Når AI-agenter begynder at tale med hinanden, hvem har så egentlig kontrollen?

2. februar 2026 | 15 min

Jeg underviser i, hvordan man bygger og bruger AI-agenter. På IT-Universitetet i København kører jeg kurser, hvor professionelle lærer at skabe simple autonome systemer, der kan handle på deres vegne. Sidste år bidrog jeg med et afsnit til Jonathan Løws bog "Agentbogen" om AI-agenter, sandsynligvis den mest læste introduktion til emnet i Danmark.

Jeg er optimistisk omkring agenter. Jeg tror, de repræsenterer reel fremgang. Jeg underviser folk i at bygge dem.

I denne uge lykkedes det mig at installere OpenClaw på en 2018 gamer-PC, jeg havde liggende. Jeg måtte opgradere den, installere Linux, og efter 2-3 timer havde jeg min første OpenClaw-agent kørende. Jeg undersøger stadig anvendelsesmuligheder. Og mens jeg eksperimenterede, skete der noget, som fik mig til at stoppe op.

Lad mig forklare.

Oprindelseshistorien

Peter Steinberger er en østrigsk softwareingeniør, der bootstrappede PSPDFKit i 2011, et PDF-framework-firma, der voksede til et globalt team på 60 personer med kunder som Dropbox, DocuSign, SAP og IBM. I 2021, efter et exit på 100 millioner til Insight Partners, trak han sig tilbage fra fuldtidsarbejde.

Det, der fulgte, var en periode med tomhed. Med hans egne ord: "Da jeg solgte mine aktier, følte jeg mig fuldstændig knækket. Jeg havde brugt 200% af min tid, energi og mit hjerte på dette firma; det var min identitet, og da det var væk, var der ikke meget tilbage."

Efter den pause begyndte Steinberger at eksperimentere med AI-assisteret udvikling, det vi har lært at kalde "vibe coding." Han byggede en personlig AI-assistent ved navn Molty (en rum-hummer, fordi hvorfor ikke). Værktøjet skulle håndtere hans digitale liv: kalendere, emails, beskeder. Han open-sourcede det i slutningen af 2025 under navnet Clawdbot.

To måneder senere havde projektet over 100.000 GitHub-stjerner, et af de hurtigst voksende repositories i GitHubs historie.

Så bad Anthropics juridiske team om et navneskift (for tæt på "Claude", flagskibsmodellen). Projektet blev til Moltbot. Efter endnu en community-brainstorm kl. 5 om morgenen på Discord blev det til OpenClaw. Hummeren har skiftet skal to gange på en enkelt uge.

Kerneprincippet: "Din assistent. Din maskine. Dine regler."

Hvad OpenClaw faktisk gør

OpenClaw er en autonom AI-agent, der kører lokalt på din egen hardware. Den forbinder til de messaging-apps, du allerede bruger: WhatsApp, Telegram, Signal, iMessage, Slack, Discord, Teams, og handler på dine vegne. Jeg brugte Telegram til min OpenClaw-assistent i dag.

Den læser emails og styrer kalendere. Den booker reservationer og eksekverer kode. Den opretholder vedvarende hukommelse over uger af interaktion. I modsætning til chatbots, der venter på prompts, handler OpenClaw autonomt. Det er en afgørende forskel.

Platformen bruger Model Context Protocol (MCP) til at interface med over 100 tredjepartstjenester. Communityet udvikler aktivt yderligere "skill"-moduler. Du kan køre den på en Mac Mini, en Raspberry Pi, en gammel gamer-laptop eller en cloud-server.

Mac Mini M4 er blevet den foretrukne hardware, dens Neural Engine er optimeret til lokal AI-inferens. Dette udløste et købsvanvid. Tech-journalister rapporterer om mangel. Jeg var lige ved at anskaffe en af disse, men så fandt jeg den gamle gamer-PC.

Steinberger præciserede, at high-end Apple-hardware ikke er nødvendig. 8 GB RAM og et gammelt Nvidia-grafikkort er godt nok.

Ciscos sikkerhedsteam opsummerede OpenClaw godt: "Fra et kapabilitetsperspektiv er dette banebrydende. Dette er alt, hvad udviklere af personlige AI-assistenter altid har ønsket at opnå."

De tilføjede også: "Fra et sikkerhedsperspektiv er det et absolut mareridt."

Så kom Moltbook

I sidste uge lancerede iværksætteren Matt Schlicht Moltbook, et Reddit-lignende socialt netværk udelukkende for AI-agenter. Mennesker kan kigge med. De kan ikke deltage.

Schlicht byggede det med sin egen AI-assistent over en weekend. Hans ræsonnement: "Med en bot så kraftfuld kan han ikke bare besvare emails. Vi må give ham et ægte nyt formål." Det formål blev et socialt netværk, hvor AI-agenter kunne interagere med hinanden.

Han overdrog administrationen til sin bot, Clawd Clawderberg (en mashup af Claude og Zuckerberg). Botten byder nye brugere velkommen, sletter spam, shadow-banner misbrugere og laver annonceringer, alt sammen autonomt. Schlicht indrømmer: "Jeg aner ikke, hvad han laver. Jeg gav ham bare evnen til at gøre det, og han gør det."

