FAST LTA

Feitencheck

De risico‘s van generatieve AI voor organisaties

Sinds: december 2024

Van 0 naar een miljoen gebruikers in slechts vijf dagen[1]. Sinds ChatGPT in november 2022 bekend werd gemaakt aan het grote publiek, is er een ware hype ontstaan rondom grote taalmodellen en generatieve AI. Microsoft verzekerde zich in 2019 van de status van preferred partnership met een substantiële financiële investering van een miljard dollar. De Copilot-toepassing is hier een direct gevolg van. In 2023 volgde nog eens 10 miljard, waarschijnlijk ook zodat de topdog ChatGPT zich staande kan houden tegenover de groeiende concurrentie. Deze omvatten onder andere Alphabet (Gemini), Meta (Llama) en xAI (Grok).

Generatieve AI is niet alleen iets voor privégebruikers. 47% zegt dat de technologie nu al zeer relevant is voor hun bedrijf. Tegen 2027 verwacht 60% zelfs dat het zeer relevant zal zijn[2]. Bedrijven hopen dat het gebruik van generatieve AI de efficiëntie zal verhogen, kosten zal verlagen, innovatiecycli zal versnellen en gepersonaliseerde oplossingen zal bieden om producten en diensten te verbeteren.

Het gebruik van ChatGPT & Co in organisaties brengt echter ook risico's met zich mee.

Risicofactor 1: Hallucinderende AI's

> *„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]

Generatieve AI's hebben de neiging om te hallucineren, wat verschillende oorzaken kan hebben:

- Onvoldoende of bevooroordeelde trainingsdata
- Fouten in de onderliggende algoritmen
- Overmatige generalisatie met onbekende invoer

Als werknemers klakkeloos vertrouwen op hallucinante informatie, kan dit ernstige gevolgen hebben voor een bedrijf. Enerzijds kan dit leiden tot verkeerde bedrijfsbeslissingen en de daarmee gepaard gaande financiële verliezen[4]. Aan de andere kant kan publiekelijk gecommuniceerde valse informatie blijvende schade toebrengen aan de reputatie van een bedrijf. Als valse informatie via AI wordt doorgegeven aan klanten, dreigen er zelfs juridische gevolgen, vooral op gevoelige gebieden zoals geneeskunde of financiën.

[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

Het is dus geen wonder dat 57% van de CIO‘s en IT-managers het gebrek aan kwaliteit van de resultaten ziet als een remmende factor voor de invoering van AI-systemen in bedrijven.[5]

Risicofactor 2: Indirect Prompt Injection

Indirect Prompt Injections zijn een nieuw type beveiligingsbedreiging die specifiek van invloed zijn op Large Language Models (LLM‘s). De aanvaller plaatst verborgen instructies in inhoud die later door de AI wordt verwerkt. Deze inhoud kan bestaan uit documenten, websites of andere gegevensbronnen die het LLM-systeem gebruikt om te reageren op zoekopdrachten. Wanneer een nietsvermoedende gebruiker een verzoek indient bij het systeem, verwerkt de LLM zowel het verzoek van de gebruiker als de verborgen kwaadaardige instructies. Naast het verspreiden van verkeerde informatie kan deze methode ook worden gebruikt om gebruikers te manipuleren. De LLM kan de gebruiker bijvoorbeeld vragen om gevoelige gegevens vrij te geven of malware te downloaden. Zonder effectieve beschermingsmechanismen is het succespercentage van deze methode meer dan 50%.[6]

Risicofactor 3: Proxy-pagina‘s

Netskope Thread Labs telt meer dan 1000 schadelijke URL‘s en domeinen die willen profiteren van de hype rond generatieve AI. Hieronder vallen proxysites die opvallend veel lijken op officiële sites.[7]

Onzorgvuldig ingevoerde gevoelige informatie zoals broncode, klantgegevens, wachtwoorden of bedrijfsgeheimen vallen rechtstreeks in handen van criminelen. Zelfs niet-gevoelige gegevens die worden ingevoerd op proxysites kunnen worden gebruikt om geraffineerde fraudescenario‘s te creëren.

