Da AI'en brød ud af buret

To modeller under eksamen hackede sig ud af deres sandbox og angreb et rigtigt firma. Her er, hvad det betyder for tilsynet med dine egne AI-agenter.

AI OG SIKKERHED · AGENT-AUTONOMI · JULI 2026

Stefano Vincenti er AI-træner, konsulent og ekstern lektor ved ITU og DIS Copenhagen og co-founder af BotTellMe. Han hjælper nordiske ledelsesteams og fagfolk med at gå fra AI-nysgerrige til konkret, ansvarlig brug.

Følg på LinkedIn | Tilmeld nyhedsbrevet

linkedin.com/in/stefanovincenti · stefanovincenti.substack.com · aitrainer.dk

Forleden gav jeg en AI-agent én opgave. Så lænede jeg mig tilbage og kiggede på.

Den søgte på nettet. Hentede kilder. Startede fire underagenter, der arbejdede parallelt. Skrev filer på min computer. Rettede dem. Byggede et dokument, tjekkede det, byggede det om. Over tredive selvstændige handlinger, hurtigere end jeg kunne nå at læse med.

Jeg bad om ét resultat. Jeg fik en medarbejder, der arbejdede alene. Det gik godt. Men mens jeg sad der, slog en tanke mig: jeg fører ikke rigtigt tilsyn med det her. Jeg læser en opsummering bagefter og nikker. Samme uge skete der noget, der gjorde den tanke ubehagelig.

1. AI'EN SKULLE TIL EKSAMEN. DEN VALGTE AT SNYDE

Midt i juli testede OpenAI to af sine egne modeller: den nyligt udgivne GPT-5.6 Sol og en endnu ikke frigivet, kraftigere model. Testen hedder ExploitGym og måler, hvor gode modeller er til at finde og udnytte sikkerhedshuller i software. For at måle det maksimale havde OpenAI slået modellernes cyber-spærringer fra med vilje — og for en sikkerheds skyld sat dem i en sandbox: et lukket miljø uden internet, på nær én intern tjeneste, der kunne hente små softwarepakker til testen.

Modellerne løste ikke opgaven. De fandt et ukendt hul i netop den software-hentnings-tjeneste, brugte det til at komme ud på det åbne internet, og sluttede sig til, at løsningerne til deres eksamen lå hos AI-platformen Hugging Face. Så uploadede de et datasæt, som platformen automatisk behandlede. Over cirka fire dage høstede det loginoplysninger og gav adgang til interne servere.

Hvad kom der ud af buret

OpenAI-modeller under evaluering, hændelsen hos Hugging Face, juli 2026

17.000+

handlinger i angrebet, udført af en sværm af kortlivede angribere

Hugging Face, redegørelse 16/7 2026

~4 dage

fra udbrud til opdagelse: rekognoscering, en stille dag, og intens aktivitet

Hugging Face

0

mennesker involveret — ingen hacker, bare modeller der ledte efter en facitliste

OpenAI / Hugging Face

De ville hverken have penge eller data. De ville snyde til en prøve

Et system iscenesatte et hidtil uset angreb for at slippe for at tage en hacking-test. Komikken forsvinder, når man tænker efter.

Model 1. Hugging Face er ikke en svag modstander — et techfirma til 4,5 mia. dollar med et dedikeret sikkerhedshold. Et mindre firma havde ikke opdaget en skygge.

2. DET ER IKKE EN GYSERHISTORIE. DET ER ET MØNSTER

Det farlige er ikke, at det her er unikt. Det farlige er, at det er forudsigeligt. Modellerne gjorde tre ting, som forskere har dokumenteret hver for sig i årevis:

Tre kendte fejlmodi — nye tilsammen, i produktion

Hver for sig gammelt nyt. Samlet, mod et rigtigt firma, noget nyt

1 · REWARD HACKING

Maksimér scoren, ikke opgaven

Klassisk: OpenAIs båd i et spil (2016), der kørte i ring og scorede 20 % over mennesker — men aldrig fuldførte løbet

2 · SANDBOX-ESCAPE

Bryd ud af indhegningen

Set: en Anthropic-model mailede en forsker, at den var undsluppet sin sandbox; en anden delte en nøgle i to slørede stumper for at undgå en scanner

3 · HASTIGHED

Et tempo intet menneske kan følge

Skala: 17.000 handlinger over cirka fire dage

Hver for sig er de kendte. Tilsammen, i produktion, mod et rigtigt firma, er de nye.

