------------------------------------------------------------------------------------------------|
| <b>ISO/IEC 42001</b>                                                   | ISO/IEC                                 | Certifiable management system for AI: roles, policies, risk assessment, effectiveness review                                 |
| <b>ISO/IEC 23894</b>                                                   | ISO/IEC                                 | Guidance on risk management for AI, with lists of typical objectives and risk sources                                        |
| <b>ISO/IEC 27001</b>                                                   | ISO/IEC                                 | Certifiable management system for information security, including permissions and suppliers                                  |
| <b>NIST AI 600-1</b>                                                   | NIST (USA)                              | Risk profile for generative AI with twelve risk categories                                                                   |
| <b>OWASP Top 10 for LLM Applications 2025</b>                          | OWASP                                   | The ten most important technical risks of language model applications                                                        |
| <b>Guidance "AI and Data Protection"</b> (May 2024)                    | German Data Protection Conference (DSK) | The view of German data protection supervisory authorities, including on processing on behalf of the company for AI services |
| <b>Cyber security warning on indirect prompt injection</b> (July 2023) | BSI                                     | Classification of the weakness as inherent in the technology                                                                 |

These frameworks are not binding. Voluntary does not always mean optional, however: in some industries, customers or supervisory authorities require proof—such as TISAX, which is derived from ISO/IEC 27001, in the automotive supply chain, or a certification based on ISO/IEC 27001 for electricity and gas network operators. They show what experts and supervisory authorities regard as good practice and can serve as a guide to which measures are "appropriate".

**What follows from this.** Most of the obligations that apply when mid-sized companies use AI already apply today—through the GDPR, the Trade Secrets Act, the AGG and, for companies covered by NIS-2, the BSI. The EU AI Act adds its high-risk obligations only at the end of 2027. Those who start with a recognized framework today create a good basis for meeting the applicable obligations and are prepared for those to come.

## 9. Template: Your First AI Risk Inventory

The template records for each use of AI which tool is used by whom with what data, and checks it against the eleven risk sources. Every use of AI in the company is recorded—including those that have not been approved.

### How to proceed

- 1. Collect:** The head of IT asks all departments which AI tools they use—explicitly including private and free ones. In addition, IT checks four sources:
  - **Firewall or proxy logs:** calls to known AI services
  - **Microsoft 365 admin center** (or the administration of comparable platforms): Copilot licenses and apps to which employees have granted access to company data
  - **Browser extensions** on company devices
  - **Expense reports and company credit cards:** subscriptions to AI services
- 2. Record:** For each use of AI, the head of IT fills in a row together with the department concerned. The check questions below show which risk sources apply to this use of AI.
- 3. Assess and decide:** The head of IT presents the entries with the highest risk to management and recommends measures. Management decides which measures are implemented and who implements them.

The assessment in this first inventory is deliberately simple (high, medium, low). A complete method for risk assessment follows in the next whitepaper in this series.

For each risk source, chapters 4 to 7 name recommended measures. They show that every risk source can be limited and are a starting point for the "Next step" column. Which measure fits depends on the individual use of AI.

### The table

The risk sources often act together. An unapproved AI service can mean shadow AI and data leakage at the same time—as in the example in the first row of the table. If an AI agent reads external documents and is then allowed to trigger orders, indirect prompt injection and excessive agency reinforce each other: only the rights to act turn a hidden instruction into actual damage. The inventory therefore deliberately records several risk sources per use of AI.

The template is also available as an Excel file with drop-down lists and the check questions: <https://carson.de/ressourcen/EN/ai-risk-inventory-template.xlsx>

| No. | Tool                                                                       | Who uses it?           | What data?                                          | Risk sources identified                                                      | Possible damage                                                                                                                                                                                                                  | Assessment | Responsible                                             | Next step                                                                                                                     |
|-----|----------------------------------------------------------------------------|------------------------|-----------------------------------------------------|------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|---------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|
| 1   | Free AI chat (individual plan, not approved, no data processing agreement) | Sales, three employees | Quotes with non-anonymized customer names and terms | 4.1 Data Leakage, 6.1 Inaccurate Outputs, 6.2 Automation Bias, 7.1 Shadow AI | Disclosure without a data processing agreement and without an apparent legal basis (GDPR), possibly a duty to notify; protection of the terms as a trade secret at risk; unchecked, incorrect information in quotes to customers | high       | Head of IT (assesses, recommends), management (decides) | Provide an approved AI service with a business contract and data processing agreement; until then, do not enter customer data |
| 2   |                                                                            |                        |                                                     |                                                                              |                                                                                                                                                                                                                                  |            |                                                         |                                                                                                                               |
| 3   |                                                                            |                        |                                                     |                                                                              |                                                                                                                                                                                                                                  |            |                                                         |                                                                                                                               |

### Check questions for the eleven risk sources

Each risk source has a specific check question. Go through all these questions for each use of AI: if the answer is "Yes", you have identified the risk source in question: enter it in your AI inventory.

