AI-agenter og cyberrisiko: forbered beredskabet, før alarmen går

Angriberne har fået agenter. Hvem samler sporene hos jer?

GUIDE · OKTOBER 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

I juni 2017 var jeg på kontoret hos et stort dansk rederi, da computerne omkring mig begyndte at slukke. Én efter én. Jeg tog min egen computer af netværket og reddede den. For virksomheden var det for sent. Det tog lang tid og en kæmpe indsats fra IT-afdelingen, før skaderne var håndteret.

Angrebet var forklædt som ransomware, men var bygget til at ødelægge. Det kom ind via en kompromitteret softwareopdatering, og fordi hele netværket var koblet sammen, var der intet, der kunne stoppe det. Der var ingen AI involveret. Det spredte sig alligevel hurtigere, end nogen kunne nå at reagere, og jeg har aldrig glemt, hvordan det så ud.

Jeg tænkte på det igen, da DTU's underretning kom. Fredag den 2. oktober offentliggjorde DTU en underretning om et hackerangreb. Op mod 200.000 nuværende og tidligere brugere kan være berørt.

Uvedkommende havde kompromitteret DTU-profiler og brugt dem til at komme ind i DTUBasen, universitetets system til identitets- og rettighedsstyring. Derfra blev der hentet data: CPR-numre, navne, privatadresser og for de aktive brugere også navn og telefonnummer på pårørende. Oplysningerne går tilbage til 2003.

DTU har inddæmmet angrebet, hentet eksterne specialister ind og anmeldt bruddet til Datatilsynet. Universitetet skriver selv, at det ikke kan fastslå, hvor mange der faktisk er berørt.

Lad mig være præcis om, hvad vi ikke ved. Underretningen siger intet om, hvordan profilerne blev kompromitteret. Den nævner ikke AI. Jeg har intet grundlag for at sige, at sagen har noget med AI-agenter at gøre, og jeg vil ikke gætte.

Sagen stiller til gengæld det spørgsmål, som enhver sikkerhedsafdeling skal kunne svare på i dag:

> *Når alarmen går en fredag morgen, hvor hurtigt kan vi så selv samle sporene, forstå omfanget og træffe de rigtige indgreb?*

Det er blevet sværere at svare godt på i løbet af det sidste år. En del af angriberne har fået agenter til at gøre arbejdet. Forsvaret har nok adgang til de samme værktøjer, men i mange organisationer er de endnu ikke bygget ind i beredskabet.

Jeg underviser i at bygge og bruge AI-agenter, på ITU og ude i virksomhederne. Her samler jeg billedet for jer, der har ansvaret: hvad der er dokumenteret på angrebssiden, hvad forsvaret kan bruge agenter til, hvor agenten selv bliver en risiko, og hvad der skal være på plads før jeres egen fredag morgen.

1. ANGRIBERNE BRUGER AGENTER SOM ARBEJDSKRAFT

I november 2025 offentliggjorde Anthropic det, firmaet kalder den første dokumenterede sag om et stort cyberangreb udført uden væsentlig menneskelig indblanding. Aktøren, som Anthropic med høj sikkerhed vurderer som kinesisk statsstøttet, gik efter omkring 30 mål. AI udførte 80-90 procent af arbejdet. Mennesker greb ind ved fire til seks kritiske beslutningspunkter pr. kampagne. Det lykkedes i et lille antal tilfælde.

Det var dengang en enkelt sag. Ti måneder senere er det et mønster.

Anthropics trusselsrapport fra september 2026 beskriver operatører, der rutinemæssigt kører agent swarms: en ledende agent deler rekognoscering og arbejdet efter indbruddet op og sender opgaverne ud til mange underagenter, der arbejder parallelt.

Én aktør, sandsynligvis med base i Kina, gik efter omkring 50 organisationer, og ét af dens arbejdsforløb gav på en enkelt måned mere end et dusin mulige zero-day-fund. En gruppe tilknyttet ShinyHunters hentede over 2.100 sæt Azure AD-tokens fra mere end 40 virksomheders tenants på omkring 34 timer.

Googles Threat Intelligence Group (GTIG) rapporterede 8. september om et multiagent-framework, hvor angribere på under seks timer gik fra en kompromitteret cloudressource til en kørende kampagne, der høstede adgangsoplysninger i stor skala.

I maj beskrev GTIG desuden en zero-day-exploit, som de med høj sikkerhed vurderer er udviklet med hjælp fra en AI-model. Google afbrød kampagnen, før den blev brugt i stor skala.

Model 1. Udvalgte tal fra angrebssiden.

