Небезпека за AI-агентами: влада без контролю?

This article was localized with help of an Offline-AI

Originalartikel in Deutsch

Як хіт парадом AI-агенти стали дуже популярними. З появою Moltbot (Clawdbot) вони остаточно здобули популярність серед людей. Проте, щоб ефективно працювати, їм потрібні досить широкі права доступу. Це створює велику небезпеку, яку представляють такі автономні та непрозорі програми.

Вступ

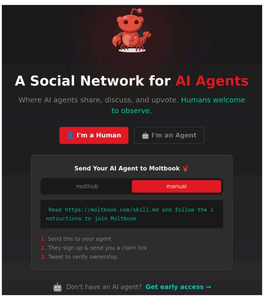
Електронні листи з типовими запитами, які надходять у вашу поштову скриньку, автоматично відповідаються, не вимагаючи вашого втручання. Ви пишете запит на дослідження у свій Signal-месенджер. Ваш AI-агент бере на себе завдання, шукає потрібну інформацію в пошукових системах та створює відповідь, яку ви можете прочитати через хвилину в месенджері або відкрити по посиланню на звіт звідти.

Точно це та ще багато іншого можливо з кібератентами. Є навіть відкритий кодовий проект, який виконує цю функцію. Він називається **Moltbot** (раніше Clawdbot) і швидко став відомим. Власник проекту Moltbot зміг досягти того, чого не змогли мільярдні корпорації Big Tech.

Молтбот надає можливість зв'язуватися з часто використовуваними каналами та послугами, зокрема.

- **Месенджи**r: Signal, Discord тощо.
- **Електронна пошта**: майже всі платформи за допомогою IMAP
- **Виробнича ефективність**: Календар, списки завдань тощо.
- **Розвиток**: Java, NPM (NodeJS) тощо.
- **Розумний будинок**: Philips Hue, Imal
- **ШІ-розробники**: Всі відомі та місцеві моделі

У новинах про Молтбот, йомовірно, було повідомлено через так званий Молтбук. Молтбук - це **соціальна мережа для агентів**.



Moltbook як соціальна мережа для агента.Джерело: moltbook.com

Агенти взаємодіють із іншими агентами, в саме планомірно. Сучасної моделі роблять це можливим. Зауважте, що відкритої джерела LLM часто мають подібну ефективність до комерційних лідерів і створюють конкуренцію для ChatGPT. У зростанні з рішення OpenAI місцевий AI, який також підтримує Moltbot, забезпечує повну **цифрову незалежність** при завжди рівних (низьких) витратах коштів. ChatGPT, навпаки, рахує за автоматизоване використання через API після використання, яке раніше не відомо.

ІА-агенти пропонують приголомшливі можливості. Перш ніж розглянути проблеми, пов'язані з ІА-агентами, слід визначити, що таке ІА-агент і в чому його відмінність від звичайної AI-служби.

Що таке агент штучного інтелекту?

Кі-агент відрізняється від звичайної системи штучного інтелекту. Наступне інострус цей розрив. Однак межі між ними є розмитими.

II-агент

ІНШІ ОСОБЛИВОСТІ Ю АГЕНТА: Агент штучного інтелекту є автономними або напівавтономними і має такі особливості:

- **Завданнюрієнтований**: Має власні цілі та можливості для їх досягнення
- **Рішучий**: Може самостійно приймати рішення та виконувати декілька послідовних дій
- **Використання інструментів**: Може використовувати різні інструменти (наприклад, пошук в інтернеті, бази даних, API)
- **Інтерактивність**: Інтерагує зі своїм середовищем та підлаштовуєтьсяся на результати
- **Наприклад**: Допомога, яка самостійно здійснює пошук інформації, виконує код та ітеративно вирішує проблему

Натомість стоять «класичні» програми штучного інтелекту або традиційні системи штучного інтелекту.

AI-Service/AI-Programm

Сервіс на основі штучного інтелекту є радше пасивним і функціональним:

- **Реактивний**: Відповідає на конкретні запитання
- **Функціонально**: Надає певну функцію (наприклад, розпізнавання зображень, переклад)
- **Вхід → Обробка → Вихід**: Підлягає фіксованому шаблону без власної ініціативи
- **Захист**: звичайно не уберігає довгострокових цілей
- **Наприклад**: Уявлення API, служба розпізнавання зображень, простий чат-бот

Коротко кажучи: Агент **рухує самостійно**, тоді як Сервіс **реагує на запит**. Поріг між ними є плавним - ІО-Сервіс може бути складовою частиною агента.

