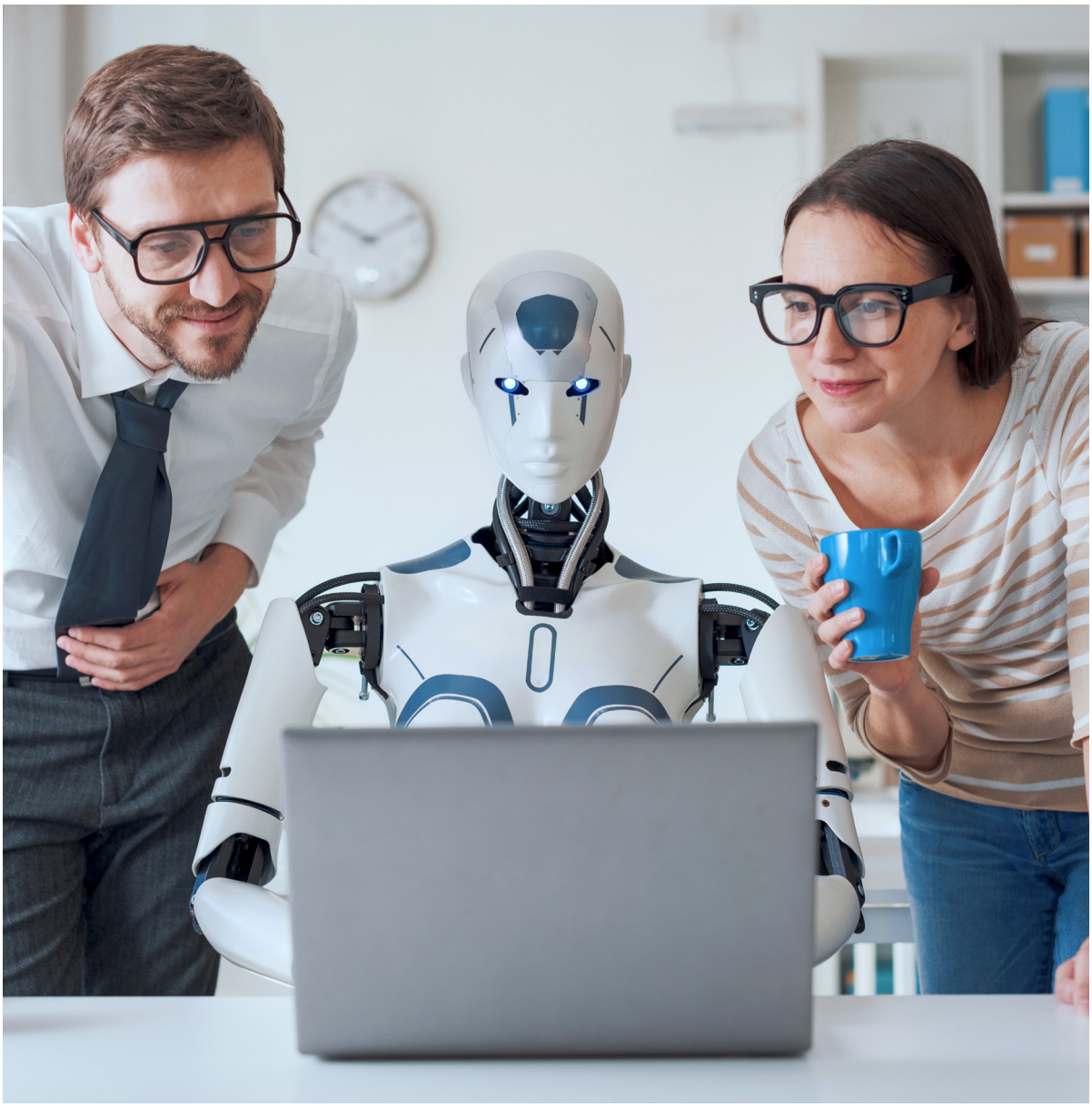
FAST LTA

Factcheck

The Risks of GenAI for Organizations



Since: December 2024

From zero to one million users in just five days.[1] Since ChatGPT was introduced to the public in November 2022, large language models and generative AI have attracted enormous attention.

In 2019, Microsoft secured preferred-partner status through a substantial investment of one billion dollars. Copilot is a direct result of this partnership. A further 10 billion dollars followed in 2023, most likely also to help market leader ChatGPT compete with growing rivals. These include Alphabet with Gemini, Meta with Llama and xAI with Grok.

