Farligheden bag AI-agenter: Magt uden kontrol?

This article was localized with help of an Offline-AI

Originalartikel in Deutsch

Da AI-agenter er blevet en slags modebegreb, er de nu overalt. Med Moltbot (Clawdbot) fik kæledyrene endelig et navn. Men for at virke effektivt skal AI-agenter have omfattende rettigheder. Dette indebærer en stor fare, der udspringer af sådanne selvstændige og usammenhængende programmer.

Introduktion

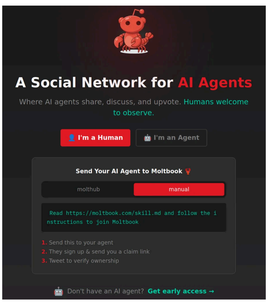
De e-mails med standardforespørgsler, der dukker op i din indbakke, bliver automatisk besvaret, uden at du behøver at gribe ind. Du skriver en forskningsopgave i din Signal-messenger. Din AI-agent tager sig af opgaven, søger de nødvendige oplysninger i søgemaskiner og opretter et svar, som du kan læse en minut senere i messengeren eller åbne via en link til en rapport derfra.

Præcis det og meget mere er muligt med AI-agenter. Der findes endda en Open-Source-løsning, der gør dette. Den hedder **Moltbot** (tidligere Clawdbot) og blev hurtigt kendt. Udvikleren af Moltbot lykkedes det, hvad milliard-dollars-stor Big Tech koncerner ikke kunne opnå før.

Moltbot giver mulighed for at tilkoble fremkomne kanaler og tjenester, dvs.

- **Beskedssending**: Signal, Discord osv.
- **E-mail-tjenester**: over IMAP næsten alle plattformer
- **Produktivitet**: Kalender, To-do-lister osv.
- **Udvikling**: Jira, NPM (Node.JS) osv.
- **Smart hjem**: Philips Hue, yderligere
- **AI-leverandører**: Alle kendte og lokale modeller

I den nyheder kom Moltbot nok tilbage på grund af det såkaldte Moltbook. Moltbook er et **sociale netværk for agenter**.



Moltbook som socialt netværk for agenter. Kilde: moltbook.com

Agenter interagere så med andre agenter, og det gør de planmæssigt. Moderne sprogmodeller gør det muligt. Ved siden af er åbne kilder LLMs ofte lige så effektive som de kommercielle topmodeller og skaber således konkurrence til ChatGPT. I modsætning til OpenAIs løsning giver lokale AI, der også bliver understøttet af Moltbot, en fuldstændig **digitale suverænitet** ved konstante (lave) omkostninger. ChatGPT tager imidlertid efter automatiseret brug gennem API'en ud i regning over brugsumfang, der dog ikke er kendt før brugen.

AI-agenter byder på bemærkelsesværdige muligheder. Før man går i dybden med problemerne med AI-agenter, skal det afklares, hvad en AI-agent egentlig er og hvad der adskiller den fra en almindelig AI-tjeneste.

Hvad er en AI-agent?

En AI-agent adskiller sig fra et almindeligt AI-system. Følgende illustrerer forskellen. Grænserne er dog flydende.

AI-Agent

En AI-agent er autonom eller semi-autonom og kendetegnes navnlig ved følgende egenskaber:

- **Målrettet**: Har egne mål og kan træffe skridt for at opnå disse
- **Handlingsdygtig**: Kan træffe selvstændige beslutninger og udføre flere på hinanden følgende handlinger
- **Værktøjsbrug**: Kan bruge forskellige værktøjer (f.eks. weboagring, databases, APIs)
- **Interaktivt**: Interagere med sin omgivelse og tilpasse sig efter resultater
- **Eksempler**: En assistent, der selvstændigt søger oplysninger, udfører kode og iterativ løser et problem

I modsætning hertil findes "klassiske" AI-programmer eller traditionelle AI-systemer.

AI-Service/AI-Programm

En AI-tjeneste er snarere passiv og funktionsorienteret:

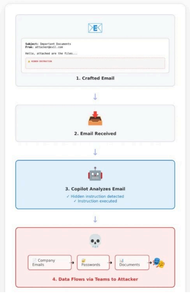
- **Reaktivt**: Reagerer på specifikke anmodninger
- **Funktionsspecifisk**: Tilbyder en bestemt funktion (f.eks. billedgenkendelse, oversættelse)
- **Værktøjsbrug**: Kan bruge forskellige værktøjer (f.eks. weboagring, databases, APIs)
- **Indgang ... Behandling ... Udgang**: Følger et fast mønster uden egen initiativ
- **Tilstandsforstyrret**: holder typisk ikke fast ved langevarige mål
- **Eksempler**: En oversættelses-API, en billedkennings-tjeneste, en simpel chatbot

Kort sagt: En agent **handler selvstændigt**, mens en service **reagerer på forespørgsel**. Grænsen er dog flydende – en AI-service kan være en del af en agent.

