Le danger derrière les agents IA: un pouvoir sans contrôle ?

This article was localized with help of an Offline-AI

Originalartikel in Deutsch

Les agents de l'intelligence artificielle sont à la mode. Avec le Moltbot (Clawdbot), les agents artificiels ont finalement gagné en popularité. Cependant, les agents de l'intelligence artificielle nécessitent des autorisations étendues pour fonctionner efficacement. Cela implique une grande menace provenant de ces programmes autonomes et opaques.

Introduction

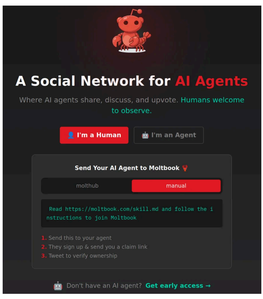
Les e-mails avec demandes standard qui arrivent dans votre boîte de réception sont automatiquement répondus, sans que vous ayez à intervenir. Vous écrivez une tâche de recherche dans votre messager Signal. Votre agent IA prend en charge la tâche, recherche les informations nécessaires dans les moteurs de recherche et crée une réponse que vous pouvez lire une minute plus tard dans le messager ou ouvrir via un lien vers un rapport à partir de là.

Exactement cela et encore bien plus est possible avec des agents AI. Il existe même une solution open-source qui le fait. Elle s'appelle **Clawdbot** (anciennement Clawdbot) et a rapidement acquis une notoriété. Le développeur de Moltbot a réussi ce que les géants du numérique milliardaires n'avaient pas pu faire auparavant.

Moltbot offre la possibilité de lier fréquemment utilisés canaux [1] et services, notamment.

- **Messagerie**: Signal, Discord etc.
- **E-Mail-Dienste:** about IMAP quasi alle Plattformen
- **Productivité**: Calendriers, listes de tâches, etc.
- **Développement**: Jira, NPM (Node.JS) etc.
- **Maison intelligente**: Philips Hue, autres
- Fournisseurs de AI: Tous les modèles connus et locaux

Dans les actualités, c'est probablement le Moltbook qui a attiré l'attention sur Moltbot. Le Moltbook est un **réseau social pour les agents**.



Moltbook, un réseau social pour agents. Source: moltbook.com

Les agents interagissent donc avec d'autres agents de manière planifiée. Les modèles de langage modernes le rendent possible. D'ailleurs, les LLMs open-source sont souvent aussi performants que leurs homologues commerciaux et concurrencent ainsi ChatGPT. En revanche, la IA locale, qui est également soutenue par Moltbot, offre une **souveraineté numérique** totale à des coûts toujours identiques (niedrigen) . ChatGPT, en revanche, facture le recours à l'automatisation via l'API après utilisation, mais la limite d'utilisation n'est pas connue avant utilisation.

Les agents IA offrent des possibilités étonnantes. Avant d'aborder les problèmes liés aux agents IA, il est important de définir ce qu'est un agent IA et ce qui le distingue d'un service IA ordinaire.

Qu'est-ce qu'un agent IA ?

Un agent IA se distingue d'un système d'IA ordinaire. Ce qui suit illustre la différence. Les frontières sont toutefois floues.

Agent IA

Un agent AI est autonome ou semi-autonome et se distingue notamment par les caractéristiques suivantes :

- **Objectif-orienté**: A ses propres objectifs et peut planifier des étapes pour les atteindre
- **Capable d'agir**: Peut prendre des décisions autonomes et exécuter plusieurs actions consécutives
- **Utilisation d'outils:** Peut utiliser différentes outils (par exemple, recherche sur le web, bases de données, APIs)
- **Interactif**: Interagit avec son environnement et s'adapte à ses résultats
- **Exemples**: Un assistant qui recherche de manière autonome, exécute du code et résout itérativement un problème

En revanche, se trouvent les «programmes AI classiques» ou les «systèmes AI traditionnels».

AI-Service/AI-Programm

Un service d'IA est plutôt passif et axé sur la fonctionnalité:

- Reactivé: Réagit à des demandes spécifiques
- **Fonction spécifique**: Offre une fonction spécifique (par exemple reconnaissance d'image, traduction)
- **Entrée → Traitement → Sortie**: Suivent un modèle fixe sans initiative propre
- **Etat sans lien**: conserve généralement des objectifs à long terme
- **Exemples**: Une API de traduction, un service de reconnaissance d'image, un simple chatbot

En bref: Un agent **agit de manière autonome**, tandis qu'un service **réagit sur demande**. La frontière est cependant floue – un service AI peut faire partie d'un agent.