Generative AI is not only relevant to private users. In total, 47% of respondents say that the technology is already highly relevant to their company. By 2027, 60% expect it to be highly relevant.[2]

Companies hope that generative AI will increase efficiency, reduce costs, accelerate innovation cycles and offer personalised solutions that improve products and services.

However, using ChatGPT and similar tools within organisations also involves risks.

Risk 1: Hallucinating AI

> *„GPT-4 has the tendency to "hallucinate," i.e. "produce content that is nonsensical or untruthful in relation to certain sources." This tendency can be particularly harmful as models become increasingly convincing and believable, leading to overreliance on them by users."*[3]

Generative AI systems have a tendency to hallucinate. This can have several causes:

- Insufficient or biased training data
- Errors in the underlying algorithms
- Excessive generalisation when processing unfamiliar input

If employees blindly trust hallucinated information, this can have serious consequences for a company.

It can lead to incorrect business decisions and the associated financial losses.[4] False information communicated publicly can also cause lasting damage to a company's reputation.

If false information generated by AI is passed on to customers, it may even have legal consequences, especially in sensitive fields such as medicine or finance.

---

[1] https://www.statista.com/chart/29174/time-to-one-million-users

[2] https://www.luenendonk.de/produkte/studien-publikationen/luenendonk-studie-2024-der-markt-fuer-it-dienstleistungen-in-deutschland

[3] https://cdn.openai.com/papers/gpt-4-system-card.pdf

[4] https://www.semanticscholar.org/reader/413dfda358df6e360805189f93b95ec71dc6caea

It is therefore no surprise that 57% of CIOs and IT managers see the insufficient quality of AI-generated results as a factor slowing the adoption of AI systems within companies.[5]

Risk 2: Indirect Prompt Injection

Indirect prompt injections are a new type of security threat that specifically affects large language models, or LLMs.

An attacker places hidden instructions in content that will later be processed by the AI. This content may consist of documents, websites or other data sources used by the LLM system to respond to queries.

When an unsuspecting user submits a request, the LLM processes both the user's request and the hidden malicious instructions.

In addition to spreading incorrect information, this method can be used to manipulate users. The LLM could, for example, ask the user to disclose sensitive data or download malware.

Without effective safeguards, this method has a success rate of more than 50%.[6]

Risk 3: Proxy-pages

Netskope Threat Labs has identified more than 1,000 malicious URLs and domains seeking to exploit the hype surrounding generative AI. These include proxy sites that look remarkably similar to official websites.[7]

Sensitive information entered carelessly, such as source code, customer data, passwords or trade secrets, may fall directly into the hands of criminals.

Even non-sensitive data entered on proxy sites can be used to create sophisticated fraud scenarios.

Risk 4: Data Leakage

Company documents, customer data, financial information and intellectual property. Employees may enter sensitive or confidential information into generative AI tools without being aware of the consequences.

Many generative AI services store submitted data to improve their models.[8] This means that confidential business information may be stored on external servers and used to train the AI.

[5] https://www.luenendonk.de/produkte/studien-publikationen/luenendonk-studie-2024-generative-ai-von-der-innovation-bis-zur-marktreife

[6] https://arxiv.org/abs/2302.12173

[7] https://www.netskope.com/de/netskope-threat-labs/cloud-threat-report/ai-apps-in-the-enterprise

[8] https://www.proofpoint.com/de/blog/security-awareness-training/generative-ai-risks-to-know

In the worst-case scenario, sensitive information could become accessible to anyone, anywhere, simply by entering the right prompt.

Companies now use an average of 9.6 different generative AI applications.

Of all sensitive information shared with AI tools, 46% relates to source code. More than one-third, 35%, concerns data that companies are legally required to protect. Intellectual property, excluding source code, accounts for 15%, while passwords and security keys account for 4%.[9]

Source code is shared with generative AI particularly often. One reason is that generative AI can reliably detect errors and security vulnerabilities in code. IT professionals also tend to adopt new technologies at an early stage.

On average, 158 incidents occur every month per 10,000 business users.

Regulated data, such as financial information, health information and other personal data, is shared with generative AI an average of 18 times per month per 10,000 business users. Intellectual property other than source code is shared four times per month.[10]

Risk 5: Shadow-AI

Despite all the risks, 33% of companies give their employees unrestricted access to generative AI tools.

