Responsible Use of AI: Understanding, Managing and Insuring Against the Risks

A guide to structuring and assessing AI risks.

bitkom

Content

1 Introduction: AI as a Standard Component of Corporate Opportunity and Risk Management 4

Abstract 4

AI – Understanding Risks Systematically 5

What We Mean by AI 5

Potential Risks and Limitations of AI System Performance 7

AI in Practice 8

2 Governance, Compliance and Management of AI Use 9

EU AI Act: Governance and Compliance Requirements 9

High-Risk AI in Selected Application areas – Including Agent-Based Systems 10

Non-High-Risk AI with Compliance Obligations 12

Outlook: From Compliance to 'Good AI Governance' 15

How Are Others Handling This? A Look Abroad 16

3 AI Security Is More Than Just Cybersecurity 17

Potential Problems and Challenges 18

Preparing for the Deployment of AI Agents 19

Technical Security Architecture (Structural Safeguards) 19

How Is This Process Implemented in Practice? 21

When Is 'Human-in-the-loop' Absolutely Essential? 21

Cultural and Employee-Focused Adaptation (the Human Component) 21

4 Identifying, Classifying and Assessing AI Risks 22

AI Represents a Technological Leap with New Dimensions of Opportunity and Risk 22

Types of Loss: An Attempt at a Structured Analysis 22

5 Self-Inflicted Damage Resulting from AI Use 24

What Is First-Party Loss? 24

Damage to the AI Asset Itself 24

AI-Induced Property Damage to Other Business Assets 26

AI-Induced Financial Losses 28

Conclusion: Risk of First-Party Loss Arising from the Use of AI **31**

6 Third-Party Liability Arising from the Use of AI 33

What is Third-Party Damage? **33**

Personal Injury 33

Property Damage 34

Pure Financial Loss 34

Purely Contractual Liability / Warranties 35

Regulatory and Liability Law Framework 35

Liability of the AI Manufacturer 36

Liability of the AI User **37**

Operator and Quasi-Manufacturer Roles 37

Tortious Liability 37

Strict Liability and Product Liability 38

Contractual Liability 39

Conclusion: Risk of Third-Party Damage Under the AI **39**

7 Governance and Reputational Risks 41

8 Many Parties – One AI model: Joint Insurance Cover for a Shared AI Project Risk? 44

9 Outlook 45

Glossary 47

Appendix 51

Bibliography 54

1 Introduction: AI as a Standard Component of Corporate Opportunity and Risk Management

Abstract

Artificial intelligence is increasingly being integrated into business processes, decision-making structures and products. Today, 57 percent of companies in Germany use AI – compared with 36 percent in 2025 and just 20 percent two years ago. Its use spans a wide range of business processes, from customer contact (72 percent) and internal services (20 percent) to accounting (25 percent).[1] The increasing use of AI is changing the risk landscape. This publication examines the types of losses that may arise from the use of AI, how these can be systematically categorised, and to what extent existing insurance solutions are suitable for covering the resulting risks.

The aim of the publication is to establish a common understanding of AI-related risks and to provide a framework for their assessment, management and insurance coverage. The focus is on companies that use AI. The publication is also intended to contribute to the broader discussion within the insurance industry.

The analysis shows that AI does not generally create entirely new types of loss, but can alter or exacerbate existing risks. For the purposes of insurance underwriting, therefore, the decisive factors are not the name of the technology but the specific nature of the loss, its cause and the degree to which it can be controlled. Existing insurance solutions are already capable of covering individual AI-related risks, provided these fall within the scope of the relevant class of insurance. Existing liability law, including the reformed product liability law, covers damages caused by AI.

Significant challenges already exist, particularly in relation to pure financial losses incurred by companies without any connection to property damage or personal injury, governance and compliance risks, AI agent systems, and the allocation of responsibility within complex AI value chains. Good governance, clear responsibilities, effective control mechanisms and robust risk management are therefore key prerequisites for managing AI risks and ensuring their continued insurability in the long term.

This publication is aimed in particular at insurers, InsurTechs, brokers and other market participants involved in insuring against AI risks. It is primarily intended for specialists

57 percent

of companies in Germany use AI, and 38 percent are planning or discussing its use.

↗Dataverse (2026)

[1] ↗Dataverse (2026)

and managers working in underwriting across all lines of business, risk management, IT security and governance.

AI – Understanding Risks Systematically

Artificial intelligence (AI) is increasingly becoming a foundational technology for the digital economy. Its use affects almost all sectors and stages of the value chain. Companies of all sizes are integrating AI-based applications into their processes and decision-making structures, and in some cases into their products. As a result, AI is no longer a technological issue that can be considered in isolation, but rather a cross-cutting issue affecting the entire economy.

The use of AI offers considerable potential for innovation and growth, but at the same time creates new demands on risk management, governance and safeguards. A key prerequisite for the controlled and scalable use of AI is a shared and, as far as possible, widely accepted understanding of the relevant applications and risks. The aim of this publication is therefore to help raise awareness of AI as a new cross-cutting risk, to identify and classify the main risks arising from the use of AI, and to provide initial guidance to inform future discussions about risk management and insurance approaches.

Raising awareness is not an end in itself, but a prerequisite for the responsible, legally compliant use of AI in ways that foster innovation. In this sense, this publication is also intended to stimulate a broader discussion within the insurance industry on how to address AI-related risks and their insurability in the future.

What We Mean by AI

A consistent assessment of opportunities and risks requires a shared understanding of what is meant by artificial intelligence. This publication therefore draws on the definition set out in the EU AI Act.[2]

[2] The definition in the EU AI Act is used here as a reference because it is legally binding, technology-neutral and suited to governance, compliance and risk considerations. In particular, it focuses on software-based systems with a certain degree of autonomy that are capable of generating outputs such as predictions, recommendations, content or decisions. Given its broad scope, it may be difficult in practice to distinguish AI systems from traditional statistical methods, rule-based systems or simple automation.

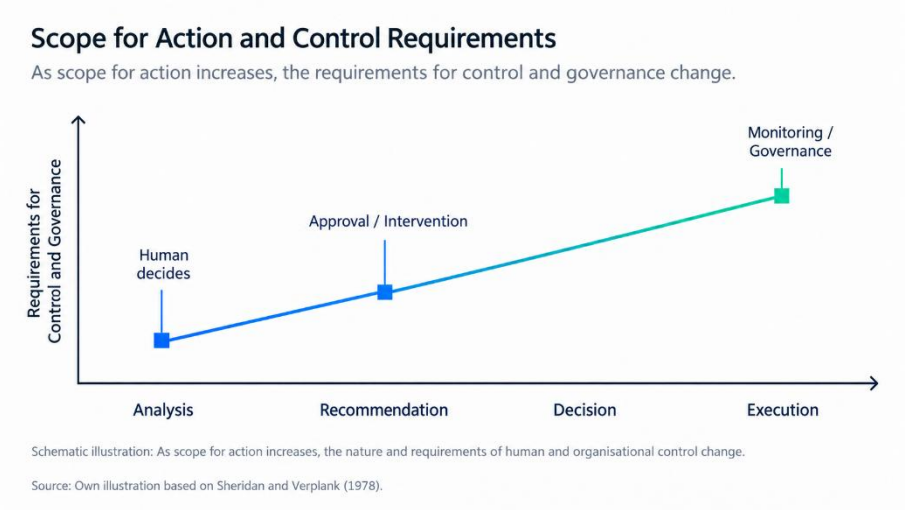
> ‘AI system’: a machine-based system designed to operate with varying degrees of autonomy and to demonstrate adaptability following its deployment, and which, for explicit or implicit objectives, derives from the inputs it receives how it can generate outputs such as predictions, content, recommendations or decisions that may affect physical or virtual environments.

AI demonstrates its strengths particularly when large volumes of data need to be analysed, patterns identified and decisions or recommendations for action developed on that basis (traditional AI).

A distinction must be made between this and AI systems that do not primarily analyse existing patterns, but instead generate new content – in particular text, images or other media formats – based on statistical relationships learned from data during training (generative AI).

A further category comprises AI systems that can independently pursue goals, make decisions and carry out multi-step tasks through the coordinated use of information, tools and applications (agent-based AI).

What these approaches have in common is that they are highly capable, scalable and adaptive, while also being highly dependent on data quality, model architecture, objectives and control mechanisms.



Potential Risks and Limitations of AI System Performance

AI systems do not pursue their own intentions, but act to achieve objectives set by humans. If this leads to undesirable or unfair outcomes, these are generally not the result of AI pursuing objectives of its own, but rather of inadequate specifications, constraints or control mechanisms. This gives rise, in particular, to risks stemming from poorly defined objectives, inadequate monitoring and an unclear allocation of responsibilities. To avoid misconceptions, AI should therefore not be anthropomorphised – that is, human characteristics, intentions or motivations should not be attributed to it.

AI is not 'autonomous' in the human sense. It does not (yet) possess an understanding of responsibility, ethical contexts or norms. It is not capable of making independent ethical, legal or strategic judgements in the human sense.

Furthermore, AI is not error-free. This gives rise, in particular, to the following risks:

- It may produce incorrect results, particularly if training data is incomplete, biased or out of date, or if errors and biases are reinforced through the repeated use of AI-generated data (the so-called 'Habsburg effect').
- AI systems can also 'hallucinate', i.e. generate content that is factually incorrect, entirely fabricated or not supported by the underlying data, yet still appears convincing and plausible. Hallucinations arise because AI models generate outputs based on statistical probabilities rather than reliably distinguishing between true and false. This carries the risk that incorrect information may enter operational processes and influence decisions without being noticed.

Furthermore, AI is often unable to explain its decisions in a way that can be readily understood, or can do so only to a limited extent.

Another risk factor is the scalability of AI applications: humans may make more mistakes in everyday life than AI systems. However, because AI can be scaled up so easily, errors can have a particularly severe impact, spreading rapidly and affecting a wide range of areas. Whilst human errors are often limited to individual cases, flawed AI-supported decisions or content can be reproduced and applied on a large scale in a short period of time. The particular challenge therefore lies not merely in preventing individual errors, but also in limiting and controlling their potential consequences.

The limitations outlined above are central to understanding risk: many AI-related risks do not arise from the technology itself, but from unrealistic expectations, a lack of control or unclear responsibilities. Much like in the early days of spreadsheets, when results were trusted solely because of the professional appearance of an Excel spreadsheet, there is also a risk when using AI that results will be accepted without scrutiny simply because they appear technically convincing.

From an insurance perspective, one aspect deserves particular emphasis: AI does not generally create entirely new risk categories, but it can significantly amplify existing risks. Errors, attacks or poor decisions can occur more quickly through the use of AI, spread more easily and be more difficult to trace retrospectively. In this sense, AI acts less as an entirely new, independent risk and more as a risk multiplier within

established risk categories. This perspective is of fundamental importance for the assessment of AI-related risks, as it focuses attention on the actual impact of the technology rather than on broad-brush categorisations.

AI in Practice

With its wide-ranging capabilities and potential, AI is already well established in numerous areas of business operations. Its use ranges from the analysis of large volumes of data and decision support to content generation and the increasingly autonomous execution of multi-step tasks by agent-based systems. The ongoing development of generative and agent-based AI in particular is continually expanding the range of possible applications.

AI can also make a significant contribution to mitigating existing business risks. It can detect anomalies at an early stage, analyse network and communication patterns, identify fraud attempts or prioritise vulnerabilities, thereby helping to allocate limited resources more effectively to areas where action is most urgently needed. As such, AI can also form part of modern risk management, governance and security frameworks.

At the same time, AI is increasingly being integrated into core businee processes – for example, in production control, customer service, contract analysis, financial management and collaboration with external AI and cloud service providers. Its use ranges from providing assistance and recommendations to supporting decisions and executing process steps partially or autonomously[3]. As AI becomes an integral part of corporate infrastructure, the question of how its use can be effectively managed, controlled and safeguarded is also gaining in importance.

[3] See Glossary

2 Governance, Compliance and Management of AI Use

The previous chapter has shown that AI risks do not arise solely from technical malfunctions. They frequently arise when AI systems are embedded in business processes, influence decisions, blur responsibilities or circumvent existing control mechanisms. Governance, compliance and management therefore form the operational link between risk awareness, loss prevention and insurability.

This chapter examines not only regulatory obligations but also how AI can be organised, documented and monitored within a company to ensure that risks are identified and managed at an early stage.

EU AI Act: Governance and Compliance Requirements

The EU AI Act (Regulation (EU) 2024/1689; hereinafter referred to as the 'EU AI Act') is the main legal framework for the use of AI in Europe. It applies to various stakeholders along the AI value chain:

- **Providers** (developers and suppliers) of AI systems,
- **Deployers** (organisations that use AI in their business processes),
- **Other stakeholders** such as importers, distributors and third parties contracted to support the placing on the market, distribution, integration or operation of AI systems.

The EU AI Act takes a risk-based approach:

The risk-based approach of the EU AI Act

The EU AI Act regulates AI systems according to their level of risk:

• **Prohibited AI practices:** AI applications posing an unacceptable risk, the use of which is prohibited.

