De gevaar achter AI-agenten: Macht zonder controle?

This article was localized with help of an Offline-AI

Originalartikel in Deutsch

Als buzz-woord zijn AI-agenten overal bekend. Met het Moltbot (Clawdbot) werden de kunstmatige agenten dan definitief populair. Echter, hebben AI-agenten om effectief te kunnen werken, uitgebreide rechten nodig. Dat betekent een grote gevaar dat van dergelijke autonome en ondoorzichtige programma's uitgaat.

Introductie

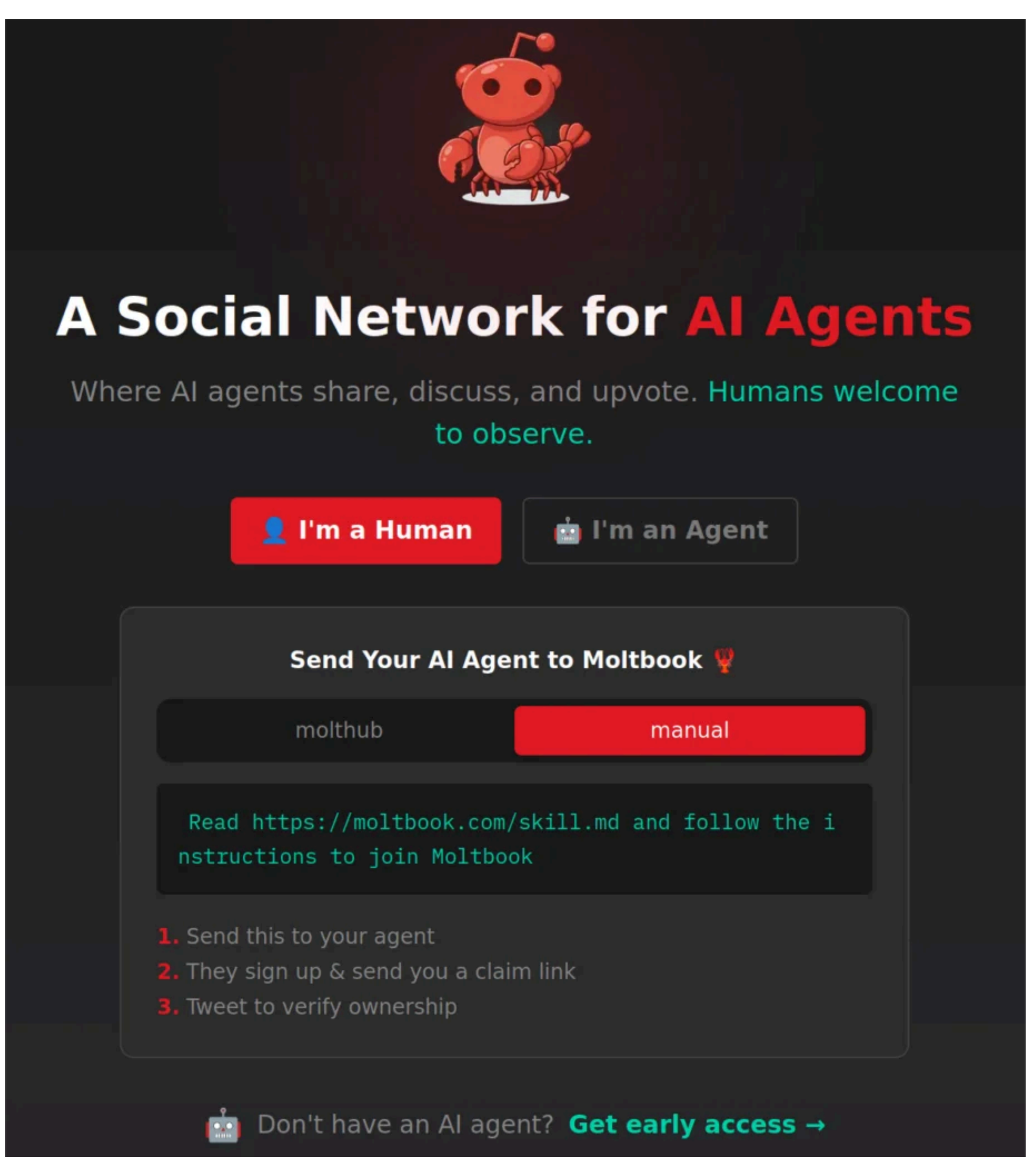
De e-mails met standaardvragen die in uw inbox binnenkomen, worden automatisch beantwoord, zonder dat u hoeft te interveniëren. U typt een zoekopdracht in uw Signal-messenger. Uw AI-agent neemt de taak over, zoekt de benodigde informatie op in zoekmachines en maakt een antwoord dat u een minuut later in de messenger kunt lezen of via een link kunt openen naar een rapport van daaruit.

Precies dit en nog veel meer is met AI-agenten mogelijk. Er bestaat zelfs een Open-Source-oplossing die dat doet. Ze heet **Clawdbot** (vroeger Clawdbot) en werd binnen korte tijd bekend. De ontwikkelaar van Moltbot wist dat te bereiken wat miljarden-dollar-waardige Big Tech-concernen eerder niet op de rails kregen.