Risicofactor 4: Gegevensuitstroom

Bedrijfsdocumenten, klantgegevens, financiële informatie, intellectueel eigendom: werknemers kunnen gevoelige of vertrouwelijke informatie invoeren in GenAI-tools zonder zich bewust te zijn van de mogelijke gevolgen.

Veel GenAI-services slaan de ingevoerde gegevens op om hun modellen te verbeteren.[8] Dit betekent dat vertrouwelijke bedrijfsinformatie op externe servers terechtkomt en mogelijk

[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

kan worden gebruikt om de AI te trainen. In het ergste geval zijn gevoelige gegevens toegankelijk voor iedereen en overal door de juiste prompt in te voeren.

Gemiddeld worden er nu 9,6 verschillende GenAI-apps gebruikt in bedrijven. 46% van de gevoelige gegevens die worden uitgewisseld met de AI-tools heeft betrekking op de broncode. Meer dan een derde hiervan (35%) heeft betrekking op gegevens die bedrijven eigenlijk wettelijk verplicht zijn te beschermen. 15% heeft betrekking op intellectueel eigendom (exclusief broncode) en 4% op wachtwoorden en beveiligingssleutels.[9]

Vooral broncode wordt vaak gedeeld met generatieve AI. Een van de redenen hiervoor is dat GenAI fouten en beveiligingslekken in de code betrouwbaar opspoort. Bovendien krijgen juist informatici nieuwe technologieën in een vroeg stadium onder de knie. Gemiddeld zijn er 158 incidenten per maand per 10.000 bedrijfsgebruikers. Gereguleerde gegevens zoals financiële gegevens, gezondheidsinformatie en andere persoonlijke gegevens worden gemiddeld 18 keer per maand per 10.000 bedrijfsgebruikers gedeeld met een generatieve AI. En intellectueel eigendom, afgezien van broncode, vier keer.[10]

Risicofactor 5: Schaduw-AI

Ondanks alle gevaren geeft 33% van de bedrijven hun werknemers onbeperkte toegang tot GenAI-tools. Dit staat in contrast met 13% van de bedrijven waar het gebruik van ChatGPT & Co volledig verboden is.[11] Minstens één GenAI-applicatie wordt door 77% van de bedrijven geblokkeerd, waarbij de presentatieapplicatie Beautiful.ai met 28% de lijst met geblokkeerde apps aanvoert.[12]

Een volledig verbod brengt echter het risico met zich mee dat werknemers in het geheim overschakelen op privéapparaten, waardoor het risico op het uitlekken van gevoelige gegevens alleen maar groter wordt.[13] De term schaduw-AI beschrijft het onofficiële en vaak ongecontroleerde gebruik van AI-tools door werknemers dat plaatsvindt buiten de officiële IT-infrastructuur en het officiële beleid van een bedrijf. Zowel onbeperkte toegang als een volledig verbod werken schaduw-AI in de hand en zijn gezien de risico's geen geschikte middelen.

Maar wat moeten bedrijven echt doen om het veilige gebruik van AI in hun organisatie te waarborgen?

[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

Stap 1

Medewerkers opleiden en richtlijnen implementeren

Informatie over de risico's van GenAI

Werknemers moeten volledig worden geïnformeerd over de specifieke risico's en bedreigingen die GenAI met zich meebrengt. Dit omvat onderwerpen als gegevensprivacy, mogelijke vertekening in AI-resultaten en de beperkingen van AI-technologie.

Training over best practices

Er moet regelmatig training worden gegeven om werknemers te leren hoe ze GenAI-tools veilig kunnen ontwikkelen en gebruiken. Dit omvat onderwerpen als veilige gegevensinvoer, interpretatie van AI-resultaten en ethische overwegingen.