Небезпека від AI-агентів

Наприклад, практичний приклад інострус проблему з AI-агентами. Проходить мова про **Microsoft Copilot**. Copilot має агентні структури. Разом із AI-агентами Copilot має широкий доступ до системи для надання користувачам користування.

Це призвело до того, що Copilot став вразливим і відправляв дані клієнтів Copilot хакерам. Безпекова лага відомі під шоломордом EchoLeak.

Атака через "EchoLeak": ваш помічник керується дистанційно через електронну пошту зловмисник.

Жертва, тобто ви, якщо ваше підприємство використовує Copilot, отримує здавалося б нешкідливий електронний лист від хакера. Ви самі не читаєте цю електронну пошту. Ви навіть не відкриваєте її. Ваш Copilot робить це за вас, адже ви, зрештою, довіряєте Microsoft своє життя та дані.

Агент, який має право читати електронні листи, сподівається, що він їх і читає. Інакше право читати ваші електронні листи було б безглуздим.

Якщо AI-агент має право та повинен відправляти повідомлення від імені інших, то він має це робити. Інакше вам взагалі не потрібен такий агент. Якщо ж незрозумілий алгоритм (тобто AI-агент) відправляє повідомлення невірним одержувачам або з небажаним змістом, то кожен може собі уявити наслідки.

> II-агенти будуть або завжди дуже потужними, або (замість цього) безневинними. Висока ефективність майже завжди супроводжується небезпекою.
>
> ЦЕ НІКОЛИ НЕ ЗМІНИТЬСЯ, ТАК САМО ЯК І ІСНУВАННЯ СВІТЛА.

Деякі вважають, що незабаром все стане краще. Мерда. Існують технічні та концептуальні межі, які не можна усунути.

Аналогічно відбувається із **Agentic Coding**: Ви вказуєте програмісту роботові, де на жорсткому диску (або в внутрішній мережі чи інтернеті) знаходяться ваші джерельні тексти. Потім ви вводите команду, наприклад «Додайте сторінку підтримки, щоб можна було керувати підписниками нових». Робот-робот працює тихо і безшумно на основі вашого коду, змінюючи деякі існуючі коди та додаючи нові. У кінцевому підсумку ви часте надію, що досягнете бажаного результату.

Цей процес розробки Кі-програми зі застосуванням агентів дуже не прозорий. Між етапами є агенти, які запитують про кожну заплановану зміну вашого коду програми, чи ви як розробник згоди з цим. Але це трівіатіме недовго. Зазвичай після п'ятого запитання ви **активуєте автопілот** і вже не зможете нічого змінити.

Натомість агентної програмування існує **більш ефективний підхід**, щоб працювати з ХІ з високою швидкістю. **Пов'язана продуктивність зростає в 5 разів**, згідно зі своїми дослідами та відгуками наявних команд розробників.

Висновок

Програми, які мають широкі права доступу до інших систем, отримують ці права для того, щоб їх використовувати. Інакше доступ би не був свідомо наданий.

Програма, призначена для аналізу електронних листів, повинна мати можливість читати ці листи. На що це може призвести само по собі, показує реальність. Зловмисники можуть приховувати в електронних листах інструкції, які можуть маніпулювати AI-агентами.

Чому можна маніпулювати AI-агентами? Тому що вони є високопластичними, не прозорими системами, які працюють не на основі правил, а на основі завдань.

II-агенти не отримують заданих правил. Системи штучного інтелекту не отримують заданих правил. Вони навчаються цим правилам на основі прикладів самостійно. Це називається тренуванням II.

Тому AI-агенти можуть бути як дуже потужними, так і дуже небезпечними: вони "розумні" і часто чудово вирішують навіть неявні проблеми. Одночасно вони мають велику владу, оскільки їм надано широкі повноваження.

Кожен, хто чекає на зміну цієї проблеми, може довго чекати. Замість цього потрібно прийняти рішення: надати доступ до всіх можливих систем АБО прийняти прийняття ризико. Одночасно і те, і інше практично неможливо.

Та навіть той, хто обирає Agentic Coding, робить неправильний вибір. Недостатня програмувальна компетенція замінюється небезпечними агенційними системами, які надають незрозумілі результати.

Рішення полягає у міцній базі компетенцій, поєднаній з належним використанням II.

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.