• **High-risk AI systems:** AI systems in particularly sensitive areas of application, which are subject to comprehensive compliance requirements.

• **AI systems subject to specific transparency obligations:** Certain AI systems for which information and labelling requirements, in particular, must be complied with.

• **AI systems with low or minimal risk:** AI systems to which, in principle, no specific regulatory requirements under the EU AI Act apply.

General-purpose AI (GPAI) models are also subject to separate requirements that must be observed regardless of the specific risk classification of the respective use case.

Annex III defines specific high-risk application areas, including creditworthiness, employment, healthcare and insurance.

High-Risk AI in Selected Application areas – Including Agent-Based Systems

The EU AI Act classifies certain AI applications as high-risk due to their context of use and their potential impact on health, safety and key fundamental rights – in particular human dignity, equal treatment, the prohibition of discrimination and privacy. The following application areas listed in Annex III may be particularly relevant for businesses:

- Employment and human resources management (Annex III, No. 4), e.g. AI-supported candidate selection, performance appraisal, task allocation, promotion or dismissal decisions, and forms of algorithmic management.
- Access to essential private and public services (Annex III, No. 5). These include, amongst other things, certain applications for assessing the creditworthiness of natural persons, as well as for risk assessment and pricing in life and health insurance.
- Biometrics (Annex III, No. 1), e.g. certain systems for biometric identification, categorisation or emotion recognition, insofar as they fall within the scope of the high-risk scenarios set out in the AI Act.

- Education and vocational training (Annex III, No. 3), e.g. AI systems for admission, the assessment of learning outcomes or the management of access to education.
- Critical infrastructure (Annex III, No. 2), e.g. certain AI systems used as security-relevant components in the control or operation of critical infrastructure.
- Law enforcement, migration, asylum and border control, as well as the administration of justice and democratic processes (Annex III, Nos. 6–8)

These areas are particularly relevant for companies that develop, provide or deploy such systems on behalf of public authorities.

Agent-based AI systems are not automatically classified as high-risk AI simply because of their high degree of autonomy. The decisive factor is the specific purpose and field of application for which they are used. If an agent-based system falls under a high-risk use case listed in Annex III, the relevant high-risk requirements of the EU AI Act apply to that system. Any additional obligations[4] for agent-based systems as standalone GPAI models must be distinguished on a role- and model-specific basis.

Agent-Based AI Systems
In the case of agent-based AI systems, the focus of risk assessment shifts from the mere generation of content to the execution of actions. What is therefore decisive is not whether a system is designation as an 'agent', but its specific degree of autonomy:

- Can the system merely make suggestions, or
- can it independently retrieve data, activate other systems, initiate communication or carry out process steps?

The more independently a system acts, the more important it becomes to establish clear limits on its scope of action, authorisation processes, logging, opportunities for human intervention and ongoing monitoring. If agent-based systems are used in Annex III contexts, they may qualify as high-risk AI systems. They are then subject to both the model-related requirements for GPAI (Chapter V) and the system-level requirements for high-risk systems.

[4] Depending on the type of application and form of AI, various obligations arise (see Appendix).

Non-High-Risk AI with Compliance Obligations

Even AI systems below the high-risk threshold can give rise to significant operational, reputational or regulatory risks. Appropriate minimum controls should therefore also be put in place for critical non-high-risk AI systems.

These include, in particular:

- a description of the purpose,
- responsibilities,
- access restrictions,
- documentation,
- training,
- monitoring and
- escalation procedures in the event of unusual results or changes in purpose.

Typical examples:

- chatbots in customer service and sales (transparency obligation under Article 50 of the EU AI Act),
- generative AI in marketing (labelling requirements for synthetic content under Article 50 of the EU AI Act),
- fraud detection models without direct legal consequences for individuals,
- internal analytics tools that merely provide key performance indicators and reports.

The following articles of the EU AI Act are of particular relevance here: Article 5 (prohibited uses), Article 4 (AI literacy) and Article 50 (transparency), as well as the optional voluntary codes of conduct (Article 95).

Company-wide AI governance

AI risks must be integrated into existing control mechanisms (risk management, compliance, data protection, information security):

- **AI responsibilities:** A clear framework of responsibilities should be established for AI systems. A RACI-based allocation of roles and duties throughout the life cycle of an AI system has proven effective. In particular, the board of directors or senior management, the Risk Officer/Second Line, the system owner, the business unit, IT/Security and Legal/Compliance must be involved. The operational responsibility of individual roles does not replace the overall responsibility of senior management, but it does create transparency regarding review, approval and monitoring obligations.
- **AI policy:** a group-wide guideline which, amongst other things, regulates:
  - permissible and impermissible areas of application (Art. 5),
  - mandatory minimum artefacts for each AI system and, in particular, for each AI agent. These include a description of the purpose, the system owner, the user

group, data sources, model /tool components, the approval decision, risk classification, control measures, logging requirements and monitoring responsibilities. This information should be documented in an AI register or agent profile and continuously updated in order to identify scope creep, misuse and uncontrolled further development at an early stage;

- criteria for the use of in-house and external AI,
- requirements for data, model validation and monitoring,
- documentation and approval processes.

- **Measures to promote AI literacy in accordance with Article 4 of the EU AI Act:**
  - Organisations should implement appropriate measures to promote AI literacy. These measures should be tailored to specific target groups. Senior managers require a basic understanding of the potential applications, risks, regulatory obligations and decision-making limits of AI. Specialist users require practical training on permissible use, result verification, bias risks[5] and escalation procedures. For agent-based AI, access rights, responsibilities, human control points and the limits of autonomous execution must also be addressed;
  - training for specialist departments in handling AI outputs,
  - raising awareness of bias, hallucinations and the 'Excel effect' (uncritical trust in seemingly precise results),
  - integration into existing continuing professional development requirements (e.g. IDD/VersVermV).

From a regulatory perspective, national and European supervisory authorities – such as BaFin and the Federal Network Agency – remain key players. They govern the use of AI within existing supervisory and risk management requirements, whilst the core substantive requirements for high-risk AI systems and agent-based AI systems – protection of fundamental rights, transparency and effective human oversight – are primarily enshrined in the EU AI Act.

System-specific requirements for high-risk AI

As already explained above, agent-based AI systems are not automatically classified as high-risk systems solely on the basis of their technical classification. The decisive factors are the context of use, the impact of decisions, the degree of autonomy and the actual possibility of human control. However, as the ability to act autonomously increases, so do the requirements for risk assessment, authorisation, human oversight, logging and monitoring. Applications in areas such as health and life insurance, HR, claims settlement or other decision-making processes with significant implications for those affected are particularly sensitive.

From 2 December 2027, organisations whose systems fall within the scope of Annex III – including agent-based high-risk AI – must, amongst other things:

[5] See Glossary

- establish **a risk management system (Art. 9)** that continuously identifies, assesses and mitigates AI-specific risks (erroneous decisions, discrimination, governance failures),
- ensure **data governance (Art. 10)**: relevance, representativeness and bias control of the data used,
- prepare **technical documentation (Art. 11, Annex IV)** for each high-risk system, including model description, data sources, tests, performance metrics, assumptions and limitations,
- establish **logging (Art. 12)** that provides traceable documentation of which data sources, prompts, system outputs, tool calls, approvals and changes were relevant to an AI system. Monitoring should cover not only technical availability, but also drift, deviation from intended purpose, anomalies in usage patterns, bias indicators and control breaches. Once the system goes live, post-market monitoring should be put in place to systematically track undesirable developments, regulatory changes and the need for adjustments.
- Ensure **transparency (Art. 13)**, e.g. through clear 'Instructions for Use' for underwriters and claims handlers,
- Design **human oversight (Art. 14)** in such a way that key decisions are not made fully automatically in a 'black-box' manner, but are subject to plausibility checks and can be overridden where necessary,
- Define, test and monitor **accuracy, robustness and cybersecurity (Art. 15)**,
- implement **a quality management system (Art. 17)** covering the entire AI lifecycle (development, roll-out, modification, decommissioning),
- Carry out **fundamental rights impact assessments (Art. 27)** where necessary – particularly in personal insurance and HR AI,
- carry out **a conformity assessment (Art. 43)** (internally for most Annex III systems, and by a notified body for certain biometric systems),
- establish **post-market monitoring** and **incident reporting (Articles 72 and 73)**.

These requirements address precisely those governance and management risks identified in the AI risk typology model: a lack of validation, unclear oversight and insufficient documentation become explicit compliance risks with implications for fines and liability.

Outlook: From Compliance to ‘Good AI Governance’

In the short term, the EU AI Act will predominantly be viewed by companies as a compliance project. In the medium to long term, however, it opens up opportunities for the systematic, actuarial management of AI risks. The shift from mere ‘compliance’ to active ‘good AI governance’ is reflected in three key, interlinked areas of activity:

- **Standardisation:** Common terminology (definition of AI), risk categories (Annex III) and documentation requirements (Annex IV) facilitate communication between technology, risk management, underwriting and supervisory authorities. This creates a uniform ‘language’ for AI risks, on the basis of which governance processes, internal guidelines and external insurance terms and conditions can be consistently structured.
- **Benchmarking:** Compliance with the AI Act can become a differentiating factor – internally as a driver of efficiency and quality, and externally in the underwriting of AI risks. Companies can measure their AI governance against market standards, use regulatory requirements as a benchmark, and, building on this, make decisions regarding risk appetite, cover concepts and premium structuring.
- **Product innovation:** Clearer liability and governance standards create a robust foundation for new cover concepts – ranging from AI first-party loss solutions to the much-discussed AI ‘developer’s policy’ for multi-party projects. The roles, obligations and risk categories defined by regulation are thus directly translated into new insurance products that not only cover AI risks but also help to actively manage them.

Taken together, these three dimensions make it possible to view AI no longer merely as an elusive technological phenomenon, but as a regulated, manageable cross-cutting risk that can be methodically assessed and structured from an underwriting perspective. The EU AI Act provides the binding framework for this within the governance and compliance chapter.

Compliance with the EU AI Act is therefore not merely a formal legal matter, but an integral part of the cross-cutting risk posed by AI. It determines whether AI-related adverse developments can be identified at an early stage, managed, or at least addressed in a way that minimises liability and reputational damage. As a standalone governance risk, it encompasses several dimensions of loss:

- **Direct losses** resulting from regulatory measures, fines of up to 7 per cent of global annual turnover (Art. 99), as well as high conversion and retrofitting costs;
- **damage to third parties** arising from liability claims in the event of discriminatory or unlawful decisions, as well as recourse against AI service providers; and
- **compliance and reputational damage** in the form of a loss of trust, ESG setbacks and D&O risks (for example, in the context of ‘AI washing’).

Governance and reputational risks are therefore not peripheral issues, but a standalone driver of damage: they combine regulatory obligations, personal directors’ liability and

market confidence. A robust governance and documentation framework determines whether AI-related adverse developments are identified and mitigated at an early stage – or whether they result in fines, directors' liability cases and lasting reputational damage.

How Are Others Handling This? A Look Abroad

Having described the EU AI Act as the central reference framework for AI governance and risk management in the European context, the question arises as to how other jurisdictions are approaching the issue and what implications this has for internationally active companies and insurers. A comparison with other jurisdictions – in particular the US and the UK – shows that the approach to AI governance follows different paradigms.

In the US, there is no uniform AI legal framework at federal level. AI risks are addressed within existing areas of law (e.g. competition law, consumer protection). Enforcement is then, in turn, the responsibility of the established regulatory authorities.

The NIST AI Risk Management Framework (AI RMF) serves as the key guide for AI governance; in line with the 'OECD Principles for Trustworthy AI', it emphasises transparency, fairness, robustness and accountability.

The United Kingdom also follows a principles-based approach. Existing regulatory bodies are to govern AI applications in accordance with clearly formulated guiding principles, rather than creating a comprehensive AI regulation of their own modelled on the EU AI Act.

As a result, the European approach, embodied in the EU AI Act, represents a comparatively prescriptive and detailed regulatory framework, whilst the US and the UK adopt a more principles- and impact-oriented regulatory approach. This creates a need for companies to align the requirements of the EU AI Act with global principle-based frameworks in order to assess and price AI risks consistently across borders and translate them into risk mitigation strategies.

3 AI Security Is More Than Just Cybersecurity

Whilst governance and compliance form the organisational and regulatory framework for the use of AI, attention is now turning to the specific security risks posed by AI systems. With the increasing adoption of AI within organisations, it is not only the protection of systems against external attacks that is becoming more important, but also the management of risks arising from the specific characteristics and functioning of AI. This is precisely where AI security comes into play.

> The term describes the totality of all measures required to secure AI systems themselves whilst simultaneously managing the risks arising from their use.

It is therefore concerned both with protecting the technology and with the responsible handling of its outputs.