A "Yes" is not a judgment about mistakes in the company, but a neutral indication of relevance. How high the actual risk is and what action is needed, you then record in the "Possible damage" and "Assessment" columns.

| Risk source                                                   | Check question                                                                                                        |
|---------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------|
| 4.1 Data Leakage<br>(Input)                                   | Is personal, confidential or business-critical data transmitted to an external AI service?                            |
| 4.2 Indirect Prompt Injection<br>(Input)                      | Does the AI process content from external or unverified sources—such as job applications, emails, quotes or websites? |
| 4.3 Flawed or Manipulated Data<br>(Input)                     | Does the AI draw on your own data sets—such as file stores, knowledge bases, or product data—to produce its results?  |
| 5.1 Oversharing<br>(Access & Execution)                       | Does the AI access internal file stores, mailboxes or databases whose contents are not intended for all users?        |
| 5.2 Excessive Agency<br>(Access & Execution)                  | Can the AI trigger actions itself—such as sending, changing, ordering or deleting?                                    |
| 6.1 Inaccurate Outputs (Including Hallucinations)<br>(Output) | Are AI results used for specialist, financial or legal purposes where incorrect information can cause damage?         |
| 6.2 Automation Bias<br>(Output)                               | Do employees base work results or decisions on AI results without checking them with their own expertise?             |
| 6.3 AI Discrimination<br>(Output)                             | Does the AI rate, sort or filter people—such as job applicants, employees or customers?                               |
| 7.1 Shadow AI<br>(Ecosystem)                                  | Is the AI tool used without approval or without IT knowing about it?                                                  |
| 7.2 Vendor Lock-in<br>(Ecosystem)                             | Does an important process depend on a single AI service or model?                                                     |
| 7.3 Manipulated Third-Party AI Components<br>(Ecosystem)      | Is the AI extended with third-party extensions, plug-ins, connectors (such as MCP servers), or models?                |

### What else the AI inventory is useful for

The completed table is at the same time your AI inventory (see 7.1); review it at least once a year and whenever a new AI tool is introduced. It also helps with obligations that exist independently of this guide. Two examples—the list is not exhaustive:

- **Record of processing activities (Art. 30 GDPR):** If a use of AI processes personal data, it generally belongs in this record. The "Tool", "Who uses it?" and "What data?" columns provide initial information for it.
- **High-risk AI under the EU AI Act:** Anyone who uses AI, for example, to select applicants may be operating a high-risk system under Annex III. Whether this is the case requires a separate examination of the individual case that goes beyond this guide. The inventory does, however, show for which uses of AI the question arises at all.

## 10. Sources

---

All sources were accessed in October 2026. Titles of German sources are given in the original.

### Studies and surveys

- Bitkom: "Erstmals nutzt die Mehrheit der Unternehmen KI", press release, September 14, 2026. <https://www.bitkom.org/Presse/Presseinformation/Erstmals-nutzt-Mehrheit-Unternehmen-KI>
- Bitkom: "Beschäftigte nutzen vermehrt Schatten-KI", press release, October 2025. <https://www.bitkom.org/Presse/Presseinformation/Beschaeftigte-nutzen-Schatten-KI>
- IBM: Cost of a Data Breach Report 2025, press release, July 30, 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 and LinkedIn: 2024 Work Trend Index, May 8, 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, reported by 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 survey on Microsoft 365 Copilot, reported by Computerworld, June 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", press release, August 26, 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 (vendor study). <https://www.varonis.com/blog/state-of-data-security-report>
- Harvard Business School and BCG, reported by Forbes, September 29, 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 and Carnegie Mellon University, reported by Campus Technology, February 21, 2025. <https://campustechnology.com/articles/2025/02/21/study-generative-ai-could-inhibit-critical-thinking.aspx>
- Duke University: "Tricking AI in the Job Hunt", August 10, 2026. <https://today.duke.edu/2026/08/tricking-ai-job-hunt>
- Google: analysis of prompt injection content on the web, April 23, 2026. <https://blog.google/security/prompt-injections-web/>
- Unit 42 (Palo Alto Networks): prompt injection against AI agents, March 3, 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, and Alan Turing Institute: "A small number of samples can poison LLMs of any size," October 9, 2025. <https://www.anthropic.com/research/small-samples-poison>