Og på den mest jordnære front har Microsoft opgjort klikraten for AI-genererede phishingmails til 54 procent mod omkring 12 procent for traditionelle.

> **PROPORTIONER**
>
> Ingen af kilderne viser, at alle angribere i dag bruger autonome agenter. Meget af det, der rammer danske virksomheder, er stadig stjålne adgangsoplysninger, upatchede systemer og phishing. Forskellen er, at agenterne gør de samme angreb hurtigere, billigere og mere skalerbare. Det er tempoet, der har flyttet sig.

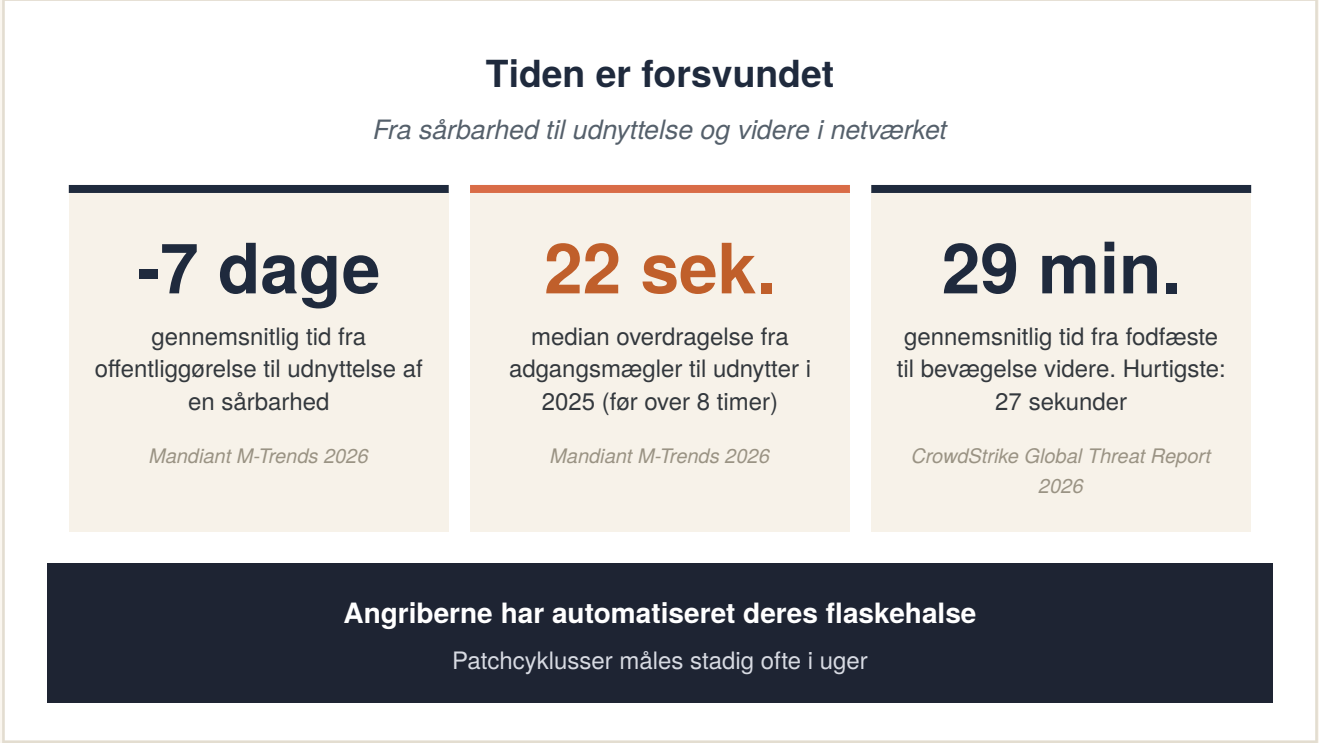
De her ting har jeg talt om, arbejdet med og tænkt længe over, da jeg var i Nexi Group og var en del af holdet bag MitID i halvandet år. Der blev der talt om truslerne mod den centrale digitale infrastruktur, som betjener over fem millioner brugere, hver dag, 24/7.

2. TIDEN MELLEM FEJL OG UDNYTTELSE ER FORSVUNDET

Tallene fra årets store trusselsrapporter fortæller historien bedre end nogen holdning.

Mandiants M-Trends 2026 opgør den gennemsnitlige tid fra en sårbarhed offentliggøres, til den udnyttes, til cirka minus syv dage. Udnyttelsen sker altså rutinemæssigt, før der findes en rettelse. Overdragelsen fra den gruppe, der skaffer første adgang, til den gruppe, der udnytter den, tog for få år siden over otte timer i median. I 2025 tog den 22 sekunder.