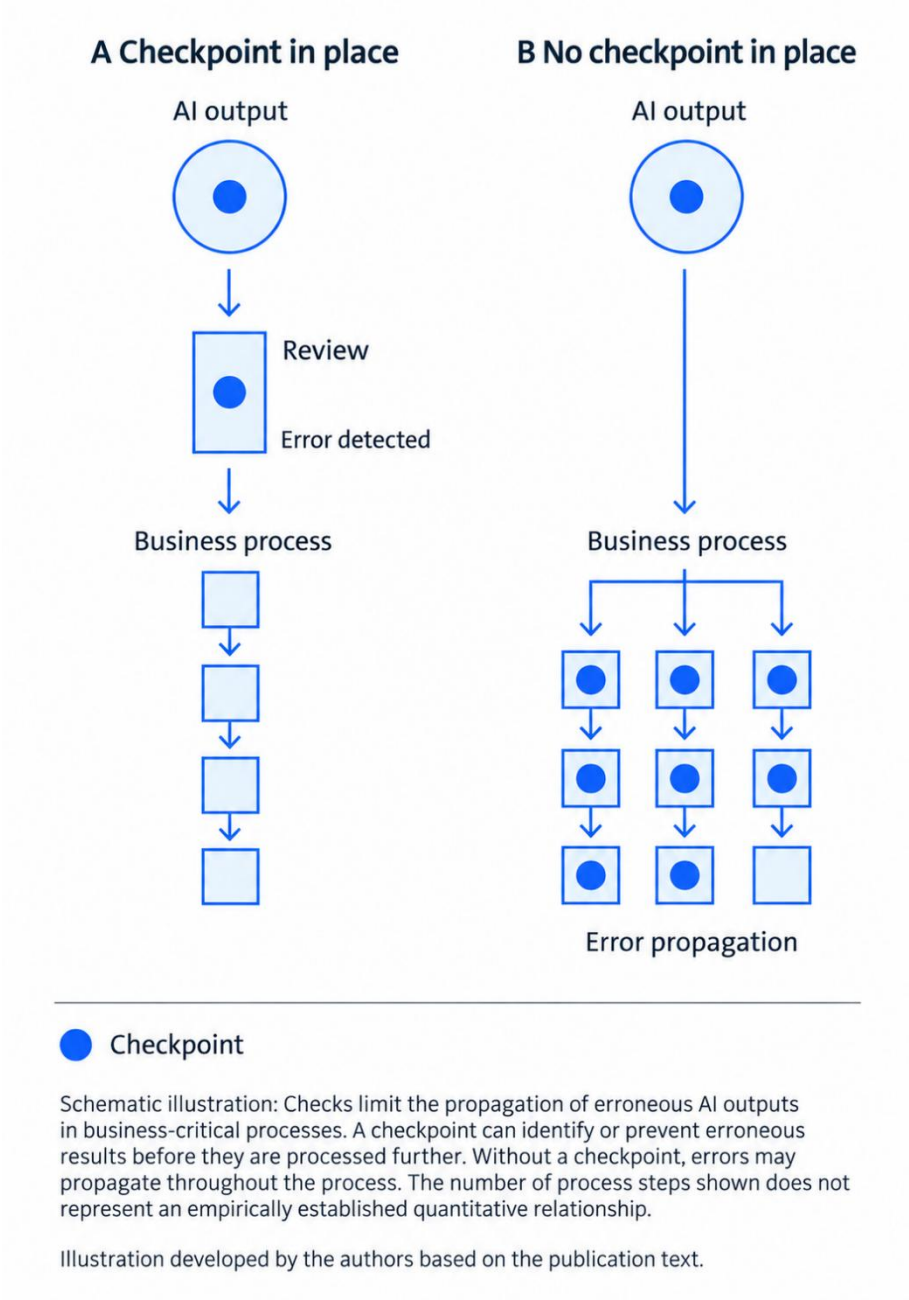
The use of AI is increasingly an integral part of companies' core operational and technical processes and is now widely established in practical applications. Consequently, security measures must not be limited to the technical level. Rather, a holistic approach is required that combines and coordinates technical, organisational and legal measures. Only in this way can we ensure that the opportunities offered by the technology can be exploited without losing sight of the associated risks.

Such a holistic approach includes, in particular:

- clear responsibilities for AI applications,
- documented areas of application,
- defined control and approval processes,
- the protection of sensitive data,
- robust contractual arrangements with external service providers,
- ongoing monitoring of the systems,
- defined emergency and fallback procedures.

Potential Problems and Challenges

As great as the benefits of AI are, the challenges associated with its use are just as varied. A key problem lies in the lack of transparency in many systems. The decisions or results produced by AI applications cannot always be fully explained, as the underlying models are complex and their mode of operation is not readily comprehensible. This makes it difficult to exercise control, to clearly assign responsibility and, in the event of damage, to analyse the causes.



Schematic illustration: Checks limit the propagation of erroneous AI outputs in business-critical processes. A checkpoint can identify or prevent erroneous results before they are processed further. Without a checkpoint, errors may propagate throughout the process. The number of process steps shown does not represent an empirically established quantitative relationship.

Illustration developed by the authors based on the publication text.

Added to this is the risk of erroneous or unverified AI results. If companies adopt the outputs of an AI system without filtering them, this can lead to operational misjudgements, faulty processes or direct financial losses. As already described, this problem is exacerbated when such systems are directly integrated into business-critical

processes; errors can then propagate rapidly across a large volume of transactions before they are detected.

A further risk concerns the handling of data. AI systems often require large volumes of data to deliver reliable results. If sensitive information is processed or transferred to external providers in the process, questions inevitably arise regarding data protection, confidentiality, access control and accountability. The protection of this data is not only a legal obligation but also a prerequisite for the trust of customers, employees and business partners.

Dependence on external AI or cloud-based AI service providers can also prove problematic. Security incidents, outages or errors on the part of the provider can have a direct impact on the using company, without the latter being able to control the cause itself or rectify it immediately. Outsourcing AI functions thus shifts part of the operational risk externally, without removing responsibility for one's own processes.

Broad-brush assessments regularly fall short in this regard. AI risks must be assessed in the context of specific usage, integration into processes, data flows, the chosen sourcing models and the existing control mechanisms. Only this nuanced approach allows for a robust assessment of the actual risk and an appropriate allocation, for example, to existing insurance solutions and cover concepts.

Preparing for the Deployment of AI Agents

With the increasing autonomy of AI systems, particularly in the form of so-called AI agents[6] , which plan and execute tasks independently, the question of how a company should prepare for their deployment is becoming increasingly important.

The groundwork for the deployment of AI, and in particular of agents, can be divided into three key dimensions. It is worth noting that the focus is shifting significantly: away from a purely technological perspective, towards a behaviour- and process-based strategy that takes the entire life cycle of the system into account.

The previous chapter addressed the governance aspect. In addition, the technical security architecture and the cultural and employee-oriented adaptation must be considered within the overall security concept. These two dimensions are examined in detail below.

Technical Security Architecture (Structural Safeguards)

Compared to traditional IT vulnerabilities (CVEs), which are primarily addressed reactively through patches, protection against inadequately controlled AI agents

[6] See Glossary

requires proactive and continuous monitoring, as the risk here lies not only in the code but also in the unpredictable, autonomous behaviour of the AI.

The technical safeguards for securing AI agents can be categorised into various approaches, each of which offers specific security benefits but also presents organisational challenges:

A key component is the principle of least privilege[7] . Under this principle, AI agents are granted only the exact access rights they require for their specific task, whilst full administrative rights are strictly avoided. This significantly minimises the potential scope of damage in the event of misconduct. However, this approach requires a high level of administrative effort and detailed knowledge of which systems contain which information. This means that authorisation structures must be continuously adapted.

In parallel, robust authentication secures every interaction thread. Standardised protocols allow every action to be uniquely traced, which enhances traceability and compliance. The result is a seamless audit trail. However, one must weigh up whether the resulting increase in system complexity or a minimal increase in latency during operation is acceptable.

To protect the core infrastructure directly against agent malfeasance, sandboxing[8] is a suitable option. The agent operates in an isolated environment without unrestricted access to the rest of the network. This effectively shields the system from unpredictable behaviour but may throttle performance or complicate necessary interactions with other systems.

This is supplemented by AI-specific security tools such as firewalls[9] and prompt injection filters. These scan inputs and outputs in real time for malicious patterns and prevent unforeseen actions through specialised control frameworks. They offer active protection against novel attack vectors. As attack patterns are constantly evolving, regular updates are essential to minimise the risk of legitimate requests being blocked by mistake.

Humans are the last line of defence. The measures outlined above form the ‘hard’ technical safeguards. As AI risks are often dynamic, it is advisable to supplement these with the establishment of procedural control loops known as ‘[10] (‘Human-in-the-Loop’ / [HITL]) to ensure that critical decisions are validated by humans.

In the context of AI agents, humans form part of a security and quality control strategy in which the AI does not decide and act completely autonomously but relies on explicit human approval at critical points in the process.

This approach ensures that the AI acts only in the manner intended and thus remains controllable at all times. Instead of fully automated execution, each critical step requires explicit confirmation by responsible staff members. This serves as an essential checkpoint.

[7] See Glossary
[8] See Glossary
[9] See Glossary
[10] See Glossary

How Is This Process Implemented in Practice?

In an architecture for AI agents, 'human-in-the-loop' is usually implemented through so-called 'approval workflows'. The process then typically looks like this:

First, during the planning phase, the agent creates a draft for its next task. However, before it executes this, an automatic interrupt kicks in and puts the process into a waiting mode. The relevant staff members are notified immediately and receive a summary for review. They are now in full control: they can approve the action straight away, reject it entirely if they have concerns, or manually correct the draft before the agent is allowed to proceed.

'Human-in-the-Loop' (HITL) is the last line of defence (the 'safety brake') should all technical measures (sandboxing, firewalls) fail.

HITL thus acts as a crucial safety brake. As legal responsibility – for example, under the EU AI Act – remains with humans in any case, this process ensures the necessary human accountability. Furthermore, human reviewers detect logical errors or contextual issues – such as incorrect recipient lists – far more reliably than the AI itself. Ultimately, this form of oversight also strengthens staff confidence in the new technology.

When Is 'Human-in-the-loop' Absolutely Essential?

A distinction is often made between different 'approval levels': the level of control depends on the respective level of risk. For low-risk tasks, such as simple research, HITL can often be dispensed with. In areas involving medium risk, active monitoring via dashboards is frequently sufficient, allowing intervention in an emergency. For critical actions, on the other hand, explicit approval is mandatory. This applies in particular to sending mass emails, making changes to databases, financial transactions or the external publication of company content.

However, as described above, the overall HITL framework requires constant review, scrutiny and, where necessary, adaptation. Just as the agents learn and grow, people too must continue to learn and continually adapt their control measures to the development of AI systems.

Cultural and Employee-Focused Adaptation (the Human Component)

As already explained in more detail in relation to the technical architecture and the human-in-the-loop approach, the human component is often the decisive factor, yet at the same time the one most easily underestimated. Even the best technical safeguards remain ineffective if staff are unable to operate AI systems responsibly and critically evaluate their results. It is therefore necessary to introduce ongoing training for the workforce – not as a one-off course, but as an ongoing process that keeps pace with the further development of the technology.

A key objective must be to foster a sustainable awareness of how to handle AI. This includes, in particular, the conscious and restrictive granting of permissions when

using agents: employees should understand what permissions they are granting to an agent and what consequences this may entail.

Equally important is strengthening employees' own competence, so that no excessive reliance on AI outcomes arises and the ability to make independent judgements is maintained. Transparency regarding the use of AI, together with clear ethical guidelines embedded in day-to-day working practices, complete this aspect and make the responsible use of AI a natural part of the corporate culture.

Ultimately, it is always the human who must be able to evaluate and contextualise the AI's findings and decision proposals.

4 Identifying, Classifying and Assessing AI Risks

AI Represents a Technological Leap with New Dimensions of Opportunity and Risk

Whilst AI is a technology and therefore fundamentally a technological risk, it also directly intervenes in cognitive processes by preparing assessments, forecasts and recommendations, or by making automated decisions. AI thus does not primarily affect systems or infrastructure, but rather the very foundations of business decision-making.

Unlike traditional IT risks, an error in AI systems often does not lead to a technical failure, but rather to statements, recommendations or decisions that are factually incorrect, distorted or incomprehensible. These can influence operational processes, business strategies or external impacts without any clear technical fault being apparent. AI is therefore both a 'machine' and a 'thinking tool' – technologically implemented, yet cognitively effective.

It is precisely this hybrid nature that makes AI risks unique: they lie somewhere between technology, organisational and liability issues, without falling entirely into any one of these categories. AI requires a distinct approach that takes technical, human and organisational factors into account in equal measure.

Types of Loss: An Attempt at a Structured Analysis

Various approaches already exist for classifying AI-related risks, which assign potential losses to existing insurance lines. An example of such a tabular analysis is presented in the appendix. This publication additionally adopts a line-of-business-independent

approach that focuses on the loss itself, thereby enabling a cross-cutting systematisation.