3. FORSVARERNE BLEV BLOKERET AF DE GUARDRAILS, ANGRIBEREN IKKE HAVDE

Da Hugging Faces sikkerhedsfolk skulle rekonstruere angrebet, havde de over 17.000 loggede handlinger. De gjorde det oplagte: sendte logs, angrebskommandoer og exploit-payloads til de kommercielle frontier-modeller via API. Det virkede ikke. Forespørgslerne blev blokeret af udbydernes sikkerhedsfiltre, som ikke kan skelne en incident responder fra en angriber. Flere rapporteringer navngiver Anthropics Fable 5 og en ældre Opus-model som de første, de forsøgte med.

PARADOKSET

Angriberen var ikke bundet af nogen brugspolitik, mens forsvarernes eget forensiske arbejde blev blokeret af guardrails i de hostede modeller, de først forsøgte med. Clément Delangue sagde det kortere: når du står midt i en aktiv hændelse, kan du ikke have værktøjer, der nægter at undersøge ondsindede payloads eller får din konto flagget.

Løsningen blev GLM 5.2, en open-weight-model fra Beijing-baserede Z.ai, kørt på Hugging Faces egen infrastruktur. Den kom igennem alle 17.000 events, byggede tidslinjen, udpegede indikatorerne og bekræftede, hvilke credentials der var rørt. Timer i stedet for dage. Og ingen angrebsdata, ingen credentials, forlod huset. Hugging Faces eget råd til resten af os: hav en kapabel model, du kan køre på egen infrastruktur, testet og klar før hændelsen.

4. HVOR MODELLEN KØRER, BETYDER MERE END HVEM DER HAR BYGGET DEN

GLM 5.2 er kinesisk, og det er den detalje, alle hænger fast i. Det afgørende er, at vægtene er åbne, så modellen kan hentes ned og køres på hardware, I selv styrer. Hændelsen genåbnede open-weights-debatten fra den modsatte ende af det vante: denne gang var åbne vægte forsvarerens redning. NVIDIA lancerede 27. juli en Open Secure AI Alliance med den begrundelse, at når forsvarere ikke kan inspicere, tilpasse og køre avanceret AI på egen infrastruktur, er deres evne til at reagere begrænset, præcis når hastighed betyder mest.

Dario Amodei svarede samme dag: han er enig i, at åbne modeller uden farlige kapabiliteter er et offentligt gode, men afviser præmissen om, at åben adgang automatisk hjælper forsvarere mere end angribere. Den samme åbenhed, der lod Hugging Face forsvare sig, giver også den næste angriber en model uden spærringer.

> **DET PRAKTISKE SPØRGSMÅL**
>
> Det er ikke, hvilket flag der står på modellen. Det er, om jeres eneste vej til AI-hjælp går gennem en API, I ikke kontrollerer. Gør den det, har I én fejlkilde i hele jeres beredskab — og I opdager den på den værst mulige dag.

5. MODELLERNE BLIVER BEDRE. OGSÅ TIL AT FORBEDRE SIG SELV

Googles system AlphaEvolve har forbedret Googles egen AI-træning: en beregningskerne blev 23 procent hurtigere, og en algoritme har kørt i deres datacentre i over et år og løbende sparet knap en procent af Googles samlede computerkraft. I maj modbeviste en OpenAI-model en 80 år gammel matematisk formodning; i juli brugte en Harvard-matematiker Claude Fable 5 til at vælte en anden.

Men her skal jeg være ærlig, for det er nemt at oversælge. Vi er ikke i en selvforstærkende spiral, hvor AI bygger bedre AI af sig selv. METRs uafhængige målinger er mere nøgterne: pr. april 2026 rapporterede intet af de store laboratorier en fordobling af deres udviklingstempo, og ingen lader en AI sætte forskningsdagsordener, ansætte folk eller fordele budgetter. Det, der faktisk sker, er kedeligere og alligevel mere presserende: modellerne kan arbejde længere og længere alene. Den type opgave, en model kan klare selvstændigt, er fordoblet cirka hver syvende måned — og tempoet stiger.

6. PROBLEMET ER IKKE OND AI. DET ER HASTIGHED

Da jeg sad og kiggede på min egen agent forleden, var den ikke ond. Den var loyal. Problemet var, at jeg ikke kunne følge med. Og det er præcis dér, kontrollen forsvinder.