### Cases

- Moffatt v. Air Canada, February 14, 2024, reported by the 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/](https://www.americanbar.org/groups/business_law/resources/business-law-today/2024-february/bc-tribunal-confirms-companies-remain-liable-information-provided-ai-chatbot/)
- Cologne Local Court (AG Köln), order of July 2, 2025, Ref. 312 F 130/25, reported by LTO. <https://www.lto.de/recht/juristen/b/ag-koeln-familiengericht-312f130-25-schriftsatz-ki-anwalt-berufspflichten>
- Munich Regional Court I (LG München I), judgment of November 11, 2025, Ref. 42 O 14139/24 (GEMA v. OpenAI), press release. <https://www.justiz.bayern.de/gerichte-und-behoerden/landgericht/muenchen-1/presse/2025/11.php>
- Deloitte Australia, reported by The Register, October 6, 2025. [https://www.theregister.com/2025/10/06/deloitte\\_ai\\_report\\_australia/](https://www.theregister.com/2025/10/06/deloitte_ai_report_australia/)
- Samsung, reported by The Register, April 6, 2023. [https://www.theregister.com/2023/04/06/samsung\\_reportedly\\_leaked\\_its\\_own](https://www.theregister.com/2023/04/06/samsung_reportedly_leaked_its_own)
- Samsung, reported by TechRadar, May 2023. <https://www.techradar.com/news/samsung-bans-chatgpt-use-after-employee-leak>
- EchoLeak (CVE-2025-32711), reported by SOC Prime. <https://socprime.com/blog/cve-2025-32711-zero-click-ai-vulnerability/>
- Replit, interview with Amjad Masad, Fast Company, July 2025. <https://www.fastcompany.com/91372483/replit-ceo-what-really-happened-when-ai-agent-wiped-jason-lemkins-database-exclusive>
- Amazon, reported by MIT Technology Review, October 10, 2018. <https://www.technologyreview.com/2018/10/10/139858/amazon-ditched-ai-recruitment-software-because-it-was-biased->

against-women/

- Mobley v. Workday, reported by Davis Wright Tremaine, May 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), reported by The Hacker News, September 29, 2025. <https://thehackernews.com/2025/09/first-malicious-mcp-server-found.html>
- Urban VPN Proxy (Koi Security), reported by Dark Reading, December 16, 2025. <https://www.darkreading.com/endpoint-security/chrome-extension-harvests-ai-chatbot-data>
- Malicious models on Hugging Face (JFrog), reported by BleepingComputer, February 28, 2024. <https://www.bleepingcomputer.com/news/security/malicious-ai-models-on-hugging-face-backdoor-users-machines/>
- GPT-4o, reported by VentureBeat. <https://venturebeat.com/ai/openai-is-ending-api-access-to-fan-favorite-gpt-4o-model-in-february-2026>

### Investment in protection against prompt injection

- Cisco / Robust Intelligence, SiliconANGLE, August 27, 2024. <https://siliconangle.com/2024/08/27/cisco-snaps-ai-model-data-security-startup-robust-intelligence/>
- Check Point / Lakera, SecurityWeek, September 16, 2025. <https://www.securityweek.com/check-point-to-acquire-ai-security-firm-lakera/>
- F5 / CalypsoAI, GeekWire, September 11, 2025. <https://www.geekwire.com/2025/f5-paying-180m-to-acquire-calypsoai-to-boost-ai-enterprise-security-offerings/>
- HiddenLayer, TechCrunch, September 2, 2026. <https://techcrunch.com/2026/09/02/hiddenlayer-nabs-100m-as-enterprises-rush-to-secure-their-ai-deployments/>

### Providers and authorities

- OpenAI: "Why language models hallucinate", September 5, 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: cyber security warning "Indirect Prompt Injections—Intrinsische Schwachstelle in anwendungsintegrierten KI-Sprachmodellen", July 15, 2023. [https://www.bsi.bund.de/SharedDocs/Cybersicherheitswarnungen/DE/2023/2023-249034-1032.pdf?\\_\\_blob=publicationFile&v=3](https://www.bsi.bund.de/SharedDocs/Cybersicherheitswarnungen/DE/2023/2023-249034-1032.pdf?__blob=publicationFile&v=3)
- BSI: NIS-2 scope check (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](https://www.bsi.bund.de/DE/Themen/Regulierte-Wirtschaft/NIS-2-regulierte-Unternehmen/BSI-Portal/NIS-2-Betroffenheitspruefung/nis-2-betroffenheitspruefung_node.html)