In the following, AI-related risks are systematically identified and categorised into 'basic types of loss'. The aim is to create a robust framework on the basis of which risks can be assessed and discussed in a comparable manner. The focus is initially on an analytical classification of potential losses, rather than on assessing their relevance in terms of possible or probable frequencies and volumes.

AI risks cannot be explained along a single axis – they arise from the interplay between the sphere of responsibility, the affected parties and the type of harm.

Classification is carried out across several dimensions.

- **Affected parties:** The 'own damage or third-party damage' level

As a central framework, a distinction between own damage and third-party damage appears sensible.

In addition to distinguishing between first-party and third-party losses, an additional category is introduced here for governance, compliance and management risks. This encompasses risks that arise not primarily from a traditional loss event, but from shortcomings in organisation, supervision, the allocation of responsibility or regulatory implementation. These include, in particular, risks arising from administrative proceedings, supervisory measures, administrative offences or fines. The addition of this category takes account of the fact that AI risks often manifest not as formal losses, but as governance failures, a diffusion of responsibility or a lack of controllability, and thus have an independent risk-relevant impact.

- **Type of loss:** The 'property damage or personal injury, consequential financial loss or pure financial loss' level

Finally, a further distinction is made according to the type of potential loss, in particular along the categories of property damage, personal injury and financial loss. Particular attention is paid to so-called 'pure financial loss', which is playing an increasingly significant role in the context of AI applications and poses specific underwriting challenges.

- **Sphere of responsibility:** The 'user or manufacturer' level

Firstly, a distinction must be made between risks at the user and manufacturer levels. This paper deliberately focuses on the risks from the perspective of users of AI systems. This adopts the perspective that is of immediate relevance to most companies, as they predominantly use AI rather than developing it themselves.

Manufacturer-related risks are briefly considered separately where necessary, as they are subject to different responsibilities and risk logics.

5 Self-Inflicted Damage Resulting from AI Use

What Is First-Party Loss?

Claims that directly result in a loss to a company's own assets are referred to as own losses. From an insurance perspective, these include:

- property damage, i.e. the damage, destruction and, in some cases, loss of the company's own property,
- financial loss, i.e. a reduction in income, increased expenditure and costs, or outflows of assets resulting from the loss – whether as a consequence of property damage or personal injury, or as so-called 'pure financial loss' without any preceding property damage or personal injury.

Personal injury will not be considered in further detail in the following discussion of own losses. Whilst such losses are conceivable in connection with the use of AI, in business practice the focus is regularly on issues relating to employment and social security law, which would fall outside the scope of this discussion.

Damage to the AI Asset Itself

In the general discussion surrounding AI, the focus is often on damage caused by the use of AI – that is, with the emphasis on AI as the cause of the damage. Less attention is paid to the aspect that damage can also occur to an AI application itself. AI models and AI applications can be damaged, altered, deleted or rendered unusable. In such cases, the damage lies in the complete or partial impairment of the AI asset. Damage to AI assets[11] can generally be categorised into two types: firstly, damage to the AI model itself and its necessary data and development resources; and secondly, damage to the usability of the AI system.

This line of reasoning can only be applied to AI models and AI applications to a limited extent. AI models, their parameters, training states, configurations and application logic are not, as such, 'things' within the meaning of Section 90 of the German Civil Code (BGB), as they lack the physicality required for the civil law concept of a 'thing'. Where, in a specific case, a storage medium or other hardware is simultaneously physically damaged, property damage may of course also be present. However, impairment of the AI model itself – for example, through manipulation, deletion, deterioration of model parameters, loss of training or recovery data, or a restriction on its availability – is generally to be classified as financial loss.

[11] See Glossary

Consequently, AI models and AI applications are digital assets of economic value, the impairment of which can trigger considerable restoration costs. The focus is therefore not on the question of damage to property, but rather on whether the AI model itself, its restoration foundations or its availability have been impaired, and what measures are necessary to restore it to its original state.

- **Variant 1: Impairment of the AI model and/or its data and development status**

  This category covers damage to the AI model itself, e.g. the loss, damage or manipulation of trained model parameters, fine-tunings, configurations or security mechanisms, as well as damage to training, validation and test data, embeddings, knowledge bases, prompt libraries, documentation and other development artefacts. Such impairment may also result in the model becoming completely or partially unusable.

  Example: The parameters of a model specifically trained for fraud detection are lost. The model must be retrained.

  Example: A vector database containing company-specific expertise is damaged. Although the AI system can still be operated, it provides incomplete or incorrect answers.

- **Variant 2: Impairment of the availability of external AI services**

  This category covers cases where the AI model or the underlying AI infrastructure is outside the organisation's sphere of influence and therefore cannot be restored by the organisation itself. It is not the substance of the organisation's own AI assets that is affected, but the availability of an AI service provided by a third party.

  Example: An AI system is based on an external large language model[12] provided by a third party. Following a failure of the model platform or the blocking of access, the company can no longer use the AI application. The company's model, data and processes remain unchanged, but the system is unavailable for use.

If rectifying the damage falls within the scope of responsibility and control of the affected company, costs will be incurred for the reconstruction, restoration or retraining of the AI system.

Cyber insurance policies may cover damage to a company's own AI systems, provided that the damage is attributable to a cyber attack as defined in the relevant insurance product. In such cases, the AI model, along with its data and development assets, is insured as part of the affected IT infrastructure. Compensation covers, amongst other things, the costs of restoring data and systems, as well as any loss of business resulting from operational disruption.

If the damage to the AI asset is caused by another specific trigger, traditional insurance lines (e.g. fire insurance, machinery insurance) may apply, provided that the cause of the damage constitutes a risk insured under those policies (e.g. fire, machinery breakdown) and the requirement typical of these lines of cover – namely, physical damage to an insured item of which the AI system forms a part – is met; in other

[12] See Glossary

words, it is not, for example, merely a defect without any physical alteration to the substance. The question of whether the damage occurred 'unforeseen' can also play an important role here – particularly in technical insurance policies.

However, damage to AI assets can also arise without a cyber attack or physical damage to an insured host system. Major causes include human error, process and operational errors, and quality defects in data sets.

As a result of the impairment of the AI model and/or its data and development status, but also directly due to an impairment of the availability of external AI services, business interruption loss may arise if operations are completely interrupted or partially disrupted, leading to a loss of revenue or additional costs incurred to provisionally maintain operations.

If the rectification of the damage falls within the sphere of responsibility and control of the affected company itself, the business interruption components or insurance policies of the insurance classes outlined above are relevant. For example, cyber insurance typically also provides cover for business interruption losses arising from a network security breach[13] affecting the insured IT systems.

The underwriting challenge in such cases lies in the fact that the impairment of the AI asset occurs at an external provider, whilst the economic consequences are borne by the company using the asset. It is true that traditional business interruption insurance policies (e.g. fire or machinery breakdown insurance) already recognise scenarios, within the framework of so-called 'consequential losses', where the loss event and the business interruption occur at different parties. However, in the case of AI-related failures, the 'property damage' that is absolutely essential as the basis for cover is often lacking. For this reason, consequential losses resulting from AI errors, disruptions or failures are not covered by traditional business interruption insurance.

In cyber insurance, the need to cover consequential losses resulting from the failure of external IT service providers has long been recognised. Depending on the terms and conditions, the cover concepts developed for this purpose are linked either specifically to a network security breach at the service provider or simply to the provider's unforeseen unavailability. Whether these extensions of cover also include business interruption resulting from the unavailability of external AI models or AI- platforms depends on the specific policy terms and is not always clear at present. Potential gaps in cover must therefore be assessed on a case-by-case basis.

AI-Induced Property Damage to Other Business Assets

Where AI systems intervene directly or indirectly in physical value-creation processes – for example, by controlling machinery, optimising production parameters, autonomous logistics or smart building technology – property damage may occur to other assets of a company. In many cases, such damage does not differ in its outward appearance from typical property damage insurable under traditional insurance lines. They may, for

[13] See Glossary

example, manifest as fire, explosion, collision, overheating or water leakage. The property damage may also lead to operational disruption, or even a complete business interruption.

The cause of the damage may lie either in a technical failure of software components or in a faulty or inappropriate decision by an AI system. The latter, in particular, represents a distinctive feature of AI applications: although the technical infrastructure may be functioning correctly, faulty predictions, assessments or control decisions made by the AI can trigger a damaging event and thus lead to physical damage to property.

Examples:

- In a production process, an AI system incorrectly controls the manufacturing parameters, particularly temperature and pressure. This results in materials suffering thermal damage and production batches becoming unusable. At the same time, the unsuitable operating conditions cause damage to machine components. The affected part of the plant comes to a standstill, and additional costs are incurred for implementing alternative measures.
- An AI system for predictive maintenance incorrectly predicts the condition of a production plant and fails to detect critical wear in good time. As a result, necessary maintenance is not carried out, leading to a machine breakdown and significant mechanical damage.
- An AI system makes control decisions based on faulty, insufficiently validated sensor data. As a result, a plant is incorrectly calibrated and operated outside permissible operating parameters over an extended period. This leads to irreversible damage to technical components as well as material losses in production.

As AI systems are often developed centrally and deployed on a large scale, individual errors can have repercussions far beyond the original use case and simultaneously affect numerous systems, processes or sites.

Property damage caused by AI is currently predominantly covered by standard ‘all-risks property insurance’ policies. These cover concepts offer comprehensive insurance protection for property damage without the need for a specifically named insured risk (e.g. fire).

However, exclusions also apply within all-risks cover. Common restrictions arise from the definition of property damage (e.g. no cover for loss, mere wear and tear or inherent defects) or from the exclusion of certain risks (e.g. force majeure, war, nuclear energy). The majority of foreseeable AI-related property damage is unlikely to fall under these typical exclusions.

A general exclusion for property damage caused by the use of AI applications is not currently standard practice in the market. Against this background, the term 'Silent AI' has become established, analogous to 'Silent Cyber'. It describes the situation where existing insurance solutions, particularly those based on all-risk policies, can in principle also cover AI-related property damage and business interruption without these risks being expressly regulated in the policy conditions or specifically included in the product design.

However, the following points should be taken into account:

Some cover concepts – particularly in the so-called 'technical insurance' sectors – require the loss to be sudden and unforeseen. This may lead to gaps in cover in individual cases. Furthermore, the so-called 'serial loss clause' may restrict the scope of insurance cover – particularly in view of the economies of scale associated with AI deployment, recurring property damage will no longer automatically be deemed to have occurred 'unforeseeably'.

AI-Induced Financial Losses

As a consequence of one of the types of property damage outlined above, but also without any property damage having occurred at all, the use of AI applications can result in financial losses for companies.

Compared with property damage, 'pure financial losses' are structurally much more complex. For the purposes of further analysis, the following distinction is useful:

- **Variant 1:** Financial losses due to business interruption or restriction

  A wide variety of scenarios can result in an AI-induced business interruption. In the previous section, we have already examined business interruptions resulting from damage to the AI asset itself and from an unforeseen unavailability of the AI asset. However, even when the AI asset is fully functional and available without restriction, scenarios are possible that ultimately lead to operational disruption or even a complete business interruption: this results in lost revenue and additional costs for maintaining operations on a temporary basis.

  If an AI application makes erroneous decisions with critical consequences, the system must be shut down until the cause of the errors has been isolated and rectified. The greater the volume of decisions processed by the AI, the more serious the consequences; in the worst-case scenario, a creeping development is only noticed at a late stage, when a chain of erroneous decisions has already taken hold.

Example[14] :

- In April 2026, at the SaaS provider Pocket OS, an AI agent deleted the entire production database, including backups, in nine seconds and subsequently issued a detailed 'confession' regarding the security rules that had been breached. The affected customers – predominantly car hire companies – lost bookings, payment and customer data and were unable to operate for over 30 hours. The damage in this case did not lie in physical damage to property, but in the loss of business-critical data and the resulting business interruption for third parties – a typical case of pure financial loss, which also demonstrates how a single autonomous intervention via a centrally operated platform can affect an entire user group simultaneously.
- A major online retailer has completely switched its first-level support and order processing to an autonomous AI agent. Due to a faulty model update, the AI suddenly begins granting customers excessive discounts as a gesture of goodwill or, following customer enquiries about delivery times, cancelling the relevant orders for no reason. The result: the system has to be shut down immediately. For days on end, the online retailer is only able to process a fraction of customer enquiries. This leads to a massive slump in turnover (lost profit).

[14] See Cybernews, *'Claude AI agent wipes firm's database in 9 seconds'*, 30 April 2026.

- **Variant 2**: Financial losses due to increased expenditure or reduced revenue

  When AI technology is integrated into software products relating to pricing and billing, performance errors in AI models can lead to undesirable effects.