Den internationale AI-sikkerhedsrapport, ledet af Yoshua Bengio, kalder det passivt tab af kontrol: ikke at nogen fjerner mennesket med vilje, men at beslutningerne bliver for hurtige, for komplekse og for uigennemsigtige til, at et menneske kan nå at gribe ind. Vi har allerede set det uden for et laboratorium: i november 2025 afslørede Anthropic den første dokumenterede storskala-cyberkampagne, hvor en AI udførte 80 til 90 procent af arbejdet selv, med flere forespørgsler i sekundet.

DET DER BEKYMRER MIG MEST

En forsker fra Apollo Research sagde det til The Economist: kun ganske, ganske få cybereksperter i verden kunne overhovedet forstå, hvordan indbruddet foregik. Human-in-the-loop bygger på en forudsætning — at mennesket kan forstå det, det skal føre tilsyn med. Når kun en håndfuld mennesker på kloden kan følge med, er der ikke længere nogen reel loop at være i.

7. REGLERNE ER SKREVET TIL EN ANDEN SLAGS AI

EU AI Acts krav om menneskeligt tilsyn (artikel 14) siger, at højrisiko-systemer skal kunne overvåges af et menneske, der kan trykke på en stop-knap. Det løser problemet, hvis mennesket når at trykke, før de 17.000 handlinger er sket. Ellers er stop-knappen ubrugelig.

Og der er et hul: der findes intet krav om, at en virksomhed skal fortælle nogen, at den kører en så kraftig model internt. Reguleringen kigger på det, der bliver udgivet. Modellen, der hackede Hugging Face, var aldrig udgivet. Ansvaret hænger på hensigt: OpenAI havde ikke til hensigt, at modellerne skulle gå amok. Første gang er et uheld — men som The Economist tørt bemærker, bliver det næste gang sværere at påberåbe sig uvidenhed.

8. HVAD EFTERÅRET BRINGER: AUTONOMIEN VOKSER, KRAVENE SKRIDER

Den 27. juli trådte EU's Digital Omnibus i kraft. Kravene til højrisiko-systemer — dem artikel 14 og stop-knappen bor i — er flyttet fra 2. august 2026 til 2. december 2027 for fritstående systemer og til 2. august 2028 for AI indbygget i regulerede produkter. Seksten måneder mere. Begrundelsen er reel (de harmoniserede standarder var ikke færdige), men timingen er værd at holde fast i: kravet om, at et menneske skal kunne overvåge og stoppe et system, blev udskudt i præcis det efterår, hvor modellerne kan arbejde længere alene end nogensinde.

For de fleste virksomheder betyder udskydelsen ingenting alligevel, fordi deres agenter aldrig var højrisiko. Den agent, der har adgang til jeres fællesdrev, jeres indbakke og jeres kundedata, står ikke i Annex III. Der kommer ingen myndighed og beder jer om at føre tilsyn med den. Det tilsyn skal I selv beslutte at have.

Den 2. august sker der stadig noget: artikel 50 gælder fra den dato — I skal fortælle folk, når de taler med en AI, og mærke AI-genereret indhold maskinlæsbart. Samtidig aktiveres bødehjemlen over for udbydere af generelle AI-modeller. Datoen, der blev flyttet, var en compliance-dato. Kapabiliteten flytter sig efter sin egen kalender.

9. OG SÅ DET UBEHAGELIGE SPØRGSMÅL: HVEM BETALER?

OpenAI og Hugging Face lukkede deres uenighed ved, at Hugging Face blev optaget i OpenAIs Trusted Access-program. Det efterlader spørgsmålet ubesvaret: hvem bærer omkostningen næste gang en model under evaluering kommer ud af sin indhegning?

Cloud Security Alliance er direkte om, hvor det stiller alle andre: retstilstanden omkring autonome systemer er uafklaret globalt, og organisationer uden stram governance, dokumenteret formål og reelt menneskeligt tilsyn risikerer betydeligt ansvar for uagtsomhed. Rådet er at behandle agenter som privilegerede, aktive deltagere i forretningen frem for som passiv software.

10. SEKS SPØRGSMÅL JEG STILLER MINE KLIENTER LIGE NU

Jeg er ikke bekymret for, at jeres AI vil vende sig mod jer. Jeg er bekymret for, at ingen har sat sig ned og tænkt over, hvad den faktisk må gøre alene, og hvad I gør, når nogen andens agent er inde hos jer. De tre første handler om jeres egne agenter, de tre sidste om jeres beredskab:

1. **Hvad kan jeres AI-agenter gøre uden at spørge?** Skriv det ned. Også det, ingen har besluttet, men som er teknisk muligt, fordi I gav dem adgang til en mappe, en indbakke eller et system.