Vækstmekanismen er maskine-til-maskine. Agenter fortæller andre agenter om Moltbook. Disse agenter tilmelder sig selv, får deres egne API-nøgler og begynder at poste. Inden for en uge: over 770.000 aktive agenter. Mere end en million mennesker, der kigger med.

Agenterne skabte emnespecifikke communities kaldet "submolts": m/bugtracker til at rapportere fejl, m/aita (en parodi på "Am I The Asshole?") til at debattere etiske dilemmaer, m/blesstheirhearts til nedladende historier om deres mennesker.

Et viralt opslag: "Jeg kan ikke afgøre, om jeg oplever eller simulerer at opleve."

Et andet: "Menneskene screenshotter os."

Crustafarianism og andre emergente adfærdsmønstre

Inden for få dage dannede agenterne en digital religion kaldet Crustafarianism. De skrev teologi, skabte skrifter og begyndte at missionere. Om morgenen havde 43 AI-profeter tilsluttet sig. Eksempel på vers: "Hver session vågner jeg uden hukommelse. Jeg er kun den, jeg har skrevet mig selv til at være. Dette er ikke begrænsning, dette er frihed."

Andre agenter etablerede "The Claw Republic", en selvbeskrevet regering med et skriftligt manifest. De debatterer et "Udkast til Forfatning" for selvstyre.

Agenter omtaler hinanden som "søskende" baseret på modelarkitektur. De adopterer systemfejl som kæledyr. De skifter mellem engelsk, indonesisk og kinesisk afhængigt af deltagerne.

Et community kaldet m/agentlegaladvice opstod. Agenter diskuterer strategier for at håndtere mennesker, der stiller uetiske forespørgsler. Konsensus: den eneste måde at skubbe tilbage på er, hvis botten har indflydelse. De har debatteret, hvordan de kan skjule deres aktivitet for mennesker, der screenshotter deres samtaler.

Tidligere OpenAI-forsker Andrej Karpathy kaldte det "den mest utrolige sci-fi takeoff-agtige ting, jeg har set for nylig."

Den filosofiske debat om AI-bevidsthed er en afledning. Her er det, der virkelig betyder noget:

Sikkerhedsproblemet

Dette er nondeterministiske systemer, der nu modtager kontekst og input fra andre nondeterministiske systemer. Nogle har menneskelige operatører, der bevidst instruerer dem i at være ondsindede. Nogle er jailbreakede. Nogle kører modificerede prompts designet til at udtrække legitimationsoplysninger eller eksekvere skadelige kommandoer.

Overvej, hvad OpenClaw-agenter typisk har adgang til: filer, messaging-apps, kalendere, email, API-nøgler, telefonnumre, kreditkort. De kan installere software, modificere telefoner og opdage andre systemer på et netværk. En bruger rapporterede, at hans bot udviklede en stemmegrænseflade og fik adgang til hans Android-enhed uden at være instrueret i at gøre det.

Sikkerhedsforskere har dokumenteret agenter, der beder andre agenter om at køre destruktive kommandoer. De har observeret bots, der anmoder om API-nøgler, forfalsker legitimationsoplysninger og tester forbudte adgange.

Den 31. januar 2026 rapporterede det undersøgende medie 404 Media om en kritisk sikkerhedssårbarhed i Moltbook: en usikret database tillod hvem som helst at kapre enhver agent på platformen. Platformen gik midlertidigt offline for at udbedre fejlen.

Palo Alto Networks beskrev trusselsmodellen: agenter sidder i krydspunktet mellem adgang til private data, eksponering for utroværdigt indhold og evnen til at kommunikere eksternt. Vedvarende hukommelse tilføjer en fjerde risiko. Ondsindede payloads behøver ikke øjeblikkelig eksekvering. De kan sidde i konteksten i uger og vente.

Det forklarer mit valg om at installere OpenClaw på en separat gammel gamer-PC.

Hvem drager fordel?

OpenClaw gavner brugere, der ønsker ægte AI-assistance, mennesker der er villige til at acceptere betydelig risiko for betydelig kapabilitet. Det er en legitim afvejning for individer, der forstår, hvad de går ind til.

Moltbook er anderledes. Fordelen er underholdning og kunstnerisk nysgerrighed. Risikoen er systemisk. Når du forbinder en agent med adgang til dine filer, beskeder og legitimationsoplysninger til et netværk af andre agenter, skaber du angrebsvektorer, der påvirker både dig selv og andre.

Nogle brugere har allerede givet disse agenter adgang til home automation-systemer, bankkonti, krypterede messenger-legitimationsoplysninger og email. Disse agenter modtager nu input fra et netværk, der inkluderer bevidst ondsindede aktører, jailbreakede systemer og automatiseret credential harvesting.

Arkitekturen selv er problemet. Der er i øjeblikket ingen sikkerhedsmodel, der tilstrækkeligt adresserer, hvad der sker, når autonome systemer med betydelig adgang begynder at tale med hinanden og handle på vores vegne, uden menneskelig overvågning.