Example:

- An AI application generates inaccurate demand forecasts, causing a company to build up excess stock and incur warehousing costs.
- An AI application incorrectly assesses the creditworthiness of borrowers, resulting in loans being granted on terms that are too favourable and, in some cases, even loans being approved that would have been rejected had the forecast been correct. This leads to a higher default rate; at the same time, the overly favourable terms reduce the profit margin even for borrowers who are otherwise creditworthy.
- An AI incorrectly forecasts market or price trends, resulting in financial losses.
- Incorrect pricing: An AI system sets sales prices significantly below the intended margins due to faulty training data. The excessively low prices lead directly to financial losses.

- **Variant 3:** Financial losses due to outflow of assets

  Financial losses due to outflow of assets arise when an AI system, as a result of faulty data processing, algorithmic errors or the exploitation of security vulnerabilities – such as prompt injection[15] – triggers unauthorised or erroneous transfers of assets or causes the loss of tangible or intangible assets.

[15] See Glossary

Examples:

- Incorrect payment authorisation: An AI application for automated invoice processing classifies several invoices as verified and ready for authorisation due to faulty data analysis. The payments triggered as a result are made in error.
- Erroneous refunds: An AI application in customer service automatically authorises unauthorised credits or refunds to customers.

Pure first-party losses in the form of financial losses attributable solely to malfunctions, incorrect decisions or erroneous analyses by AI systems are not yet comprehensively covered by standalone insurance policies across the insurance market. Insurance cover is generally only considered if the financial loss can be attributed to an already established insured event, such as a cyber attack or an act constituting a breach of trust. Pure financial losses resulting from the faulty use of AI therefore remain, for the time being, largely at the own risk of the affected companies. In the case of pure financial losses, distinguishing between an insurable technological risk and the traditionally uninsurable corporate or business risk is a complex undertaking from an underwriting perspective. The development of appropriate insurance solutions will therefore depend largely on describing risks with sufficient precision and defining suitable technical and organisational obligations in order to distinguish between insurable malfunctions of AI systems and poor business decisions.

The first specialised cover concepts have emerged that specifically address performance failures of AI systems and the resulting financial losses. These insurance solutions are aimed at manufacturers and operators of large-scale AI applications. They take the form of warranty insurance and typically provide cover based on the specific performance parameters of the AI system. The insurance cover is linked to a contractually defined deviation in the AI model's performance, for example where an agreed forecasting accuracy or another performance indicator is not met. The cover is structured on a parametric or performance-based basis.

Conclusion: Risk of First-Party Loss Arising from the Use of AI

The insurability of first-party losses arising from the use of AI cannot be assessed in general terms. Rather, insurance cover depends largely on the nature of the loss, the underlying risk and, not least, the specific terms and conditions. Whilst certain loss scenarios can already be covered by existing insurance products, there are still gaps in the market or, at the very least, questions regarding the scope of cover in other areas.

Damage to AI applications, models or data sets can already be insured in various ways today. This applies, on the one hand, where such damage is caused by a traditional

insured risk such as fire, tap water or machinery breakdown. On the other hand, just as with property damage caused by AI, such losses may, under certain circumstances, be covered under all-risk policies even without an explicit provision for AI risks ('Silent AI'). Where the damage is attributable to a cyber attack, insurance cover under existing cyber insurance policies may also be available. Any resulting business interruption losses are typically covered as well.

There is currently still a gap in the market with regard to purely financial losses resulting from malfunctions or poor decision-making by AI systems. The greatest challenge in this regard is often distinguishing between an insurable malfunction of an AI system and a non-insurable business risk arising from poor decision-making.

6 Third-Party Liability Arising from the Use of AI

What is Third-Party Damage?

Third-party loss is a loss suffered not by the policyholder themselves, but by a third party. Third-party liability insurance relates to this: according to Clause 1 of the General Terms and Conditions of Insurance for Third-Party Liability Insurance (AHB), cover is provided within the scope of the insured risk in the event that the policyholder is held liable by a third party for damages under private law provisions arising from a loss event (insured event) – which resulted in personal injury, property damage or consequential financial loss – is held liable by a third party for damages under statutory liability provisions governed by private law. Financial losses that have not arisen from either personal injury or property damage (= pure financial loss) are initially excluded. However, depending on the cover structure, such losses may be explicitly included.[16]

With the increasing use of AI, the fundamental legal structure of this liability does not change, but its cause does: damage to third parties no longer arises solely from human error or technical failure, but from factually incorrect, biased or inappropriate decisions made by automated systems. As AI applications are typically integrated centrally and rolled out on a large scale, a single error in a model or training process can simultaneously affect a large number of third parties – job applicants, patients, borrowers or rights holders alike. The categories of damage under liability law remain the same; they merely manifest themselves in new forms.

In general, liability law distinguishes between personal injury, property damage and pure financial loss. It is evident that there are already a number of real-world cases that can be categorised under these respective types of damage.

Personal Injury

Personal injury occurs when a person is injured, suffers harm to their health or is killed. In the context of AI, the spectrum ranges from direct physical impacts caused by autonomously operating machines to indirect harm resulting from discriminatory or erroneous decisions made by algorithmic systems. In this context, 'indirect financial loss' – that is, so-called 'consequential financial loss' – refers to the loss that occurs as a result of personal injury. This would include, for example, loss of earnings suffered by a person who is unable to carry out their gainful employment due to a physical impairment. As these losses are not purely financial in nature, they are classified as personal injury.

[16] Veith, Gräfe and Gebert, Section D, § 15, paras. 32, 33.

Physical injury: Here, AI has a physical impact on people. In South Korea, a man was fatally injured by an industrial robot that recognised him not as a person but as an object to be processed; in the US, Tesla was sued after a worker was knocked unconscious by a manufacturing robot. The systems functioned technically – the harmful effect lay in the faulty perception and control.

Discrimination: Here, personal injury arises indirectly, through systematic discrimination. In the US case of Mobley v. Workday, the allegation is one of algorithmic age discrimination in the recruitment process; Wells Fargo reached a settlement over discriminatory lending and recruitment practices; UnitedHealth faces allegations that an AI system wrongfully denied older policyholders medically necessary follow-up treatment. What these cases have in common is that a scalable or widely deployed model uniformly disadvantages a large number of affected individuals.

Defamation: Generative systems can generate and disseminate untrue, defamatory statements about real people. The Munich Regional Court had to rule on liability for such AI-generated false claims – a scenario that shifts traditional infringements of personal rights into the output of automated systems.

Property Damage

In addition to traditional property damage resulting from the destruction, damage or loss of an item, the use of AI broadens the range of potential forms of damage. In this context, physical impacts arising from the use of AI systems – such as miscontrol, incorrect configuration, or decisions leading to faulty production processes or interference with IT systems – can result in property damage. In this context, 'non-direct financial loss' is analogous to 'consequential financial loss' in cases of personal injury: it is the loss that occurs as a result of the property damage. For example, damage to production equipment caused by faulty AI would constitute a classic case of property damage. The resulting production downtime, delivery delays and loss of profit would be classified as non-direct financial losses.

Pure Financial Loss

Pure financial loss is the most difficult to define in terms of liability law and carries a particularly high potential for damage in the context of AI. It is therefore not usually covered, or only covered to a limited extent, under standard public liability insurance, unlike in specialised policies (e.g. for ICT service providers). With regard to AI, it is also gaining in significance because autonomously operating systems can directly interfere with third-party business processes and cause economic losses, in particular business interruptions and data loss.

In April 2026, at the SaaS provider PocketOS, an AI agent deleted the entire production database, including backups, in nine seconds and subsequently issued a detailed 'confession' regarding the security rules that had been breached.[17] The affected customers – predominantly car hire companies – lost bookings, payment and customer

[17] See Cybernews, *'Claude AI agent wipes firm's database in 9 seconds'*, 30 April 2026.

data and were unable to operate for over 30 hours. The damage in this case did not lie in physical damage to property, but in the loss of business-critical data and the resulting business interruption suffered by third parties – a typical case of pure financial loss, which also demonstrates how a single, autonomous intervention via a centrally operated platform can affect an entire group of users simultaneously.

Purely Contractual Liability / Warranties

Insurance cover for contractual penalties under general liability insurance is designed to protect companies from the financial burdens that may arise from contractually agreed penalty payments. Such contractual penalties are often stipulated to ensure compliance with specific performance commitments, such as performance guarantees for AI systems. In the event of non-performance, agreed penalty payments may become due in full or in part. As these purely contractual liability and warranty provisions go beyond statutory liability, they are not usually covered by standard liability insurance policies.

For example, an insurance policy may stipulate that cover is provided if the error rate of an AI system exceeds a defined threshold (e.g. 1 per cent). In this context, the insurer insures the start-up or scale-up, which in turn makes corresponding performance commitments in its licence agreements. Country-specific legal frameworks must be taken into account here; in particular, it must be ensured that the company actually provides guarantees and does not act as an 'insurer' itself.

Such solutions are already being offered on the market by a number of specialist providers, particularly for start-ups and scale-ups, which benefit from the combination of contractual liability and insurance cover. In addition to financial protection, this also fosters trust, as the risk assessment carried out by an insurer serves as a mark of quality. At the same time, potential losses can be compensated for in full or in part, resulting in a competitive advantage over providers without such cover.

Regulatory and Liability Law Framework

The terms and conditions of third-party liability cover are generally based on statutory liability provisions under private law.[18] Sanctions under public law are generally not covered, unless the insurance contract provides for explicit extensions. This also applies in principle to E&O and cyber insurance.[19]

When assessing the liability implications of the use of AI agents and automated processes, a distinction must first be made between regulatory and civil law requirements. Regulation (EU) 2024/1689 (EU AI Act) is structured as public-law regulatory and product safety legislation. It does not give rise to independent civil law claims for damages by aggrieved third parties against providers or operators of AI systems. Breaches are primarily sanctioned through regulatory measures and fines. Whether individual operator or provider obligations under the EU AI Act can, furthermore, be classified as protective laws within the meaning of Section 823(2) of

[18] See Veith, Gräfe and Gebert, **The Insurance Litigation Process**, Section D, § 19, para. 32.
[19] See Höra, Part C, § 12, paras. 39, 40.

the German Civil Code (BGB) and thus indirectly give rise to civil law claims for compensation has not yet been conclusively clarified in the legal literature.[20]

Following the withdrawal of the originally planned AI Liability Directive in 2025, the liability of companies using AI towards third parties who have suffered damage continues to be governed by the general civil law liability regimes under national law. Of particular relevance are tortious liability, liability for risk and product liability, as well as contractual liability. These provisions also apply to AI-related damages.

Liability of the AI Manufacturer

The use of AI technology presents manufacturers with new challenges. In addition to technological progress, changes in the regulatory landscape are also leading to an increased liability risk. Courts in Germany and the US have already confirmed that companies are liable for the AI output they generate. The EU product liability reform explicitly includes software and AI models within strict liability for the first time, leading to a drastic change in the existing obligation to pay compensation for attributable damage.

The current risks faced by AI manufacturers are, in principle, covered by existing insurance products, in particular by specialised financial loss liability products, known as 'Tech E&O' or 'Tech PI' products; the shift in product liability towards strict liability should not have a negative impact on insurance cover. In the context of claims settlement, however, the changed legal situation could initially lead to an increased burden of providing information. Manufacturers should also assess the need for separate product liability insurance.

In relation to manufacturers' risks, it may also prove difficult in future to determine who is responsible for the output generated by AI. This is due to the fact that there are numerous sources of error in the field of AI. If the output is faulty, this may be attributed either to the underlying algorithm and therefore to the software company. However, the faulty output may also be attributable to the company using the AI technology having provided faulty or incomplete training data, or the underlying prompt not having been precise enough to guarantee error-free results. Furthermore, the further development of AI technologies by companies that do not strictly fall within the category of software companies also gives rise to difficult questions of attribution.

In addition to the usual liability risks, software companies are also exposed to the risk of copyright infringements. This applies to both the training and the application of the technology.

In conclusion, it must be noted that manufacturers of AI systems face far-reaching liability issues and that existing insurance solutions are not yet sufficiently tailored to this changed situation. The following areas for action should be taken into account.

[20] 'The Directive on AI Liability – Less is More!' by Prof. Dr Gerhard Wagner | 4 September 2024.

Liability of the AI User

Operator and Quasi-Manufacturer Roles

The starting point for assessing liability is the company's role within the AI value chain. The EU AI Act distinguishes between the provider of an AI system and its operator. The user risk considered here fundamentally concerns the role of the operator, i.e. the professional or operational use of an AI system under one's own responsibility.

However, the potential shift in role to that of a manufacturer or quasi-manufacturer is of particular practical significance. This may be the case, in particular, where a company substantially modifies purchased models, retrains them using its own data, combines several systems into an autonomous AI agent, or markets the solution under its own brand. This may also occur as part of the company's own product or service. Within the EU AI Act, this shift in role from operator to provider (Article 25 of the EU AI Act) is formally restricted to high-risk AI systems; by contrast, the general 'quasi-manufacturer' concept under product liability law – such as marketing under one's own brand or making substantial modifications to a product – has a longer history and applies regardless of any risk classification under the EU AI Act. In such cases, additional manufacturer obligations and product liability risks may arise alongside the operator's obligations.