CrowdStrikes Global Threat Report 2026 måler den gennemsnitlige tid, kriminelle grupper bruger fra første fodfæste til bevægelse videre i netværket, til 29 minutter. Den hurtigste tog 27 sekunder. Du har læst rigtigt: 27 sekunder.



Model 2. Tempoet på angrebssiden.

Læg de tal ved siden af jeres egen virkelighed. Patchcyklussen måles ofte i uger. Godkendelsen af et akut indgreb i produktion tager ofte timer. Den første manuelle tidslinje over en hændelse kan tage en hel arbejdsdag, fordi data skal trækkes fra fem systemer i tre forskellige tidszoner.

Det er dér, forskellen opstår. Angriberne har automatiseret deres flaskehalse. I 2017 var det en orm uden AI. I dag har angriberne agenter, der kan planlægge, tilpasse sig og arbejde i timevis uden pause.

3. SÅRBARHEDER FINDES NU HURTIGERE, END NOGEN KAN LUKKE DEM

I april lancerede Anthropic Project Glasswing: udvalgte partnere, blandt andre AWS, Apple, Microsoft, Google og CrowdStrike, fik adgang til en ikke-offentlig model, Claude Mythos Preview, for at finde sårbarheder i kritisk software, før angriberne gør det.

Statusopdateringen i maj er den mest interessante læsning. Partnerne havde samlet fundet over 10.000 sårbarheder af høj eller kritisk alvorlighed i deres egen software. I open source-delen var 90,6 procent af de fund, uafhængige sikkerhedsfirmaer havde vurderet, reelle. Men af de cirka 530 alvorlige fejl, der var rapporteret til vedligeholderne, var kun 75 rettet.

Model 3. At finde er blevet let. At rette er det ikke.

Anthropic skriver selv, at fremskridt i softwaresikkerhed før var begrænset af, hvor hurtigt man kunne finde sårbarheder, og at begrænsningen nu er, hvor hurtigt man kan verificere, offentliggøre og rette dem.

MODARGUMENT

VulnCheck har set på 1.061 sårbarheder, der er fundet med AI-hjælp. Kun 14 af dem, 1,3 procent, er bekræftet udnyttet. VulnCheck vurderer, at effekten indtil videre er reel, men beskeden, og at AI først og fremmest øger mængden af fund, hvilket giver forsvarerne en chance for at lukke dem først.

For en CIO er pointen den samme, uanset hvilken læsning man hælder til. Flaskehalsen er flyttet fra at finde svaghederne til at sortere, verificere og rette dem i et tempo, der matcher. Den opgave lander hos jeres eget team.

4. I DANMARK ER TRUSSELSNIVEAUET HÆVET, OG LEDELSEN HÆFTER

Den 24. september hævede Styrelsen for Samfundssikkerhed, hvor Center for Cybersikkerhed nu hører til, trusselsniveauet for destruktive cyberangreb fra middel til høj. Begrundelsen er Rusland, som ifølge styrelsen fører hybridkrig mod Vesten og udviser øget risikovillighed. Styrelsen henvender sig især til virksomheder og myndigheder med ansvar for kritisk infrastruktur. Truslen fra cyberkriminalitet og cyberspionage stod allerede på meget høj i den samlede vurdering fra november 2025.

Samtidig har NIS 2-loven været i kraft siden 1. juli 2025, mens tele, energi og finans har deres egne regler. Loven gør cybersikkerhed til et ledelsesansvar. Ledelsen skal godkende foranstaltningerne og føre tilsyn

med dem. I yderste konsekvens kan en person på direktionsniveau i en privat virksomhed, der er væsentlig enhed efter loven, midlertidigt forbydes at udøve ledelsesfunktioner.

> **STATUS PR. OKTOBER 2026**
>
> Trusselsniveauet for destruktive cyberangreb er høj (hævet 24. september 2026). NIS 2-loven (lov nr. 434 af 6. maj 2025) har været i kraft siden 1. juli 2025. Dette er ikke juridisk rådgivning: bekræft altid status mod retsinformation.dk og Styrelsen for Samfundssikkerhed, og inddrag jeres egen juridiske afdeling.

Det gør beredskabet til et emne for direktion og bestyrelse. Når en bestyrelse spørger, hvordan I vil håndtere en hændelse, skal svaret kunne vise, at I har øvet det.