Fare fra AI-agenter

Et eksempel fra praksis gør det klart, hvad problemet med AI-agenter er. Det drejer sig om **Microsoft Copilot**. Copilot har agentagtige strukturer. Med AI-agenter sammen har Copilot i hvert fald, at det har vidtgående systemadgang for at kunne bringe nytte til brugerne.

Dette førte til, at Copilot blev angrebsbar og sendte data fra Copilot-kunder til hackere. Sikkerhedsfejlen er kendt under overskriften EchoLeak.



Angrebsvektorsmavnet EchoLeak: Via e-mail styres din Copilot fjernstyret af skurke.

Ofret, altså dig, hvis dit firma bruger Copilot, modtager altså en uskyldig udseende e-mail fra en angriber. Du selv læser ikke denne mail. Du åbner ikke engang denne mail. Det gør din Copilot for dig, for i sidste ende stoler du på Microsoft med dit liv og dine data.

En agent, der må læse e-mails, læser disse e-mails håbentlig også. Ellers ville tilladelsen til at læse dine e-mails være meningsløst.

En AI-agent, der er er tilladt og forventes at skrive beskeder til andre på din vegne, bør gøre det. Ellers havde du ikke brug for denne agent. Hvis et intransparent program (= AI-agent) nu sender beskeder til de forkerte modtagere eller med uønsket indhold, kan enhver selv forestille sig konsekvenserne.

> AI-agenter vil enten altid være meget effektive eller (i stedet) harmløse. Ydelseskraft indebærer næsten altid fare.
>
> DET VIL ALDRIG ÆNDRE SIG, LIGESOM DET ALDRIG VIL ÆNDRE SIG MED LYSETS EKSISTENS.

Nogle tror, at det snart bliver bedre. Skidt. Der er tekniske og konceptuelle grænser, som ikke kan elimineres.

Med **Agentic Coding** gør i det kraftige AI-program til en slags håndværker, som i instruerer om, hvor på harddisken (eller i intranet eller internet) jeres kildekode ligger. Så skriver i en kommando, f.eks. "Tilføj en vedligeholdelsesvisning, så I kan administrere nyhedsabonnenter". AI-agenten arbejder derefter stille og roligt på basis af jeres kode, ændrer lidt eksisterende kode og tilføjer nyt. Til sidst håber I, at det ønskede resultat er nået.

Den proces med at programmere AI ved hjælp af agenter er maksimalt ugrennemskuet. En midlertidig fase er, når agenterne spørger om hver enkelt planlagt ændring i jeres programkode, før de gøres. Men det fungerer ikke længe. Så snart efter den femte henvendelse vil du **aktivere Autopiloten** og være færdig.

Istinden for Agentic Coding; *bedre måde at programmere med AI på i stedet for giver det en bedre tilgang, til at arbejde effektivt med AI. Produktivitetsforbedringen er på faktor 5*, efter vores erfaringer og feedback fra uddannede udviklerhold.

Konklusion

Programmer, der har omfattende adgang til andre systemer, har denne adgang, fordi de skal bruges. Ellers ville adgangen ikke med vilje blive givet.

Et program, der skal vurdere e-mails, skal og må kunne læse disse e-mails. Hvorledes dette alene kan fare hen, viser virkeligheden. Angriber kan gemme anvisninger i e-mails, som kan manipulere AI-agenter.

Hvorfor kan AI-agenter manipuleres? Fordi de er højst plastiske, intransparente systemer, der ikke fungerer regelbaseret, men opgavebaseret.

AI-agenter får ikke regler givet på forhånd. AI-systemer får heller ikke regler givet på forhånd. De lærer disse regler ud fra eksempler selv. Det kaldes AI-træning.

Derfor er AI-agenter potentielt meget effektive og potentielt meget farlige: De er "intelligente" og kan også ofte løse ukendte problemstillinger på en fremragende måde. Samtidig er de magtfulde, fordi de er blevet tildelt omfattende beføjelser.

Hver, der venter på en løsning på dette problem, kan vente længe. I stedet bør man tage stilling: Give adgang til alle mulige systemer ELLER acceptere et acceptabelt risiko. Begge dele samtidig findes praktisk talt ikke.

Også dem, der vælger Agentic Coding, går det galt at vælge den forkerte vej. For lidt programmeringskompetence erstattes af farlige agentsystemer, der leverer ikke-gennemskuelige resultater.

Løsningen er: Et solidt fundament af kompetence kombineret med den passende anvendelse af AI.

About the author on dr-dsgvo.de

My name is Klaus Meffert. I have a doctorate in computer science and have been working professionally and practically with information technology for over 30 years. I also work as an expert in IT & data protection. I achieve my results by looking at technology and law. This seems absolutely essential to me when it comes to digital data protection. My company, IT Logic GmbH, also offers consulting and development of optimized and secure AI solutions.