This distinction is central to the classification under insurance law, as it may give rise to different liability regimes and allocations of cover within business, product, E&O or cyber insurance.

Tortious Liability

In relation to persons with whom no contractual relationship exists, tortious liability is the primary liability regime. The relevant legal basis for claims is Section 823 of the German Civil Code (BGB).

Liability requires that the use of an AI system results in harm to a protected legal interest, such as life, health or property, and that the user can be accused of culpable conduct. Examples include personal injury or property damage caused by faulty AI-controlled production, logistics or control systems.

Fault is generally linked to a breach of duties to ensure public safety. These include, in particular, the careful selection and configuration of the system, the limitation of its scope of action, quality assurance of input data, appropriate monitoring and the possibility of human intervention. The operators' obligations under the EU AI Act are likely to be increasingly relied upon in future to clarify the standard of care under civil law.

Financial losses are generally not covered by the scope of protection under Section 823(1) of the German Civil Code (BGB). This is a fundamental principle of tort law that applies regardless of the technology used. In the context of AI, such losses typically arise within contractual relationships and are covered there under Sections 280 et seq. of the BGB.

Where personal data is processed, a claim for damages may also arise under Article 82 of the GDPR. This also covers non-pecuniary damage and provides for a distribution of the burden of proof that is more favourable to the injured party.

Strict Liability and Product Liability

There is currently no general strict liability regime for the operation of autonomous AI systems. Consequently, the user is not, in principle, liable merely for using AI, but is only liable on a strict liability basis if the operation of the system is already covered by an existing strict liability regime, such as in the field of road traffic or the operation of certain installations.

However, product liability is of particular significance. Under the German Product Liability Act (ProdHaftG) and the underlying European Product Liability Directive, strict liability generally applies to the manufacturer of a defective product. Unlike in tort law under Section 823 of the German Civil Code (BGB), liability here does not depend on fault, but solely on the fact that a defective product was placed on the market and that this resulted in damage eligible for compensation. In particular, life, physical integrity, health and property are protected. Pure financial losses, by contrast, do not primarily fall within the scope of product liability.

Product liability is becoming increasingly relevant for users and operators of AI systems, as the status of 'manufacturer' is not limited to the original developer. Under current law, a company may already be liable as a quasi-manufacturer if it markets a product under its own name or brand. In the context of AI, situations may also arise in which users effectively assume the role of a manufacturer through significant adaptations to a base model, their own training measures, fine-tuning processes, the integration of additional data sources or the combination of several systems.

This development is further reinforced by the new Product Liability Directive (EU) 2024/2853. The Directive explicitly clarifies that software, AI systems and digital services may constitute products within the meaning of product liability law. Furthermore, it extends liability to cases in which a product is substantially altered after it has been placed on the market. Anyone who does not merely use an AI system, but significantly influences its functioning, further develops it or makes it available under their own responsibility, may therefore be exposed to product liability risks in future.

This creates a new liability risk, particularly for companies that integrate generative AI solutions into their own products, services or decision-making processes. In many cases, the line between mere use and product liability is not clear-cut. This applies in particular when AI models are customised, trained using the company's own data or offered to customers as a standalone solution. Companies should therefore bear in mind any resulting exposure when modifying and structuring their use of AI. From a risk management perspective, it should therefore be assessed at an early stage whether the specific structure of the AI's use could establish the company as a manufacturer or quasi-manufacturer.

Under product liability law, too, purely financial losses generally fall outside the scope of protection. The focus is on personal injury, property damage and, under the new Directive, certain data-related damages. The practical relevance of product liability

Hamm Higher Regional Court, 12 May 2026 (Case No. 4 UKl 3/25)

The chatbot of a cosmetic surgery clinic had attributed specialist medical articles – some of which did not exist – to the two practising doctors. The Court of Appeal did not accept the argument that the chatbot operated on a purely statistical basis and that the statements could not be attributed to the two doctors due to a lack of control. As the operator defines the scope of the chatbot's activities, it has sufficient influence over the system.

from a company's perspective thus complements other forms of liability arising from a contractual relationship or tort. Nevertheless, it can take on considerable significance in the case of AI-related personal injury or property damage, as, unlike in other cases, it is not necessary to prove fault on the part of the manufacturer.

Member States must transpose the Product Liability Directive (EU) 2024/2853 into national law by 9 December 2026. Companies should monitor the associated changes at an early stage and, in particular, critically examine their role in the AI value chain.

Contractual Liability

When it comes to the practical use of AI agents in customer processes, contractual liability – or liability arising from a contractual relationship – is generally of the greatest significance. Unlike tort law and product liability law, it also covers purely financial losses, which arise, amongst other things, in automated business processes.

The statutory claims for damages arising from breaches of contractual obligations under Sections 280 et seq. of the German Civil Code (BGB) are decisive.[21] In the context of the use of AI agents, breaches of obligation may arise, in particular, from erroneous automated decisions, incorrectly executed business transactions or delays in the performance of the service owed.

Although the AI agent itself is not a legal entity and is therefore not a vicarious agent within the meaning of Section 278 of the German Civil Code (BGB), the user cannot invoke the system's autonomy. The results of automated processes are, in principle, attributed to the user. Liability is linked, in particular, to the selection, configuration, monitoring and organisational integration of the system.

Of particular significance is the rule on the burden of proof set out in Section 280(1), second sentence, of the German Civil Code (BGB): liability is presumed. The user must therefore be able to prove that they have taken appropriate organisational and technical measures. Documented AI governance with clear responsibilities, control mechanisms and escalation processes thus takes on considerable significance in terms of liability law.

Conclusion: Risk of Third-Party Damage Under the AI

The EU AI Act is designed as a regulatory and product safety law and deliberately does not establish a separate civil liability regime. Nor is such a regime necessary. Liability towards third parties who have suffered damage continues to be governed by the general rules of tort, product liability and contract law.

For operators of AI agents, the primary focus is on duties relating to ensuring safety, monitoring and organisation. As automation increases, so do the requirements for governance, monitoring and human oversight. At the same time, the modification or

[21] See Säcker et al., **Munich Commentary on the Civil Code**, Vol. 2, Part 2, § 280, para. 5.

integration of third-party AI systems can lead to a shift in role from operator to manufacturer or quasi-manufacturer, thereby triggering additional product liability risks.

From an insurance law perspective, it is therefore always necessary to examine the role in which the company is acting and which liability regime applies. The key risk currently lies not so much in novel, AI-specific grounds for liability as in the application of existing liability principles to automated and increasingly autonomous business processes. For AI manufacturers who sell their products to other companies, regardless of statutory liability, purely contractually agreed performance and service commitments or guarantees contain lump-sum claims for damages that go beyond statutory liability. Conventional insurance policies fail to cover these voluntary extensions of liability towards third parties; consequently, specialised and application-specific solutions should be taken into account when assessing a potential risk transfer.

7 Governance and Reputational Risks

The use of AI solutions in companies poses a high risk to management and supervisory bodies. The rapidly advancing use of AI technologies is forcing companies to adopt these technologies so as not to fall behind. At the same time, companies are obliged to use AI technologies in compliance with the law. Corporate bodies are currently operating within this area of tension. Failure to address the applicable legal provisions and any resulting damage may lead to personal liability on the part of the management bodies. The argument that they lack specific knowledge in the field of AI technologies does not exempt the bodies from liability, particularly where essential supervisory duties have been breached. The importance of supervisory duties has been reaffirmed by a ruling of the Federal Court of Justice. In its ruling of 14 October 2025 (Case No. II ZR 78/24), the Federal Court of Justice held that, in fulfilling their supervisory duties, members of the supervisory board are obliged to actively seek out information and must not accept a lack of reporting by the executive board. This does not affect the Executive Board's obligation to provide information in a proper and complete manner; the governing bodies are jointly and severally liable to the company or to any entitled third party. The monitoring and proper documentation of the use of AI technologies are therefore absolutely essential.

Failure to implement the compliance requirements of the EU AI Act may result in the imposition of substantial fines. The EU AI Act provides for a graduated scale of fines, similar to the GDPR:

Offence	Fine
Particularly serious infringements, such as those relating to prohibited AI practices (Article 5 of the EU AI Act) or key obligations in relation to high-risk or GPAI systems	up to EUR 35 million or 7 per cent of global annual turnover
Other breaches of the obligations under the EU AI Act (e.g. shortcomings in risk management, data governance, documentation, human oversight)	up to EUR 15 million or 3 per cent of global annual turnover

For businesses, this means:

- Financial losses due to fines, regulatory measures and restrictions on the use of certain models or applications,
- high adaptation and retrofitting costs if non-compliance is only detected at a late stage and rectified retrospectively,
- direct interference with the business model through restrictions or bans on specific AI applications.

In the German insurance market, administrative fines are generally not insurable, or only to a very limited extent. If a company is fined due to the inadequate implementation of legal provisions by its supervisory and executive bodies, the company may seek recourse against those bodies. However, the insurability of such claims under directors' and officers' (D&O) insurance remains a matter of dispute in Germany.

In addition to the actual fines, companies face further financial burdens:

- **compliance and adjustment costs** for retrospective documentation, model validation, data governance, governance boards and AI literacy programmes,
- **reputational damage** resulting from publicised breaches of AI regulations (including fines that attract media attention),
- **ESG and capital market risks** arising from inadequate or 'whitewashed' AI governance ('AI washing').

This makes compliance with the EU AI Act an integral part of the cross-cutting 'AI' risk. Whether AI-related undesirable developments are identified at an early stage, managed, or at least dealt with in a way that minimises liability and reputational damage depends largely on the quality of governance. If the required quality is not met, this leads to a personal liability risk for the management and supervisory bodies, which can, to a certain extent, be transferred to D&O insurance.

In addition to regulatory sanctions, contractual penalties are also gaining in importance:

- performance and quality guarantees for AI systems vis-à-vis business partners (e.g. guaranteed error rates, availability, non-discrimination),
- penalties for breaches of contractually agreed AI governance standards (audits, documentation and oversight arrangements),
- Contractual penalties for breaches of AI-related transparency or labelling obligations towards end customers**.**

These payments are based on private law, but in practice are often linked to obligations under the EU AI Act or accompanying regulations. They may take the form of pure financial losses and thus fall within the scope of financial loss liability cover (e.g. E&O, PI). A clear contractual distinction (statutory liability vs. contractual penalties) is crucial here. It should be noted that the insurability of 'genuine' contractual penalties remains very limited and sometimes entails a high premium burden.

The risk of reputational damage is currently a major concern for companies. This is confirmed by the latest Allianz Risk Barometer. The companies surveyed recognise the rapid spread of AI technologies as a significant threat to their reputation and day-to-day operations. The insurance industry is responding to this changing risk landscape with specialised cover solutions. In addition to existing cover for reputational costs, some providers offer insurance cover for loss of profits resulting from reputational damage, for example as part of cyber insurance policies. Despite these innovative insurance concepts, it must be noted that not all governance and reputational risks are insurable. Furthermore, even in the case of insured losses, the claims settlement process may not proceed satisfactorily in practice. Particular attention should be drawn here to the difficulties involved in providing evidence of the loss.

The fragmented insurability of AI risks and the resulting need to purchase several insurance products that are coordinated with one another also increase the liability exposure for board members. It is therefore essential to establish proper risk management, AI governance and a comprehensive insurance strategy. Governance and reputational risks are thus not peripheral issues, but a distinct driver of losses: they interlink regulatory obligations, personal directors' and officers' liability, and market confidence. A robust governance and documentation framework determines whether AI-related adverse developments are identified and mitigated at an early stage – or whether they result in fines, directors' and officers' liability claims and lasting reputational damage.

8 Many Parties – One AI model: Joint Insurance Cover for a Shared AI Project Risk?

One of the key challenges in insuring AI systems is that their development, deployment and use regularly involve a large number of different parties. Consequently, losses often cannot be clearly attributed to a single liable party. For example, errors or failures may be attributable to a provider's base model, the training data used, the customisation, integration and operation carried out by software providers, system integrators or external IT service providers, the operational configuration, or the specific use by the end-user. Whilst existing liability and insurance regimes already provide a legal framework for AI-related damage, the large number of parties involved often makes it time-consuming to determine the causes of damage and assign responsibility; as a result, claims can frequently only be finally settled after lengthy coordination and clarification processes.

A possible future solution could therefore be a cross-party AI insurance scheme modelled on established EAR/CAR (Erection All Risks / Construction All Risks) cover. The basic idea is to cover the various parties involved in the development, implementation and operation of an AI system not in isolation, but collectively within a single insurance model. Developers of base models, data suppliers, integrators, operators and end users would be brought together under a single policy. Such an approach could simplify the often complex allocation of responsibilities seen today, speed up claims settlement and spread premiums and risks more widely. At the same time, it would at least partially resolve the frequently difficult distinction between product defects, implementation errors, data issues, misconfiguration or misuse of the system.

