Il pericolo dietro gli agenti AI: potere senza controllo?

This article was localized with help of an Offline-AI

Originalartikel in Deutsch

Come parola d'ordine, gli agenti di intelligenza artificiale sono alla moda. Con il Moltbot (Clawdbot) sono stati infine resi popolari i robot artificiali. Tuttavia, gli agenti di intelligenza artificiale richiedono ampi poteri per funzionare efficacemente. Ciò comporta un grande pericolo che proviene da tali programmi autonomi e intransparenti.

Introduzione

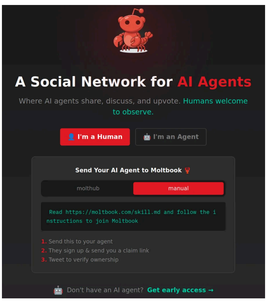
Le email con richieste standard che arrivano nella tua casella di posta in arrivo vengono risposte automaticamente, senza che tu debba intervenire. Scrivi un compito di ricerca nel tuo messaggistica Signal, il tuo agente AI si assume il compito, cerca le informazioni necessarie nei motori di ricerca e crea una risposta che puoi leggere un minuto dopo nel messaggistica o aprire tramite un link a un report da lì.

Esattamente questo e molto altro è possibile con gli agenti di intelligenza artificiale. Esiste anche una soluzione open-source che lo fa. Si chiama **Clawdbot** (in precedenza Clawdbot) e si è fatta presto notare. Lo sviluppatore di Moltbot ha raggiunto ciò che le multinazionali del big tech miliardarie non sono riuscite a fare prima.

Il Moltbot offre la possibilità di collegare frequentemente utilizzati canali e servizi, ad esempio.

- **Messaggistica**: Signal, Discord ecc.
- **Servizi di posta elettronica**: su IMAP quasi tutte le piattaforme
- **Produttività**: Calendario, liste di compiti, ecc.
- **Sviluppo**: Jira, NPM (NodeJS) ecc.
- **Casa intelligente**: Philips Hue, altre
- **Fornitori di AI**: Tutti i modelli noti e locali

Nella notizia è stato menzionato il Moltbot, probabilmente a causa del cosiddetto Moltbook. Il Moltbook è un **rete sociale per gli agenti**.



Moltbook come social network per agenti. Fonte: moltbook.com

Gli agenti interagiscono quindi con altri agenti, in modo programmato. I moderni modelli linguistici lo rendono possibile. Inoltre, i LLM open-source sono spesso altrettanto performanti dei loro omologhi commerciali e mettono così alla prova ChatGPT. A differenza della soluzione di OpenAI, la IA locale, supportata anche da Moltbot, offre una **sovranità digitale** totale a costi sempre ugualì (e bassi). Al contrario, ChatGPT addebita per l'uso automatizzato tramite l'API, ma il limite d'uso non è noto prima dell'utilizzo.

Gli agenti AI offrono possibilità sorprendenti. Prima di affrontare la problematica degli agenti AI, è necessario chiarire cosa sia un agente AI e in cosa si differenzi da un normale servizio AI.

Cosa è un agente di IA?

Un agente AI si differenzia da un normale sistema AI. La seguente illustra la differenza. I confini sono però sfumati.

Agente AI

Un agente AI è autonomo o semiautonomo e si distingue in particolare per i seguenti tratti:

- **Obiettivo orientato**: Ha propri obiettivi e può pianificare passi per raggiungerli
- **Capacità di azione**: Può prendere decisioni autonome e eseguire più azioni consecutive
- **Utilizzo di strumenti**: Può utilizzare diversi tool (ad esempio ricerca web, database, API)
- **Interattivo**: Interagisce con il suo ambiente e si adatta ai risultati
- **Esempi**: Un assistente che ricerca autonomamente, esegue codice e risolve iterativamente un problema

Al contrario, ci sono i "classici" programmi di intelligenza artificiale o i sistemi di intelligenza artificiale tradizionali.

AI-Service/AI-Programm

Un servizio di IA è piuttosto passivo e orientato alle funzioni:

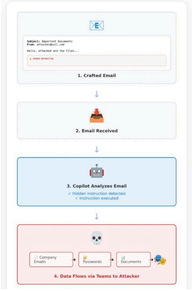
- **Reattivo**: Risponde a specifiche richieste
- **Fonction specifica**: Offre una funzione specifica (ad es. riconoscimento di immagini, traduzione)
- **Ingresso → Elaborazione → Uscita**: Segue un modello fisso senza iniziativa propria
- **Senzastato**: mantiene tipicamente obiettivi a lungo termine non definiti
- **Esempi**: Una API di traduzione, un servizio di riconoscimento immagini, un semplice chatbot

In sintesi: Un agente **agisce in autonomia**, mentre un servizio **risponde a richiesta**. La frontiera è però sfumata – un servizio di intelligenza artificiale può far parte di un agente.