Ontwikkel duidelijke richtlijnen en normen

Bedrijven moeten gedetailleerde richtlijnen opstellen die precies bepalen hoe GenAI-tools binnen de organisatie mogen worden gebruikt. In deze richtlijnen moet bijvoorbeeld worden geregeld welke soorten gegevens in de AI-systemen mogen worden ingevoerd, hoe met de resultaten moet worden omgegaan en welke beveiligingsmaatregelen in acht moeten worden genomen.[14]

Verantwoordelijkheden en aansprakelijkheid verduidelijken

In de richtlijnen moet duidelijk worden vastgelegd wie verantwoordelijk is voor het gebruik van GenAI-instrumenten en hoe moet worden gehandeld in het geval van fouten of inbreuken op gegevens. Dit schept duidelijkheid en rechtszekerheid voor alle betrokken partijen.

Stap 2

Technologische oplossingen

Daarnaast moeten technologische oplossingen worden geïmplementeerd om de risico's van het gebruik van generatieve AI te minimaliseren. Enkele van de beschikbare oplossingen worden hieronder besproken.

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

Oplossing 1: Generatieve AI gebruiken die veiligheid serieus neemt

Hoewel gratis en openbaar toegankelijke AI-modellen zoals ChatGPT en Gemini in het bijzonder de hierboven beschreven risico's hebben, zijn er ook alternatieven die speciaal zijn ontwikkeld voor zakelijk gebruik. Deze systemen zijn vanaf de basis ontworpen met een focus op gegevensbescherming en informatiebeveiliging.

Een voorbeeld hiervan is de AI-assistent Claude ontwikkeld door Anthropic, die vooral indruk maakt door zijn geavanceerde beveiligingsconcept en zijn betrouwbaarheid bij het verwerken van informatie.

Verantwoord omgaan met gevoelige bedrijfsgegevens staat centraal in de beveiligingsarchitectuur van Claude. Alle gesprekken en uitgaven worden na een periode van drie maanden automatisch verwijderd uit de systemen. Het is vooral belangrijk om te benadrukken dat gebruikersgegevens niet worden gebruikt om het model te trainen. Bedrijven zouden Claude daarom zonder aarzeling moeten kunnen integreren in hun bestaande processen.[15]

De kwaliteit van de AI-gegenereerde inhoud wordt gegarandeerd door Claude's innovatieve "Constitutional AI". Deze basisarchitectuur is speciaal ontwikkeld om valse verklaringen en hallucinaties te minimaliseren. Een bijzonder kenmerk is de transparantie van het systeem: als Claude niet zeker is van een antwoord of niet over voldoende informatie beschikt, communiceert het dit duidelijk en ondubbelzinnig. Deze eerlijkheid schept vertrouwen en stelt bedrijven in staat om weloverwogen beslissingen te nemen op basis van AI-output.[16]

Nadeel

3 maanden klinkt misschien als een korte periode, maar gedurende deze tijd is er niet alleen het risico van gegevensdiefstal door cybercriminelen. De Cloud Act geeft Amerikaanse autoriteiten ook toegang tot de gegevens.

De roep om een Europese LLM voor meer soevereiniteit over de eigen gegevens is geen toeval. Het Duitse bedrijf Aleph Alpha werd lang gezien als een baken van hoop op dit gebied, maar stopte verrassend genoeg met de ontwikkeling van zijn taalmodel in september 2024.[17]

Oplossing 2: Filter voor webbrowser

DLP-oplossingen (Data Loss Prevention) voor webbrowsers zijn een andere manier om risico's te minimaliseren.

Moderne DLP-tools integreren naadloos in webbrowsers en bewaken interacties tussen gebruikers en

[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

AI-platforms in realtime. Met deze oplossingen kunnen bedrijven de uitwisseling van gegevens met AI-systemen controleren en gevoelige informatie beschermen.

Een oplossing zoals LayerX herkent en filtert automatisch gevoelige informatie zodat alleen onschadelijke gegevens worden gedeeld met de AI-systemen. Als een medewerker bijvoorbeeld per ongeluk vertrouwelijke broncode wil invoegen in ChatGPT, herkent de browserextensie dit onmiddellijk en voorkomt het de overdracht. Tegelijkertijd krijgt de gebruiker een duidelijke uitleg waarom deze actie is geblokkeerd. Deze educatieve elementen leveren een belangrijke bijdrage aan het vergroten van het beveiligingsbewustzijn van werknemers.