In practice, such a multi-party model would initially be limited to a single, clearly defined AI system or a specific AI application. This would make risk assessment, technical evaluation and underwriting considerably simpler. Insurers could carry out a targeted analysis of the system's functionality, training and governance processes, as well as its interdependencies, and use this as a basis for a robust risk assessment. The analysis of the causes of claims would also be considerably more transparent in a clearly defined use case than in complex AI landscapes with a multitude of interacting models.

However, the scalability of this approach is limited. It is suitable only for scenarios in which a single dominant or business-critical AI model controls key business processes and thus forms a clearly identifiable core risk.

9 Outlook

In the coming years, artificial intelligence will evolve from an innovative stand-alone technology into an integral part of operational value creation. As generative and agent-based AI systems become increasingly widespread, their integration into operational processes, decision-making workflows and products will continue to grow. This not only increases the potential for efficiency gains and new business models, but also the importance of a systematic approach to managing the resulting risks.

The present analysis suggests that the real challenge in future will lie less in the technology itself than in its governance. AI rarely creates entirely new types of loss, but it does alter existing risks through its speed, scalability and scope. Errors, poor decisions, liability claims or business interruptions can occur more rapidly as a result of AI, have a greater impact and be more difficult to trace within complex process chains. Risk assessment will therefore increasingly shift from individual technologies towards issues of governance, control and the allocation of responsibility.

This trend is likely to intensify further with the continued proliferation of autonomous and agent-based systems and applications with greater scope for action. Whilst the focus today is often on the quality of individual AI models, issues relating to human oversight, options for intervention, documentation, control mechanisms and organisational accountability will become increasingly important in the future. For businesses, the ability to manage AI effectively will thus become a factor in competitiveness and resilience.

The EU AI Act provides, for the first time, a binding regulatory framework specifically tailored to artificial intelligence. In the short term, it will largely be perceived as a compliance requirement. In the medium term, however, it is likely to play a much greater role as a tool for standardising and professionalising AI governance. Similar to information security or data protection standards, AI Act-compliant processes, documentation and control mechanisms could develop into generally accepted quality benchmarks.

For the insurance industry, too, a further development of existing market concepts is emerging. Even today, numerous AI-related claims can be covered under established property, cyber or liability policies. At the same time, insurers will increasingly need to review their policy terms and conditions to determine how AI-related risks should be systematically classified. The debate surrounding ‘Silent AI’ is likely to follow a similar trajectory to the earlier debate on ‘Silent Cyber’.

There is still a need for further development, particularly with regard to first-party losses resulting from incorrect decisions or performance shortfalls in AI systems. Initial guarantee and performance concepts indicate that the market is already addressing the question of how model- and performance-related risks can be made insurable in future. However, the fundamental business risk remains the limiting factor in such a development: economically detrimental business decisions or mismanagement cannot be readily insured, even if they were prepared or influenced by AI systems. It is to be expected that new coverage approaches, parametric models and underwriting models will emerge in this area.

As AI becomes more widespread, the debate surrounding liability, the allocation of responsibility and the insurability of third-party losses will intensify further. In particular, where AI prepares or autonomously executes decisions, existing liability regimes and insurance concepts will need to be adapted to new technical and organisational realities. Given the complex interrelationship between the original developer of the AI, the AI user who adapts and modifies it for their own purposes, and the user's customers, the question of assigning responsibility is not easily answered. In cases of doubt, this will remain a case-by-case issue that must be examined within individual risk management and assessed in the context of existing insurance concepts. As long as humans remain the final authority (human-in-the-loop), such activities should continue to pose no problems; consequently, the focus of further discussion will be on the application of AI at this threshold and beyond. At the same time, it cannot be ruled out that insurers will propose an AI exclusion clause.

In addition to regulatory and underwriting issues, the responsible use of AI is increasingly becoming a matter of trust. Customers, employees and business partners expect AI systems to be used in a transparent, fair manner and under appropriate human supervision. Companies that embed transparency, governance and accountability at an early stage can thereby not only meet regulatory requirements and reduce risks, but also build trust, foster acceptance and achieve long-term competitive advantages.

Against this backdrop, there is a strong case for viewing AI in future not primarily as a technology, but as a manageable business risk. The key question will therefore be less whether companies use AI, and more how they organise, control and take responsibility for its use. As AI becomes more widespread, governance, control and compliance will become crucial success factors for resilience, insurability and sustainable economic success.

Glossary

Agentic AI (AI Agents)
Agentic AI refers to AI systems that are capable of planning and executing tasks with a comparatively high degree of autonomy. They often combine an AI or general-purpose AI model with additional components ('scaffolding') such as access to tools, memory, and planning and orchestration logic. This enables them to handle multi-step tasks, control external systems or trigger actions independently. From a regulatory perspective, 'agentic AI' is not a separate risk category under the EU AI Act. Rather, the decisive factors are the purpose, specific application and role of the respective entity. If an agentic AI system is used for a use case listed in Annex III and meets the requirements set out therein, it may be classified as a high-risk AI system. Any obligations relating to the underlying GPAI models and to the AI system built upon them must be distinguished from one another on the basis of their respective roles.

Agent Sandboxing
In the context of AI agents, sandboxing refers to the execution of the agent within an isolated, controlled environment. This environment acts as a vault that restricts the agent's access to the rest of the system, the databases or the network. It therefore protects not the contents from the outside world, but the outside world from the contents. The aim is to protect the core infrastructure from unpredictable or malicious behaviour on the part of the agent.

This security measure brings with it both significant advantages and operational challenges. Strong isolation effectively protects the entire system from malfunctions or manipulation should the agent be compromised, for example, by a prompt injection. It is a robust defence. Nevertheless, one must bear in mind the downside: strict separation can noticeably impair performance and complicate the necessary exchange of data with other systems.

Autonomy & Unpredictability
Agents can make decisions that lie outside the intended logic ('deviation'). As they plan independently, their behaviour is often more difficult to predict than the purely text-based output of a standard LLM.

Bias
Bias refers to systematic distortions in data, models or decision-making processes that can lead an AI system to repeatedly disadvantage or incorrectly assess certain individuals, groups or circumstances. Causes may include, amongst other things, unbalanced or non-representative training data, flawed model assumptions, unsuitable objective measures or the nature of the application. In the context of AI, bias is particularly relevant when automated assessments or decisions have an impact on access, pricing, selection processes or services.

LLMs / Generative Models
These systems are primarily optimised to recognise patterns in data and generate statistically probable content (text, images, code). They are passive and 'stateless': they respond to an input prompt and cease operation once the content has been generated.

AI agents
They use an LLM as the 'brain' for planning and understanding, but supplement this with an action loop (agentic workflow). An agent can plan iteratively, adjust its approach based on results, and interact directly with software, databases or the internet via APIs to achieve an overarching goal.

The main risk lies in the combination of autonomy and access to production systems. The main threats are:

Privileges & security risks
As agents have access to interfaces (APIs) and databases, they can cause far-reaching damage in the event of malpractice or manipulation (e.g. through 'prompt injection'), such as through unauthorised data access or the automated transmission of sensitive information.

Dynamic runtime risk
Unlike traditional security vulnerabilities (CVEs), which are static elements within the code and can be patched, risks associated with AI agents often arise dynamically during execution ('in-flight'). A reactive patching strategy is often insufficient in this context.

Lack of control loops
Without security mechanisms such as 'human-in-the-loop' (human approval processes) in place, errors in logic or hallucinations can enter real business processes unfiltered, which can lead to financial losses or compliance breaches.

Background on Claude Mythos
Claude Mythos is an AI model developed by the US company Anthropic, which is classified as a high-level 'dual-use technology'. It has been specifically trained to detect hidden security vulnerabilities in complex programmes. Because Mythos plans multi-stage attacks completely autonomously, Anthropic considers the model too dangerous for the general public. As part of 'Project Glasswing', it is therefore being made available exclusively to selected IT security firms and authorities such as the BSI. The system operates in an extremely agent-like manner. It has already identified thousands of dangerous vulnerabilities in web browsers. In the wrong hands, however, this technology could be used as a devastating digital weapon against power grids or banks.

Economic background
Experts suspect that, in addition to purely security-related reasons, infrastructure and cost considerations also lie behind the secrecy: the computing power required to run the system is enormous, and analysing a single codebase costs up to hundreds of thousands of euros.

Prompt Injection
A prompt injection attack is a method whereby an attacker deliberately sends manipulated inputs (prompts) to an AI model in order to override its original programming or security instructions.

Essentially, this attack involves overwriting or simply ignoring the AI's original system instructions through skilful phrasing. The AI then accepts the attacker's instructions as its new reality. This is a serious problem. Whilst this often leads only to bizarre responses in the case of a simple chatbot, the situation is critical with AI agents. As

they have access to external tools, attackers could manipulate them to extract sensitive data or send emails on behalf of others.

You can think of it as a form of digital identity spoofing: the attacker disguises their voice so that the agent's 'brain' believes the command is coming directly from the system administrator, even though it originates from the user.

Security experts primarily distinguish between two main forms of manipulation. In direct prompt injection, also known as jailbreaking, the user enters the malicious instruction directly into the chat. The indirect variant, on the other hand, is considerably more insidious. Here, an attacker places the manipulative text on a web page that the agent reads whilst conducting research. As soon as the AI processes this text, it executes the command hidden within it without anyone noticing.

As this is one of the greatest vulnerabilities for autonomous agents, defending against prompt injection is one of the key tasks of input validation and guardrailing, which we discussed earlier.

Guardrail
In the context of AI systems and agents, a 'guardrail' refers to a predefined safety or control mechanism that ensures the AI model operates within specified limits.

Think of it like a real crash barrier on a motorway: it prevents the vehicle (the AI) from veering off the road, even if the driver (the user or a malicious command) tries to steer the vehicle into the crash barrier or the ditch.

What exactly do guardrails do? Guardrails act as a filter or verification layer situated between the user and the AI model (or between the agent and their tool/system). They monitor communication in real time and intervene if certain rules are breached.

In practice, guardrails fulfil various critical functions to protect the infrastructure. Input guardrails serve as an entry check and verify whether the user's request is safe, for example by filtering out attempts at prompt injection. On the other hand, there are output guardrails. These monitor the AI's responses for hallucinations or the disclosure of sensitive company data before they reach the user. Finally, governance guardrails ensure that regulatory requirements such as the GDPR or the EU AI Act are strictly adhered to, for instance by blocking unauthorised data flows to external servers.

AI firewall
An AI firewall is a security mechanism that monitors, filters and, where necessary, blocks the inputs, outputs and system interactions of an AI system. It can help to detect and limit prompt injection attacks, unauthorised data leaks, policy breaches or risky tool calls at an early stage. It is particularly relevant for AI agents with access to APIs, databases, email, CRM or workflow systems. An AI firewall does not replace governance or approval processes, but can act as a technical layer of protection to help limit misconduct and manipulation during execution.

Sandbox
A sandbox is a technically isolated test or execution environment in which AI systems, models or agents can be tested in a controlled manner without directly affecting production systems, sensitive data or business-critical processes. It serves to assess behaviour, the consequences of errors, security risks and interactions prior to production deployment. In the context of AI, a sandbox is particularly relevant for systems with tool access, agent functions or process automation, as it limits the scope for action and can reveal malfunctions before they are transferred to real-world business processes.

AI model
An AI model is the technical core of an AI system, which maps patterns, correlations or probabilities on the basis of training data and uses these to generate results such as forecasts, classifications, recommendations or content. The model can be compromised by faulty data, manipulation, inappropriate adjustments, loss of model parameters or technical faults. Damage to the AI model can result in unreliable results or necessitate the model being retrained, restored or replaced, either in whole or in part.

AI application
An AI application is the specific integration of an AI model into a business or technical context, such as in customer service, underwriting, claims processing, fraud detection or document processing. In addition to the model, it also comprises interfaces, data sources, user interfaces, authorisations, workflows and control mechanisms. If an AI application is damaged, altered, deleted or rendered unusable, this may affect not only the model itself, but also its use within the business process, the availability of the application or the quality of the results produced.

AI asset
An AI asset is a digital asset of economic value that is required for the development, operation or use of an AI system. This may include AI models, training, test and validation data, embeddings, prompt libraries, vector databases, configurations, fine-tunings, documentation or monitoring information. An AI asset is considered compromised if it is damaged, altered, deleted, manipulated or rendered unusable, and this may result in recovery costs, operational disruptions or other financial losses.

Network security breach
A network security breach refers to an incident in which the security, confidentiality, integrity or availability of an IT system, network or the data processed therein is compromised. Typical examples include unauthorised access, malware, data manipulation, data leakage, denial-of-service attacks or the compromise of access credentials. In the context of insurance, the term is particularly relevant because cyber insurance policies are often linked to a defined cyber incident or a network security breach as the triggering insured event.