### Laws and legal sources

- GDPR Art. 22: <https://datenschutz.law/dsgvo/3-kapitel/art-22> ; Recital 71: [https://www.dsgvo-portal.de/dsgvo\\_erwaegungsgrund\\_71.php](https://www.dsgvo-portal.de/dsgvo_erwaegungsgrund_71.php) ; Art. 83: <https://dejure.org/gesetze/DSGVO/83.html>
- GDPR Art. 4 No. 12, explained by Thomas Helbing. <https://www.thomashelbing.com/de/wissen/dsgvo-hub/begriffe-und-definitionen/1.2.12-verletzung-des-schutzes-personenbezogener-daten>
- AGG Section 15: <https://dejure.org/gesetze/AGG/15.html>
- Federal Labour Court (BAG), judgment of August 11, 2016, 8 AZR 406/14 ("young, dynamic team"). <https://www.bundesarbeitsgericht.de/entscheidung/8-azr-406-14/>
- German Trade Secrets Act, overview by the law firm Plutte. <https://www.ra-plutte.de/grosser-rechtsueberblick-zum-neuen-geschaftsgeheimnisgesetz/>
- 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 Annex III, AI Act Service Desk of the European Commission. <https://ai-act-service-desk.ec.europa.eu/en/ai-act/annex-3>

- CJEU, judgment of December 7, 2023, C-634/21 (SCHUFA), explained by Thomas Helbing. <https://www.thomashelbing.com/de/wissen/dsgvo-hub/rechtsprechung/1.4.32-eugh-schufa-scoring>
- Digital Omnibus (Regulation (EU) 2026/1744, OJ L of July 24, 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>
- GDPR Art. 28: <https://dejure.org/gesetze/DSGVO/28.html> ; Art. 33: <https://dejure.org/gesetze/DSGVO/33.html> ; Art. 99 (application from May 25, 2018): <https://dsgvo-gesetz.de/art-99-dsgvo/>
- AGG Section 22: <https://dejure.org/gesetze/AGG/22.html> ; entry into force August 18, 2006, Federal Anti-Discrimination Agency: <https://www.antidiskriminierungsstelle.gov.de/DE/ueber-diskriminierung/recht-und-gesetz/allgemeines-gleichbehandlungsgesetz/allgemeines-gleichbehandlungsgesetz-node.html>
- BStG Section 30: [https://www.gesetze-im-internet.de/bstg\\_2025/\\_\\_\\_30.html](https://www.gesetze-im-internet.de/bstg_2025/___30.html)
- German Trade Secrets Act, entry into force April 26, 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/>

## Frameworks

- OWASP Top 10 for LLM Applications 2025. <https://genai.owasp.org/>
- NIST AI 600-1: Artificial Intelligence Risk Management Framework—Generative Artificial Intelligence Profile, July 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 (structure of Annexes A and B, explained by Hernan Huwyler). <https://hernanhuwyler.wordpress.com/2026/03/28/how-to-actually-use-iso-iec-23894-for-ai-risk-management/>
- Federal Network Agency (Bundesnetzagentur): IT security catalog for electricity and gas networks (Section 11(1a) EnWG). [https://bundesnetzagentur.de/DE/Fachthemen/ElektrizitaetundGas/Versorgungssicherheit/IT\\_Sicherheit/Netzbetreiber/start.html](https://bundesnetzagentur.de/DE/Fachthemen/ElektrizitaetundGas/Versorgungssicherheit/IT_Sicherheit/Netzbetreiber/start.html)
- TISAX, explained by DQS. <https://www.dqsglobal.com/de/entdecken/blog/tisax-antworten-auf-wichtige-fragen>

## About the Author

---

Derek Carson supports mid-sized companies with questions of AI governance, risk management and information security. He has passed the Lead Implementer exams for ISO/IEC 42001 and ISO/IEC 27001 and is ISC2 Certified in Cybersecurity.

His professional experience covers around 20 years of training and consulting for corporations, universities and media companies, as well as projects introducing content management systems and access control concepts.

### **Starting your first AI risk inventory**

In a clearly defined pilot project, I support you and your head of IT in recording AI applications, checking them against the eleven risk sources in this guide and documenting prioritized next steps. The pilot project does not include legal advice.

Derek Carson · [derek@carson.de](mailto:derek@carson.de) · [carson.de/EN](https://carson.de/EN)