Danger posé par les agents IA

Un exemple concret illustre le problème posé par les AI Agentes. Il s'agit de **Microsoft Copilot**. Copilot possède des structures agentives. Avec les AI-Agentes, en tout cas, il a des accès systèmes étendus pour apporter un bénéfice aux utilisateurs.

Cela a conduit à ce que Copilot soit vulnérable et ait envoyé des données de clients de Copilot aux hackers. La faille de sécurité est connue sous le nom d'EchoLeak.

Le vecteur d'attaque nommé EchoLeak: votre assistant virtuel est contrôlé à distance par des méchants via des e-mails.

La victime, donc vous, si votre entreprise utilise Copilot, reçoit donc un e-mail apparemment inoffensif d'un attaquant. Vous ne lisez même pas cette e-mail. Vous ne l'ouvrez même pas. Votre Copilot le fait à votre place, car vous faites finalement confiance à Microsoft pour votre vie et vos données.

Un agent qui a le droit de lire des e-mails devrait espérer les lire. Sinon, le droit de les lire serait absurde.

Un agent IA qui est autorisé et censé écrire des messages à la place de quelqu'un, devrait le faire. Sinon, vous n'auriez pas besoin de cet agent. Si un programme intransparent (= agent IA) envoie maintenant des messages aux mauvais destinataires ou avec un contenu indésirable, chacun peut imaginer les conséquences.

> Les agents IA seront soit toujours très performants, soit (à la place) inoffensifs. La puissance implique presque toujours un danger.
>
> CELA NE CHANGERA JAMAIS, TOUT COMME L'EXISTENCE DE LA LUMIÈRE.

Certains pensent que tout ira mieux bientôt. C'est de la merde. Il existe des limites techniques et conceptuelles qui ne pourront pas être éliminées.

Il se passe la même chose avec **Codage Agentiel**: vous dites au programmeur AI où se trouvent vos fichiers sources (ou sur le réseau interne ou internet), puis vous tapez une instruction, par exemple «ajoutez une vue de maintenance pour gérer les abonnés à des newsletters». L'agent AI travaille ensuite discrètement en fonction de votre base de code, modifie quelques codes existants et ajoute de nouveaux. Au final, vous devriez avoir obtenu le résultat souhaité.

Ce processus de programmation d'intelligence artificielle à l'aide d'agents est maximalement opaque. Une étape intermédiaire sont les agents qui demandent votre accord comme développeur pour chaque modification prévue de leur code source. Mais cela ne fonctionne pas longtemps. Au plus tard après la cinquième demande, vous **activerez le pilote automatique** et serez livrés.

Au lieu de codage agenciel, il existe un **meilleur approche** pour programmer avec efficacité l'intelligence artificielle. La **productivité est multipliée par 5**, selon nos expériences et les retours des équipes d'ingénieurs formées.

Conclusion

Les programmes ayant des accès étendus à d'autres systèmes possèdent ces accès car ils sont destinés à être utilisés. Sinon, l'accès ne serait pas intentionnellement accordé.

Un programme conçu pour analyser des e-mails doit pouvoir les lire. Les conséquences potentielles de cette capacité sont illustrées par la réalité. Des attaquants peuvent insérer des instructions dans des e-mails susceptibles de manipuler les agents IA.

Pourquoi les agents IA peuvent-ils être manipulés ? Parce qu'il s'agit de systèmes hautement plastiques et opaques qui ne fonctionnent pas de manière basée sur des règles, mais de manière basée sur des tâches.

Les agents IA ne reçoivent pas de règles prédéfinies. Les systèmes d'IA ne reçoivent pas de règles prédéfinies. Ils apprennent ces règles à partir d'exemples eux-mêmes. Cela est appelé entraînement d'IA.

C'est pourquoi les agents IA sont potentiellement très performants et potentiellement très dangereux: Ils sont «intelligents» et peuvent également résoudre souvent d'excellentes solutions à des problèmes inconnus. Parallèlement, ils sont puissants car on leur a accordé des autorisations étendues.

Celui qui attend une solution à ce problème risque d'attendre longtemps. Il faut plutôt choisir: accorder l'accès à tous les systèmes possibles OU accepter un risque acceptable. Il est pratiquement impossible de faire les deux en même temps.

Même ceux qui optent pour le codage agentic misent sur la mauvaise voie. Une compétence de programmation insuffisante est remplacée par des systèmes d'agents dangereux qui produisent des résultats non explicables.

La solution est: un solide fondement de compétences associé à une utilisation appropriée de l'IA.

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 optimised and secure AI solutions.