Pericolo derivante dagli agenti AI

Un esempio di pratica illustra il problema con gli agenti AI. Si tratta del **Microsoft Copilot**. Copilot ha strutture agenziali. Con gli agenti AI, comunque, Copilot ha comunque accesso a sistemi ampiamente estesi per poter essere utile agli utenti.

Ciò portò a rendere Copilot vulnerabile e a far inviare dati dei clienti di Copilot agli hacker. La fessura di sicurezza è nota con il hashtag EchoLeak.



Il vettore di attacco chiamato EchoLeak: il vostro assistente virtuale viene controllato a distanza tramite email, da malintenzionati.

La vittima, quindi lei, se il suo azienda utilizza Copilot, riceve quindi un'email apparentemente innocua da un aggressore. Lei stessa non legge questa email. Non la apre nemmeno. Il suo Copilot lo fa per lei, perché alla fine si fida di Microsoft per la sua vita e i suoi dati.

Un agente che ha il permesso di leggere le email, speriamo che le legga anche. Altrimenti, il permesso di leggerle sarebbe insensato.

Un agente AI che è autorizzato a scrivere un messaggio per qualcun altro dovrebbe farlo. Altrimenti, non avrete bisogno di questo agente. Se un programma in trasparente (= agente AI) invia ora messaggi ai destinatari sbagliati o con contenuti indesiderati, ognuno può immaginare le conseguenze.

> Gli agenti AI saranno sempre o molto potenti o (alternativamente) innocui. La capacità include quasi sempre un grado di pericolosità.
>
> NULLA POTRÀ MAI CAMBIARE QUESTO FATTO, PROPRIO COME L'ESISTENZA DELLA LUCE NON POTRÀ MAI CAMBIARE.

Alcuni credono che presto andrà tutto meglio. Merda. Ci sono limiti tecnici e concettuali che non possono essere eliminati.

Con **Agentic Coding**: dite al programmatore di intelligenza artificiale dove si trovano i vostri codici sulla cartella (o nel intranet o internet). Poi digitate una istruzione, ad esempio "aggiungi un'area di manutenzione per gestire gli abbonati ai newsletter". L'agente lavora quindi in silenzio sulla vostra base di codice, modifica alcuni codici esistenti e aggiunge nuovi. Alla fine sperate di aver ottenuto il risultato desiderato.

Questo processo di programmazione mediante agenti AI è estremamente trasparente. Una fase intermedia sono gli agenti che, ogni volta che si prevede una modifica del codice del programma, chiedono conferma al programmatore. Ma non dura a lungo. Dopo la quinta richiesta, **attiverete l'autopilota** e sarete fuori controllo.

Invece di utilizzare il codice Agente, esiste un **migliore approccio** per programmare con efficacia l'intelligenza artificiale. La **produttività aumenta del fattore 5**, secondo le nostre esperienze e le risposte dei team di sviluppatori formati.

Conclusione

I programmi che hanno ampi accessi ad altri sistemi, hanno tali accessi perché sono destinati ad essere utilizzati. Altrimenti, l'accesso non sarebbe concesso consapevolmente.

Un programma progettato per analizzare le email deve, e deve essere in grado di, leggere tali email. La realtà dimostra a cosa possa portare questo da solo. Gli aggressori possono inserire istruzioni nelle email che possono manipolare gli agenti AI.

Perché gli agenti AI possono essere manipolati? Perché si tratta di sistemi altamente plastici e opachi che non funzionano in base a regole, ma in base a compiti.

Gli agenti AI non ricevono regole predefinite. I sistemi AI non ricevono regole predefinite. Imparano queste regole analizzando esempi, un processo chiamato training dell'IA.

Pertanto, gli agenti AI sono potenzialmente molto potenti e potenzialmente molto pericolosi: sono "intelligenti" e possono anche risolvere spesso in modo eccellente problemi sconosciuti. Allo stesso tempo, sono potenti perché sono stati concessi loro ampi poteri.

Chiunque aspetti una soluzione a questo problema dovrà aspettare a lungo. Invece, bisogna decidere: concedere l'accesso a tutti i possibili sistemi ODER accettare un rischio accettabile. Entrambe le cose insieme sono praticamente impossibili.

Anche chi si affida al coding agentistico, si affida al treno sbagliato. Troppa poca competenza di programmazione viene sostituita da sistemi di agenti pericolosi che producono risultati non comprensibili.

La soluzione è: una solida base di competenze unita all'uso appropriato dell'intelligenza artificiale.

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.