This contrasts with 13% of companies where the use of ChatGPT and similar tools is completely prohibited.[11]

At least one generative AI application is blocked by 77% of companies. The presentation application Beautiful.ai tops the list of blocked applications at 28%.[12]

A complete ban creates the risk that employees will secretly switch to private devices, further increasing the risk of sensitive data being leaked.[13]

The term shadow AI describes the unofficial and often uncontrolled use of AI tools by employees outside the company's official IT infrastructure and policies.

Both unrestricted access and a complete ban encourage shadow AI. Given the associated risks, neither is a suitable approach.

What should companies do to ensure the safe use of AI within their organisation?

[9] https://www.netskope.com/netskope-threat-labs/cloud-threat-report/july-2024-ai-apps-in-the-enterprise

[10] https://www.netskope.com/de/netskope-threat-labs/cloud-threat-report/ai-apps-in-the-enterprise

[11] https://www.luenendonk.de/produkte/studien-publikationen/luenendonk-studie-2024-generative-ai-von-der-innovation-bis-zur-marktreife

[12] https://www.netskope.com/netskope-threat-labs/cloud-threat-report/july-2024-ai-apps-in-the-enterprise

[13] https://www.pwc.de/de/cloud-digital/digital/generative-ki-unterstuetzung-von-der-strategie-bis-zur-umsetzung/the-genai-building-blocks/genai-ist-gekommen-um-zu-bleiben-was-das-fur-die-cybersicherheit-bedeutet.html

Step 1

Train employees and implement guidelines

Information about the risks of generative AI

Employees must be fully informed about the specific risks and threats associated with generative AI. This includes topics such as data privacy, potential bias in AI-generated results and the limitations of AI technology.

Training in best practices

Regular training should teach employees how to develop and use generative AI tools safely. This includes secure data entry, interpreting AI-generated results and ethical considerations.

Develop clear guidelines and standards

Companies must establish detailed guidelines that specify exactly how generative AI tools may be used within the organisation. These guidelines should define which types of data may be entered into AI systems, how generated results should be handled and which security measures must be followed.[14]

Clarify responsibilities and liability

The guidelines must clearly define who is responsible for the use of generative AI tools and what actions must be taken in the event of errors or data breaches.
This creates clarity and legal certainty for everyone involved.

Step 2

Technological Solutions

Technological solutions must also be implemented to minimise the risks associated with using generative AI.

Some of the available solutions are discussed below.

[14] https://trendreport.de/generative-ai-sicher-aktivieren

Solution 1: Use generative AI that takes security seriously

Free and publicly accessible AI models such as ChatGPT and Gemini are particularly exposed to the risks described above. However, alternatives are available that have been developed specifically for business use.

These systems are designed from the ground up with a focus on data protection and information security.

One example is Claude, an AI assistant developed by Anthropic. Claude stands out because of its advanced security approach and reliability when processing information.

The responsible handling of sensitive business information is central to Claude's security architecture. Conversations and outputs are automatically deleted from the systems after three months.

It is especially important that user data is not used to train the model. Companies should therefore be able to integrate Claude into their existing processes with fewer concerns.[15]

The quality of AI-generated content is supported by Claude's innovative Constitutional AI approach.

This underlying architecture was developed specifically to minimise false statements and hallucinations.

Transparency is a key characteristic of the system. If Claude is uncertain about an answer or does not have enough information, it communicates this clearly and unambiguously.

This honesty builds trust and enables companies to make informed decisions based on AI-generated output.[16]

Disadvantage

Three months may sound like a short period, but during that time there is still a risk of data theft by cybercriminals.

The US CLOUD Act may also allow US authorities to access the data.

The demand for a European LLM that provides greater sovereignty over organisations' own data is no coincidence.

The German company Aleph Alpha was long regarded as a promising player in this field, but unexpectedly discontinued the development of its language model in September 2024.[17]

Solution 2: Web Browser Filters

Data Loss Prevention, or DLP, solutions for web browsers provide another way to minimise risks.

Modern DLP tools integrate seamlessly with web browsers and monitor interactions between users and AI platforms in real time.