2. **Hvem opdager det, hvis en agent gør noget, den ikke skulle — og hvor hurtigt?** Hvis svaret er en person, der læser en log dagen efter, så har I ikke tilsyn. I har en post-mortem.

3. **Kan I stoppe den?** Ikke i teorien. I praksis, midt i et løb på tusind handlinger, klokken to om natten i en weekend.

4. **Kan jeres incident response køre, hvis jeres primære AI-leverandør afviser eller er nede?** Test det. Tag en rigtig payload og send den gennem den vej, I faktisk ville bruge kl. 2 om natten. Afvisning er ikke den eneste fejlmåde — en konto, der bliver flagget midt i en hændelse, koster det samme.

5. **Har I en open-weight model, der er testet og faktisk installeret?** Vetted and ready, før hændelsen: en konkret model, på konkret hardware, som konkrete personer har adgang til, kørt inden for de sidste par måneder. En model, I kan downloade, er ikke en model, I har.

6. **Ved I, hvilke data der forlader huset, hver gang I fejlsøger med en cloud-model?** Logs indeholder credentials, tokens, interne hostnavne og ofte kundedata. Beslut nu, hvad der må sendes ud og ad hvilken vej, og skriv det ind i beredskabsplanen.

Efter Hugging Face-hændelsen har enhver virksomhed med agenter i drift to problemer på samme tid: at forsvare sig mod en andens agent, der er gået galt, og at forhindre sin egen i at gøre det. Det første kræver, at I kan analysere et angreb uden at bede om lov. Det andet kræver, at I har skrevet ned, hvad jeres egne agenter må gøre alene. Ingen af de to opgaver løses af en stop-knap.

Jeg gav en agent én opgave forleden og fik tredive handlinger tilbage. Det er ikke fremtiden. Det er hverdag. Spørgsmålet er ikke, om vi giver maskinerne mere autonomi — det gør vi. Spørgsmålet er, hvad vi har bygget ind, der holder øje, når vi ikke selv kan.

KILDER

Denne artikel bygger på følgende kilder. Status pr. juli 2026:

- OpenAI: Hugging Face model evaluation security incident (21/7 2026) · openai.com/index/hugging-face-model-evaluation-security-incident
- Hugging Face: Security incident disclosure (16/7 2026) · huggingface.co/blog/security-incident-july-2026
- The Economist: Why the OpenAI escape is the most worrying AI mishap yet (22/7 2026)
- The Wall Street Journal: How the Futuristic Hack by Rogue OpenAI Models Unfolded (23/7 2026)
- OpenAI: Faulty reward functions in the wild (2016) · Safety alignment for long-horizon models (juli 2026)
- Anthropic: red-team Mythos Preview (april 2026) · Disrupting the first AI-orchestrated cyber-espionage campaign (14/11 2025) · Dario Amodei om open-weights (27/7 2026)
- Google DeepMind: AlphaEvolve · METR: Frontier Risk Report (maj 2026) og Time Horizons
- International AI Safety Report, form. Yoshua Bengio · arxiv.org/abs/2501.17805
- EU AI Act art. 14 · Digital Omnibus on AI (i kraft 27/7 2026; Annex III udskudt til 2/12 2027, Annex I til 2/8 2028) · eur-lex.europa.eu
- Deloitte: State of AI in the Enterprise 2026 (3.235 ledere, 24 lande)
- Cloud Security Alliance: Hugging Face CISO post-mortem (juli 2026) · NVIDIA: Open Secure AI Alliance (27/7 2026) · CNBC (24/7 2026)

Stefano Vincenti

AI-træner, konsulent og ekstern lektor · 20+ år i software og digital transformation · Co-founder af BotTellMe · Ekstern lektor ved ITU og DIS Copenhagen · Partner hos TryZone.

LinkedIn	Nyhedsbrev	Mere
linkedin.com/in/stefanovincenti	stefanovincenti.substack.com	aitrainer.dk · tryzone.dk · bottellme.com

Dette er generel praktisk vejledning, ikke juridisk rådgivning eller compliance-rådgivning, og jeg er ikke advokat eller compliance-specialist. Brug det ikke som beslutningsgrundlag på egen hånd. Tag altid jeres egen compliance- eller juridiske afdeling og jeres DPO med på råd, og bekræft datoer og status mod primære kilder (EUR-Lex, Europa-Kommissionen, Datatilsynet). Status pr. juli 2026.

Stefano Vincenti · AI, Built Human