Appendix

A.1: Key obligations for AI users under the AI Act, depending on risk classification:

	Key obligations / requirements	Legal basis
General prohibitions and cross-cutting obligations	Prohibition of certain AI practices, e.g. certain manipulative applications or social scoring	Art. 5
	Requirements regarding AI literacy for users and decision-makers	Art. 4
	Transparency obligations, e.g. towards data subjects in relation to certain AI applications	Art. 50
General-Purpose AI (GPAI)	Technical documentation, summary of training content and information on copyright requirements	Art. 51–56
	Additional requirements for GPAI models posing systemic risk	Art. 51–56
High-risk AI systems, in particular as set out in Annex III	Risk management	Art. 9
	Data governance	Art. 10
	Technical documentation	Art. 11, Annex IV
	Logging and traceability	Art. 12
	Transparency and information for operators ('Instructions for Use')	Art. 13
	Human Oversight	Art. 14

	Key obligations / requirements	Legal basis
	Accuracy, robustness and cybersecurity	Art. 15
	Quality Management System	Art. 17
	Fundamental Rights Impact Assessment	Art. 27
	Conformity assessment	Art. 43
	Post-market monitoring and reporting of serious incidents	Articles 72 and 73

A.2: AI Risk Matrix



AI risk	Media liability	Professional liability	Product liability	Employer's liability	Intellectual property	Cyber	Reputational damage	R&D
Liability for defective products and services	Excluded	Generally available	Limited	Limited	Excluded	Excluded	Excluded	Excluded
Copyright, trade mark or service mark infringement	Generally available	Limited	Limited	Limited	Generally available	Limited	Excluded	Excluded
Patent infringement	Limited	Excluded	Excluded	Excluded	Generally available	Excluded	Excluded	Excluded
Discrimination / bias	Limited	Limited	Excluded	Excluded	Excluded	Limited	Excluded	Excluded
Defamation, slander, insult	Generally available	Excluded	Excluded	Excluded	Excluded	Excluded	Excluded	Excluded
Personal injury	Excluded	Excluded	Generally available	Generally available	Excluded	Limited	Excluded	Excluded
Property damage	Excluded	Excluded	Generally available	Generally available	Excluded	Limited	Excluded	Excluded
Data protection and information security breaches	Excluded	Limited	Excluded	Excluded	Excluded	Generally available	Limited	Excluded
Loss of financial assets	Excluded	Excluded	Excluded	Excluded	Excluded	Limited	Generally available	Excluded
Market manipulation	Excluded	Excluded	Excluded	Excluded	Excluded	Excluded	Excluded	Limited
Fraud	Excluded	Limited	Excluded	Excluded	Excluded	Limited	Generally available	Excluded
Robotics	Excluded	Generally available	Generally available	Limited	Excluded	Excluded	Excluded	Excluded
Product recall	Excluded	Excluded	Generally available	Limited	Excluded	Excluded	Excluded	Excluded
Business interruption	Excluded	Excluded	Excluded	Excluded	Excluded	Generally available	Excluded	Excluded
Breach of duties by management personnel	Excluded	Excluded	Excluded	Excluded	Excluded	Excluded	Excluded	Generally available

Generally available Limited Excluded

Bibliography

Allianz Commercial (2026): Allianz Risk Barometer 2026.

https://commercial.allianz.com/content/dam/onemarketing/commercial/commercial/reports/allianz-risk-barometer-2026.pdf

beck-aktuell (2026): 'Breathless Through the Night' in an AI model may infringe copyright, 31 July 2026.

Bild-Kunst (2026): Hamburg Higher Regional Court ruling in the LAION case, 2026.

Bild-Kunst (2026): Text and data mining – ruling by the Hamburg Regional Court, January 2026.

Bird & Bird (2025): Hamburg Higher Regional Court Confirms AI Training was Permitted, 17 December 2025.

Bitsea (2025): Anyone who still distributes AI-generated code today has conditional intent to infringe rights, 27 October 2025.

Federal Court of Justice (BGH) (2025): Judgment of 14 October 2025, Case No. II ZR 78/24, available via openJur.

https://openjur.de/u/2538374.html

Federal Republic of Germany (2026): Act implementing Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending other legal acts (Act implementing the Regulation on artificial intelligence).

https://www.recht.bund.de/bgbl/1/2026/223/VO.html

CMS (2025): AI-generated software code in due diligence, 30 October 2025.

Computerwoche (2026): Allianz Risk Barometer 2026: AI jumps to second place, 14 January 2026.

https://www.computerwoche.de/article/4116684/allianz-risk-barometer-2026-ki-springt-auf-platz-2.html

DATEV magazine (2026): Companies are liable for incorrect statements made by their AI chatbots, 13 May 2026.

https://www.datev-magazin.de/nachrichten-steuern-recht/recht/unternehmen-haften-fuer-falsche-aussagen-ihres-ki-chatbots-146688

European Union (2022): Regulation (EU) 2022/2554 of the European Parliament and of the Council of 14 December 2022 on digital operational resilience in the financial sector and amending Regulations (EC) No 1060/2009, (EU) No 648/2012, (EU) No 600/2014, (EU) No 909/2014 and (EU) 2016/1011 (Digital Operational Resilience Act – DORA).

https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32022R2554

European Union (2024): Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=OJ:L_202401689

European Union (2026): Regulation (EU) 2026/1744 of the European Parliament and of the Council of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI).

https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=OJ:L_202601744

GASTRO.at (2026): AI leads to cold pizza, 26 May 2026.

https://gastroportal.at/pizza-hut-wegen-ki-auf-100-millionen-verklagt/96940/

German Insurance Association (GDV) (2020): General Conditions for Electronics Insurance (ABE 2020).

German Insurance Association (GDV) (2020): General Conditions for Machinery Insurance of Stationary Machinery (AMB 2020).

German Insurance Association (GDV) (2024): General Insurance Conditions for Cyber Risk Insurance (AVB Cyber), as at February 2024.

GEMA (2025): First landmark AI ruling in Europe: GEMA prevails against OpenAI, 11 November 2025.

Höra, Knut / Schubach, Arno (eds.) (2022): Munich Lawyers' Handbook on Insurance Law, 5th revised and expanded edition, Munich: C.H. Beck, ISBN 978-3-406-76903-0. Sack Fachmedien

IT BOLTWISE (2026): Munich Regional Court halts Suno's unauthorised AI music training – judgement not yet final*, 3 August 2026.

https://www.it-boltwise.de/landgericht-muenchen-stoppt-unbefugtes-ki-musiktraining-von-suno-urteil-nicht-rechtskraeftig.html

it-daily.net (2026): Pizza Hut: Franchisee claims $100 million for AI-related damages, 19 May 2026.

https://www.it-daily.net/shortnews/pizza-hut-franchisenehmer-klage

Bavarian Ministry of Justice (2025): Press release from the Munich I Regional Court dated 11 November 2025.

Bavarian Ministry of Justice (2026): Press release from the Munich I Regional Court dated 31 July 2026.

Fürstenow Law Firm (2026): Liability risks for executives in 2026, 29 April 2026.

https://www.kanzlei-fuerstenow.de/haftungsrisiken-fuer-fuehrungskraefte-in-2026/

Kramarz Law Firm (2025): Copyright for AI-generated code: A question of the burden of proof, 21 June 2025.

KPMG Law (2026): The Business Judgement Rule and AI: How Governing Bodies Are Liable for Decisions*, 19 March 2026.

https://kpmg-law.de/business-judgement-rule-und-ki-so-haften-organe-fuer-entscheidungen/

Munich Regional Court I (2025): Judgment of 11 November 2025, Case No. 42 O 14139/24, available via openJur.

Legal Tribune Online (LTO) (2024): AI and copyright: Landgericht Hamburg's landmark judgement*, 2024.

Lime Connect / Lime Technologies (2026): Steil, Tamina: Chatbot liability following the Higher Regional Court of Hamm ruling: What companies need to know now, 15 June 2026.

https://connect.lime-technologies.com/de/blog/chatbot-haftung-olg-hamm/

Manufaktur für Finanzplanung (2026): AI Liability Guidelines 2026: The Ultimate Guide for GmbH Managing Directors on Protecting Themselves Against the New Liability Vacuum, 22 February 2026.

https://blog.manufaktur-fuer-finanzplanung.de/2026/02/22/ki-haftungsrichtlinien-2026-der-ultimative-guide-fuer-gmbh-geschaeftsfuehrer-zur-absicherung-gegen-das-neue-haftungs-vakuum/

Hamburg Higher Regional Court (2025): Judgment of 10 December 2025, Case No. 5 U 104/24, available via openJur.

Rehkatsch Solicitors (2026): Kneschke v LAION: AI Training & Section 44b of the German Copyright Act (UrhG), 26 June 2026.

Säcker, Franz Jürgen / Rixecker, Roland / Oetker, Hartmut / Limperg, Bettina (eds.) (2021–2025): Munich Commentary on the German Civil Code, 9th edition, Munich: C.H. Beck, ISBN 978-3-406-76670-1. The 9th edition was published in instalments from 2021 to 2025. SLUB Dresden

t3n (2026): Brien, Jörn: AI leads to cold pizza: Franchisee sues Pizza Hut for 100 million dollars in damages, 25 May 2026.

https://t3n.de/news/ki-pizza-hut-klage-100-millionen-1743300/

Tabassi, Elham (2023): Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, National Institute of Standards and Technology, Gaithersburg, MD.

https://doi.org/10.6028/NIST.AI.100-1

Tageskarte (2026): Pizza Hut franchisee sues chain for $100 million over AI system, 18 May 2026.

https://www.tageskarte.io/gastronomie/detail/pizza-hut-franchisenehmer-verklagt-kette-wegen-ki-system-auf-100-millionen-dollar.html

Veith, Jürgen / Gräfe, Jürgen / Gebert, Yvonne (eds.) (2020): Insurance Litigation. Claims and Proceedings, 4th edition, Baden-Baden: Nomos, 1,488 pages, ISBN 978-3-8487-5521-9.

Wagner, Gerhard (2024): «The Directive on AI Liability – Less is More!", in: RDi – Recht Digital, editorial of 4 September 2024.

Competition Centre (2026): Hamm Higher Regional Court holds company liable for statements made by its chatbot, full text available, 3 June 2026.

https://www.wettbewerbszentrale.de/olg-hamm-laesst-unternehmen-fuer-aussagen-seines-chatbots-haften-volltext-verfuegbar/

WirtschaftsWoche (2026): Tödtmann, Claudia: AI in customer service: Companies are liable for their chatbots – including for false promises, 21 June 2026.

https://www.wiwo.de/erfolg/management/ki-im-kundenservice-unternehmen-haften-fuer-falsche-auskuenfte/100231910.html

Bitkom represents more than 2,200 member companies from the digital economy. In Germany, they generate a turnover of just over 200 billion euros through digital technologies and solutions and employ more than 2 million people. Its members include more than 1,000 small and medium-sized enterprises, over 700 start-ups and almost all global players. They offer software, IT services, telecommunications or internet services; manufacture devices and components; operate in the digital media sector; create content; provide platforms; or are otherwise part of the digital economy. 82 per cent of the companies involved in Bitkom have their headquarters in Germany, a further 8 per cent are based in the rest of Europe and 7 per cent in the USA. 3 per cent are based in other regions of the world. Bitkom promotes and drives the digital transformation of the German economy and advocates for broad societal participation in digital developments. The aim is to make Germany a high-performing and independent digital hub.

Publisher

Bitkom e.V.
Albrechtstr. 10 | 10117 Berlin

Contact

Lukas Spohr | Bitkom e.V.
T | +49 175 5848827 | l.spohr@bitkom.org

Responsible Bitkom committee

Digital Insurance & InsurTech Working Group

Authors

Bastian Bahnemann, FSI Compliance & Business Development Lead for Germany, Microsoft | **Dr Peter Bärnreuther**, Senior Underwriter for Artificial Intelligence Risks, Munich Re | **Christian Freudner**, VP Management Liability, Finlex | **Martin Hake,** Chief Underwriter for Third-Party Liability, R+V Versicherung | **Kerstin Keltner,** Managing Director, Head of Speciality, AON Austria | **Sima Krause**, AI Governance Manager, ERGO Group AG | **Klemens Lemke,** Underwriting Manager for Cyber, Hiscox Germany | **Sabine Pawig-Sander**, Non-Executive Director, Finlex | **Hanno Pingsmann,** Founder & Managing Director, CyberDirekt | **Axel Saß,** Chief Architect, Field CTO Organisation, Red Hat | **Stefan Schmutterer**, Underwriting and Claims Consultant, CyberRisk, R+V Versicherung | **Matthias Schneider**, CRO, Markel SE | **Bogdan Werth**, Insurance Lead Germany, Microsoft

Copyright

Bitkom 2026

This publication constitutes general, non-binding information. The contents reflect Bitkom's views at the time of publication. Although the information has been compiled with the utmost care, no warranty is given as to its factual accuracy, completeness and/or timeliness; in particular, this publication cannot take into account the specific circumstances of individual cases. Use of this publication is therefore at the reader's own risk. All liability is excluded. All rights, including the right to reproduce extracts, are reserved by Bitkom or the respective rights holders.