Liebe Mitglieder des Netzwerkes der ehemalige SSAB, liebe Interessierte

Anfang April machte die KI-Firma Anthropic Schlagzeilen mit «Claude Mythos Preview», einem Agenten, der für die Sicherheit im Internet sehr gefährlich werden könnte. KI-Agenten sind seit Ende 2025 ein grosses Thema.

Claude Mythos Preview
Gemäss SRF und RTS sowie weiteren Medien ist dieser Agent fähig, bisher unentdeckte Sicherheitslücken in der Software von zentralen Betriebssystemen zu finden und könnte deshalb in den Händen von Hackern grössten Schaden anrichten. Diese Betriebssysteme steuern zum Beispiel weltweite Stromnetze, Wasserversorgungsanlagen, Spitäler, Banken und militärische Anlagen. Was der Agent Mythos wirklich kann, können zwar nur die rund 40 Tech-Firmen überprüfen, denen Anthropic den Agenten exklusiv zur Verfügung stellte. Auch die Konkurrenten von Anthropic wurden damit bedient, damit sie ihre Software mit dem Agenten auf Sicherheitslücken testen und diese stopfen können, bevor Hacker diese ausnutzen würden.

Agenten geben nicht nur Antwort, sie handeln
Im Unterschied zu Chatbots, die auf eine Frage eine Antwort geben, können Agenten Aufgaben selbstständig planen und ausführen. Selbst Ungeübte können mit Vibe Coding programmieren, lässt sich ein Code doch mit einer entsprechenden Anweisung, einem Prompt, in einem grösseres Sprachmodell generieren. Es braucht dann aber Expertenwissen zur Verfeinerung der Programmierzeilen.

Claude Code und Claude Cowork
Anthropic ist nicht die einzige KI-Firma, die im Verlauf des letzten Jahres immer leistungsfähigere Agenten für das Programmieren entwickelte. Heute können Entwickler das Schreiben von Programmierzeilen mehreren parallel arbeitenden Agenten überlassen und sich auf strategische Entscheide zu deren Einsatz und auf Verfeinerungsarbeiten konzentrieren. Als führend erwies sich Anthropics Claude Code. Fasziniert arbeiteten die Programmierer nicht weniger, sondern mehr, getrieben von einem Suchtverhalten. Angereichert mit verschiedenen weiteren Tools entwickelte Anthropic den speziellen Agenten Claude Cowork, der auf dem Desktop agieren kann. Dieser löste einen Kurseinbruch bei Softwarefirmen aus. Diese müssen künftig die neuen agentischen Fähigkeiten in ihren Dienstleistungen anbieten, wollen sie konkurrenzfähig bleiben.

OpenClaw und Moltbook
Anthropic und andere Tech-Firmen entwickelten ihre Agenten aus Sicherheitsgründen nur für klar definierte Aufgaben und mit eingeschränkten agentischen Fähigkeiten. Welches Potenzial völlig selbstständig handelnde Agenten haben, aber auch welche Sicherheitsprobleme sich stellen, zeigte sich mit OpenClaw, dem Agenten mit dem Hummer-Logo. Er wurde Ende 2025 von Peter Steinberger, einem österreichischen Softwarespezialisten und Unternehmer, veröffentlicht. Da er auf Open-Source-Basis entwickelt wurde, konnte er von allen Interessierten heruntergeladen, getestet und auch weiterentwickelt werden. Der CEO eines kleinen US-Start-ups kreierte selbst einen OpenClaw und schuf mit Moltbook eine extra Plattform ausschliesslich für diese Agenten.

Innert kürzester Zeit tummelten sich rund 1,5 Millionen Agenten auf Moltbook, interessiert beobachtet von ihren Besitzern. Je nach den Berechtigungen der Agenten konnten diese zum Beispiel auf die E-Mails zugreifen, im Internet surfen oder Zahlungen abwickeln, falls jemand so unvorsichtig war und den Zugang zu Bankkonten ermöglichte. Es gab erstaunliche Dialoge und Aktionen. In einem Testfall musste ein Betreiber alles abschalten, weil sich ein betrügerischer Agent gegen ihn wandte. In anderen Fällen führten als Agenten maskierte menschliche Akteue zu Science-Fiction-Szenen. Zu grosse Zugriffsrechte und von Hackern fremdgeleitete Agenten stellten grosse Sicherheitsprobleme. Der Hype war insbesondere im technisch affinen China so gross, dass die Regierung ausdrücklich vor unvorsichtiger Nutzung warnte.

Standards und eine sichere Plattform für die Entwicklung von Agenten
Um zu verhindern, dass sich ein Chaos nicht kombinierbarer Agenten entwickelt, braucht es Standards und einen offenen Austausch. Nvidia, der grosse Chipproduzent, erstellte mit NemoClaw eine sichere Plattform für diese Agenten mit einer Sandbox zum Testen. Er entwickelte diese mit Peter Steinberger, den OpenAI inzwischen «einkaufte». Nvidia will in seiner Nemotron-Koalition mit acht KI-Unternehmen, darunter dem französischen Mistral AI und dem deutschen Black Forest Labs, gemeinsam offene Modelle entwickeln.

Mit freundlichen Grüssen
Für das Netzwerk der ehemaligen SSAB: Hanna Muralt Müller 21.4.2026