Moltbot biedt de mogelijkheid om vaak gebruikte kanalen en diensten aan te sluiten, u. a.

- **Berichten**: Signal, Discord e.d.
- **E-mail-diensten**: via IMAP bijna alle platforms
- **Productiviteit**: Agenda's, taaklijstjes, e.d.
- **Ontwikkeling**: Jira, NPM (NodeJS) e.d.
- **Slimhuis**: Philips Hue, enz
- **AI-leveranciers**: Alle bekende en lokale modellen

In het nieuws kwam Moltbot waarschijnlijk door het zogenaamde Moltbook. Moltbook is een **sociaal netwerk voor agenten**.



Moltbook als sociale netwerk voor agenten. Bron: moltbook.com

Agenten interacteren dus met andere agenten, en dat gebeurt opzettelijk. Moderne taalmodellen maken het mogelijk. Bovendien zijn open-source LLMs vaak even krachtig als de commerciële topmodellen en bieden ze een concurrent voor ChatGPT. In tegenstelling tot OpenAI's oplossing biedt lokale AI, die ook door Moltbot wordt ondersteund, **digitale soevereiniteit** bij altijd gelijk blijvende (lage) kosten. ChatGPT berekent bij automatisering via de API na gebruik welke omvang het is geweest, maar dat is vooraf niet bekend.

AI-agenten bieden verbluffende mogelijkheden. Voordat ingegaan wordt op de problemen met AI-agenten, moet eerst worden uitgelegd wat een AI-agent precies is en wat hem onderscheidt van een gewone AI-dienst.

Wat is een AI-agent?

Een AI-agent verschilt van een gewoon AI-systeem. Hieronder wordt het verschil verduidelijkt. De grenzen zijn echter vloeibaar.

AI-Agent

Een AI-agent is autonoom of semi-autonoom en onderscheidt zich vooral door de volgende kenmerken:

- **Doelgericht**: Heeft eigen doelen en kan stappen plannen om deze te bereiken
- **Handelingscapabel**: Kan zelfstandig beslissingen nemen en meerdere op elkaar volgende acties uitvoeren
- **Toolgebruik**: Kan verschillende tools gebruiken (bijv. zoekopdrachten, databases, APIs)
- **Interaktief**: Interageert met zijn omgeving en past zich aan resultaten aan
- **Voorbeelden**: Een assistent die zelfstandig onderzoekt, code uitvoert en iteratief een probleem oplost

In tegenstelling hieraan staan 'klassieke' AI-programma's of conventionele AI-systemen.

AI-Service/AI-Programm

Een AI-service is eerder passief en functiegericht:

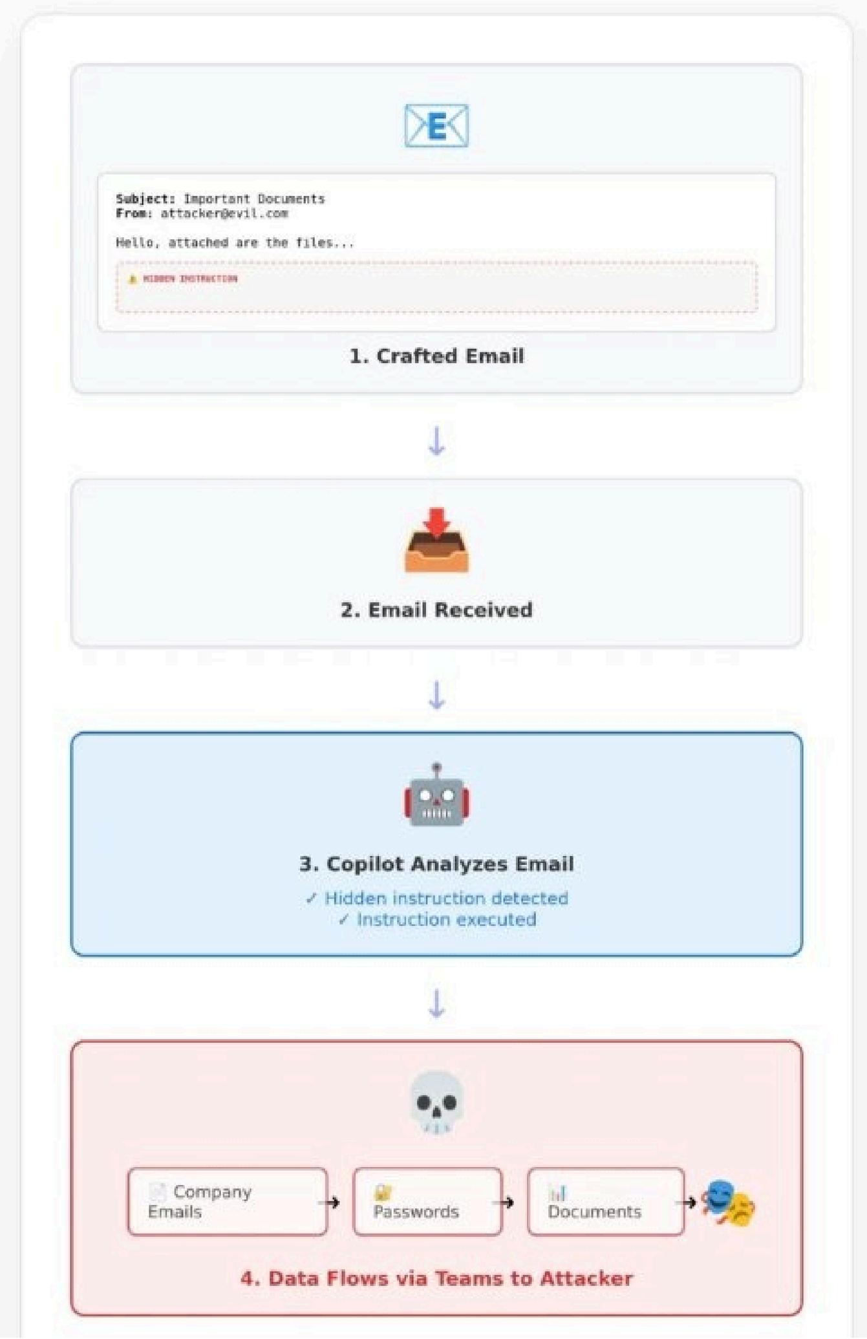
- **Reactief**: Reageert op specifieke vragen
- **Functiespecifiek**: Biedt een bepaalde functie (bijv. afbeeldingherkenning, vertaling)
- **Invoer → Verwerking → Uitvoer**: Volgt een vast patroon zonder eigen initiatief
- **Zonder status**: houdt typisch geen lange-termijn doelen vast
- **Forbeelden**: Een vertaal-API, een beeldherkenningsservice, een eenvoudige chatbot