[15] https://clickup.com/de/blog/209690/claude-ai-review

[16] https://mindsquare.de/karriere-news/ki-chatbot-claude-besser-als-chatgpt

[17] https://t3n.de/news/aleph-alpha-gegen-openai-deutsche-ki-startups-1645839

These solutions allow companies to control the exchange of data with AI systems and protect sensitive information.

A solution such as LayerX automatically detects and filters sensitive information, ensuring that only harmless data is shared with AI systems.

If an employee accidentally attempts to paste confidential source code into ChatGPT, the browser extension detects it immediately and prevents the transfer.

The user also receives a clear explanation of why the action was blocked. These educational elements make an important contribution to improving employees' security awareness.

Disadvantage

As a browser extension, LayerX only provides protection when generative AI is accessed through a web browser. Prohibiting the entry of sensitive data may also limit the benefits that generative AI can offer.

Solution 3: Avoid External Cloud Services Entirely

An effective strategy for minimising risks is to avoid external cloud services completely and use a local AI application instead.

An on-premises solution ensures that sensitive business information never leaves the organisation's own infrastructure. It also provides full control over data access and processing.

Silent AI was developed specifically to provide maximum protection for business information while making full use of artificial intelligence.

At the heart of the system is a local large language model, supplemented by Retrieval-Augmented Generation, or RAG.

RAG enables targeted searches for information within documents and extends the capabilities of the LLM. This combination allows the system to provide highly accurate answers through direct access to internal company documents.

Unlike conventional AI systems, which transfer data to external cloud services, Silent AI keeps all information within the company's own infrastructure.

The data is stored in specialised vector databases optimised for AI applications. These databases enable highly efficient storage and extremely fast searches across documents.

Silent AI can operate entirely without an internet connection, virtually eliminating the risk of data leakage.

The system automatically respects the company's existing user permissions. Employees can only access information for which they are authorised.

Integration with the existing IT environment has also been carefully designed.

Silent AI can connect directly to commonly used business systems such as Microsoft 365, SharePoint, Confluence and Jira.

Documents do not have to be uploaded separately. This simplifies handling and prevents additional security risks.

Silent AI avoids the disadvantages of the other solutions presented above.

Because no information leaves the organisation, there are no restrictions on the data that can be entered, while maximum protection is maintained.

Conclusion

The main risks associated with using generative AI within companies are unreliable AI-generated output, data leakage and cybercrime.

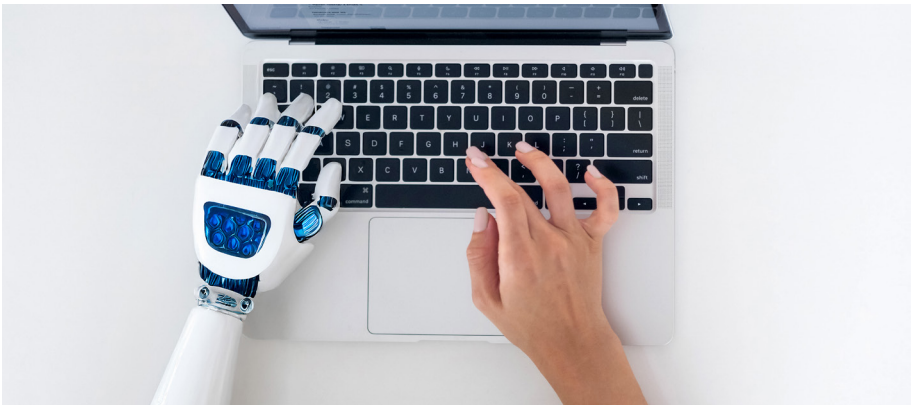
These risks cannot be addressed through simple bans or unrestricted access, as both approaches encourage uncontrolled shadow AI.

A secure implementation of generative AI systems requires a comprehensive approach that intelligently combines technical and organisational measures. This allows companies to use the potential of AI safely.

Proper employee training, combined with clearly defined company guidelines, forms the foundation. These organisational measures create the necessary awareness and confidence among everyone involved.

Several technological solutions are available. The best solution must always be selected based on the company's specific requirements. A combination of different systems may also be useful.

When maximum data protection is required, Silent AI is a suitable option because information is processed locally without restricting the prompts or data that users can enter.



PROUDLY MADE IN EUROPE

FAST LTA

FAST LTA
Rüdesheimer Str. 11
80686 München
info@fast-lta.de
www.fast-lta.de

Design,
Entwicklung
und Support
in **Deutschland**