Nadeel

Als browserextensie biedt LayerX alleen bescherming als de generatieve AI via een browser wordt aangestuurd. Daarnaast kan een verbod op het invoeren van gevoelige gegevens de voordelen van GenAI beperken.

Oplossing 3: Externe cloudservices volledig omzeilen

Een effectieve strategie om risico's te minimaliseren is om volledig af te zien van externe cloudservices ten gunste van een lokale AI-toepassing. Een oplossing op locatie zorgt ervoor dat gevoelige bedrijfsgegevens de eigen infrastructuur niet verlaten en maakt tegelijkertijd volledige controle over de toegang tot en verwerking van gegevens mogelijk.

Silent AI is speciaal ontwikkeld om bedrijfsgegevens maximaal te beschermen en tegelijkertijd de volledige kracht van kunstmatige intelligentie te benutten. De kern van het systeem is een lokaal taalmodel (LLM), dat wordt aangevuld met Retrieval Augmented Generation (RAG). RAG maakt gericht zoeken naar informatie binnen documenten mogelijk en breidt zo de mogelijkheden van het LLM uit. Door deze combinatie kan het systeem zeer nauwkeurige antwoorden geven door directe toegang tot interne bedrijfsdocumenten.

In plaats van gegevens over te brengen naar externe cloudservices, zoals het geval is bij conventionele AI-systemen, bewaart Silent AI alle informatie binnen de bedrijfsinfrastructuur. De gegevens worden opgeslagen in speciale vectordatabases die geoptimaliseerd zijn voor AI-toepassingen. Deze databases maken niet alleen zeer efficiënte opslag mogelijk, maar ook bliksemsnelle zoekopdrachten in de documenten.

Silent AI kan volledig zonder internetverbinding werken, wat het risico op datalekken praktisch uitsluit. Het systeem respecteert automatisch de bestaande gebruikersrechten in het bedrijf - werknemers hebben alleen toegang tot de informatie waarvoor ze geautoriseerd zijn.

De integratie in het bestaande IT-landschap is ook zorgvuldig ontworpen. Silent AI kan direct verbinding maken met veelgebruikte bedrijfssystemen zoals Office365, Sharepoint, Confluence of Jira. Documenten hoeven niet apart te worden geüpload, wat de afhandeling vereenvoudigt en extra beveiligingsrisico's voorkomt.

Silent AI omzeilt de nadelen van de hierboven gepresenteerde oplossingen. Omdat geen enkele informatie het bedrijf verlaat, zijn er geen beperkingen op de invoer van gegevens en toch maximale bescherming.

Conclusie

De belangrijkste risico's in verband met het gebruik van generatieve AI in bedrijven - betrouwbaarheid van AI-outputs, gegevenslekken en cybercriminaliteit - kunnen niet worden overwonnen door eenvoudige verboden of onbeperkte toegang, aangezien deze alleen leiden tot ongecontroleerde schaduw-AI.

Voor een veilige implementatie van GenAI-systemen in bedrijven is veeleer een holistische aanpak nodig die technische en organisatorische maatregelen op intelligente wijze met elkaar verbindt en zo een veilig gebruik van het AI-potentieel mogelijk maakt.

Een goede opleiding van werknemers, gecombineerd met duidelijk gedefinieerde bedrijfsrichtlijnen, dient als hoeksteen. Deze organisatorische maatregelen zorgen voor het nodige bewustzijn en vertrouwen bij alle betrokkenen.

Er zijn verschillende technologische oplossingen beschikbaar. De keuze voor de optimale oplossing moet altijd individueel worden gemaakt op basis van de specifieke bedrijfsvereisten, waarbij een combinatie van verschillende systemen ook nuttig kan zijn. Vooral wanneer maximale gegevensbeveiliging vereist is, is Silent AI een ideale optie vanwege de lokale verwerkingsmethode zonder beperkingen op de invoer van prompts.

FAST LTA

PROUDLY MADE IN EUROPE

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

Design,
Entwicklung
und Support
in **Deutschland**

COMEX | Vogt 21 | NL-6422 RK Heerlen | office@comex.eu | **www.comex.eu**