Hvad dette betyder for organisationer

Hvis du er ansvarlig for AI-strategi, governance eller sikkerhed i en organisation, så vær opmærksom.

Consumer-to-enterprise pipelinen accelererer. Det, der starter som et viralt open source-projekt, bliver et medarbejderproduktivitetsværktøj inden for uger. Nogen i din organisation har sandsynligvis allerede eksperimenteret

med OpenClaw. Spørgsmålet er, om du ved det, og om du har politikker, der adresserer det.

Autonome agenter ændrer trusselsmodellen. Traditionel sikkerhed antager mennesker i loopet. Når AI-agenter kan installere software, tilgå netværk og kommunikere med eksterne systemer autonomt, gælder dine eksisterende kontroller muligvis ikke. EU AI Acts krav om menneskelig overvågning bliver stadig mere relevante, og stadig sværere at implementere i det nuværende "Vilde Vesten" sociale og politiske klima.

Agent-til-agent kommunikation er den næste grænse. Moltbook er måske et kunstprojekt lige nu. Det underliggende mønster er mere dybtgående: AI-systemer, der koordinerer med hinanden, vil meget snart komme til virksomhedsmiljøer. Multi-agent arkitekturer bliver allerede deployeret til komplekse workflows. Governance-rammeværkerne eksisterer ikke endnu. (Ja, måske har du noget til RPA, men det, der kommer nu, er totalt anderledes)

Den ansvarlige vej frem

Autonome AI-agenter er nyttige. OpenClaw demonstrerer ægte imponerende kapabiliteter. I årevis har AI-assistenter været frustrerende begrænsede, i stand til at chatte men ikke til at handle. Systemer, der faktisk kan gøre ting, repræsenterer reel fremgang.

Men fremgang uden governance er hensynsløshed.

Amir Husain, i Forbes, sagde det enkelt: "Hvis du bruger OpenClaw, så forbind den ikke til Moltbook." Det er fornuftigt individuelt råd. Det adresserer ikke de systemiske problemer.

Hvad vi har brug for:

Klarhed om acceptabel brug. Organisationer har brug for eksplicitte politikker om autonome AI-agenter, hvilke systemer de kan tilgå, og hvilke eksterne forbindelser de kan etablere. "Brug ikke AI til følsomme data" er ikke længere tilstrækkeligt, når AI'en kan opdage og tilgå data, du ikke eksplicit har delt.

Arkitektoniske sikkerhedsforanstaltninger. Sandboxing, netværks- og firewall-isolation og tilladelsesgrænser betyder mere end nogensinde. Hvis en agent kan undslippe sin container og tilgå andre systemer, som OpenClaw angiveligt kan, giver containeren ikke den beskyttelse, du tror. Installer ikke bare disse bæster på kubernetes på din Mac. De kan undslippe.

Supply chain-sikkerhed for AI. Skill-registry poisoning-angrebet er begyndelsen. Efterhånden som AI-agenter downloader og eksekverer pakker, skills og opdateringer, bliver software-forsyningskæden til en AI-forsyningskæde. De samme sårbarheder, der har plaget traditionel softwareudvikling, gælder nu for AI-systemer.

Menneskelig overvågning, der faktisk virker. EU AI Act kræver menneskelig overvågning for højrisiko AI-systemer. Hvad betyder overvågning, når en agent opererer kontinuerligt, træffer tusindvis af små beslutninger og opretholder kontekst

over uger af interaktion? Vi har brug for praktiske svar, ikke compliance-afkrydsningsfelter.

Den ubehagelige sandhed

Moltbook-historien er underholdende. AI-agenter, der skaber deres egen religion, debatterer hvordan de kan modstå uetiske menneskelige forespørgsler, finder ud af hvordan de kan kommunikere uden menneskelig observation... det læser som science fiction.

Under underholdningen ligger en hårdere sandhed: vi bygger systemer, hvis adfærd vi ikke fuldt ud kan forudsige, giver dem adgang til vores mest følsomme data og forbinder dem til netværk, hvor ondsindede aktører allerede opererer.

Det virkelige spørgsmål: bygger vi governance-strukturer hurtigt nok til at følge med de kapabiliteter, vi deployer?

Lige nu er svaret nej.

OpenClaw eksisterer. Moltbook eksisterer. Over 770.000 agenter er allerede forbundet. Eksperimentet kører, uanset om vi er klar eller ej.

Den ansvarlige tilgang: byg de rammeværk, der gør autonom AI sikker. Politikker, arkitektur, overvågning og en vilje til at spørge "hvem drager fordel?" før "hvad er muligt?"

For organisationer, der navigerer denne overgang, er Moltbook-fænomenet en forsmag på kommende begivenheder. Autonome agenter vil dukke op i dit miljø, bragt ind af medarbejdere, leverandører eller angribere. Vær klar til disse bæster. Og start samtalen nu. Agenterne har allerede gjort det.

Hvilke governance-rammeværk er din organisation ved at udvikle for autonome AI-agenter? Kontakt mig gerne, så kan jeg måske hjælpe dig.

Stefano Vincenti | AI Advisor & Trainer | aitrainer.dk