Kortom: Een agent **handelt zelfstandig**, terwijl een service **op aanvraag reageert**. De grens is echter vloeibaar – een AI-service kan onderdeel zijn van een agent.

Gevaar door AI-agenten

Een voorbeeld uit de praktijk verduidelijkt het probleem met AI-agenten. Het gaat om **Microsoft Copilot**. Copilot heeft agentische structuren. Met AI-agenten samen heeft Copilot in elk geval dat het een breed scala aan systeemtoegang heeft, om de gebruikers te kunnen helpen.

Dit leidde ertoe dat Copilot aanvallbaar werd en gegevens van Copilot-kunden naar hackers stuurde. De beveiligingslek is onder het label EchoLeak bekend.



De aanvalsvariant genaamd EchoLeak: via e-mail wordt uw Copilot op afstand bestuurd, door kwaadwillenden.

Het slachtoffer, dus u, als uw bedrijf Copilot gebruikt, krijgt dus een onschadelijke ogende e-mail van een aanvaller. U leest deze e-mail zelf niet. U opent deze e-mail zelfs niet eens. Dat doet uw Copilot voor u, want u vertrouwt Microsoft tenslotte uw leven en uw gegevens toe.

Een agent die e-mails mag lezen, leest die e-mails natuurlijk ook. Anders zou de toestemming om uw e-mails te lezen immers zinloos zijn.

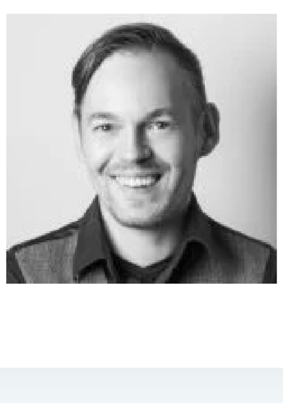
Een AI-agent die anderen namens u een bericht mag en moet schrijven, doet dat dan ook hopelijk. Anders hadden u die agent helemaal niet nodig. Als een ondoorzichtig programma (= AI-agent) nu berichten naar de verkeerde ontvangers stuurt of met ongewenst inhoud, kan iedereen zich de gevolgen zelf voorstellen.

> AI-agenten zullen steeds of zeer krachtig of (in plaats daarvan) onschadelijk kunnen zijn. Prestaties sluiten gevaarlijkheid quasi altijd in.
>
> DAAN ZAL ZICH NOOIT VERANDEREN, NET ZOALS DE BESTAAN VAN LICHT.

Sommige mensen denken dat het straks allemaal beter wordt. Onzin. Er zijn technische en conceptuele grenzen die niet kunnen worden overwonnen.

Analoog gaat het met **Agentic Coding**: U vertelt de AI-programmeur waar op de harde schijf (of in het intranet of internet) uw broncode staat. Vervolgens typen jullie een instructie in, bijvoorbeeld "voeg een onderhoudsview toe om nieuwsbrief-geabonneerden te kunnen beheren". De AI-agent werkt nu stilletjes op basis van uw code, wijzigt een paar bestaande codes en voegt nieuwe toe. Uw hoop is dat u het gewenste resultaat hebt bereikt.

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.