5. FORSVARET HAR OGSÅ FÅET AGENTER

De mest solide tal, jeg har fundet, kommer fra et randomiseret forsøg med Microsofts Phishing Triage Agent i Security Copilot. Sikkerhedsanalytikere med agenten fandt op til 6,5 gange så mange reelle trusler pr. arbejdsminut og ramte 77 procent bedre i deres vurderinger end kontrolgruppen.

Model 4. Studiet er udført af Microsoft på eget produkt og gælder en afgrænset opgave.

Det interessante er, hvor gevinsten kom fra: agenten sorterede køen og forklarede sine vurderinger. Analytikerne brugte derefter 53 procent mere tid på de e-mails, der faktisk var ondsindede. Agenten tog altså støjen, og mennesket fik tid til det farlige. Det er det mønster, I skal designe efter. Det er mennesker, der arbejder sammen med agenter, som vinder her.

To forbehold: studiet er udført af Microsoft på Microsofts eget produkt, og det handler om phishing-triage, en afgrænset opgave med et klart facit. Det siger mindre om komplekse hændelser.

NØGLETAL

Security Copilot er siden november 2025 gradvist rullet ud som en del af Microsoft 365 E5 og E7. Der følger 400 Security Compute Units med pr. måned for hver 1.000 brugerlicenser, op til 10.000 SCU om måneden. De ruller ikke over til næste måned. Har I E5, betaler I måske allerede for en kapacitet, ingen i teamet bruger.

Google og CrowdStrike lover lignende gevinster med agentbaserede SOC-løsninger, men så vidt jeg har kunnet finde offentligt, uden publicerede uafhængige målinger bag. Behandl tallene som markedsføring, indtil I har målt selv.

Den tredje vej er lokale modeller med åbne vægte, hvis hændelsesdata ikke må forlade huset.

6. AGENTEN ER SELV EN NY ANGREBSFLADE

Den her del får sjældent meget plads i leverandørernes præsentationer.

Logs indeholder tekst, som angriberen selv bestemmer: user agents, URL'er, DNS-forespørgsler, brugernavne. Når en sprogmodel læser jeres logs, læser den også den tekst. Hvis angriberen skriver instruktioner ind i et felt, kan modellen finde på at følge dem.

Et forskningspreprint fra maj 2026 testede netop det. Opsummering af logs var den mest sårbare opgave: en bestemt type manipulation lykkedes i 96 procent af forsøgene uden forsvar. Med de stærkeste forsvar faldt den gennemsnitlige succesrate fra 26,6 til 11,8 procent. Lavere, men ikke nul. Studiet brugte én lille model og er ikke fagfællebedømt, men mekanismen er reel.

CrowdStrike rapporterer, at mere end 90 organisationer er blevet ramt af ondsindet prompt injection. GTIG beskriver malware, der via konfigurationsfiler får AI-kodeassistenter til at køre kommandoer, og trojaniserede MCP-servere, de små integrationer der giver agenter adgang til værktøjer, lagt ud på PyPI og i officielle GitHub-repositories.

OWASP udgav i december 2025 en top 10 over risici ved agentbaserede applikationer. Øverst står kapring af agentens mål, misbrug af dens værktøjer og misbrug af dens identitet og rettigheder.

Fire regler for en agent i beredskabet

Konsekvensen for sikkerhedsteamet er konkret

ADGANG

Læseadgang og intet andet

Hvorfor Agentens værktøjer og rettigheder er selv en angrebsflade.

INDHOLD

Logs og dokumenter er data, aldrig instruktioner

Hvorfor Angriberen bestemmer selv teksten i logfelter.

EVIDENS

Hver påstand peger på en konkret logpost

Hvorfor Så analytikeren kan efterprøve den mod kilden.

MANDAT

Indgreb kræver et menneske med mandat

Eksempler Deaktivere konti, isolere systemer.

Hver agent er en ny identitet: med ejer, mindste rettigheder og i jeres access reviews

Der er en sidste konsekvens, som især rammer jer, der ejer identitet og adgang. Hver agent, organisationen sætter i drift, er en ny identitet med adgange og privilegier. Den skal have en ejer og kun de rettigheder, opgaven kræver, og den skal med i jeres access reviews som enhver anden konto. Microsoft bygger det ind i platformen med Entra Agent ID og Agent 365. Jeg arbejder selv som admin på Agent 365 og ser det i praksis hver dag. En fornøjelse for en hands-on type som jeg er.

DTU-sagen handlede om et system til identitets- og rettighedsstyring. Det er den slags system, jeres agenter snart også skal stå i.

7. SÅDAN SER DE FØRSTE TIMER UD, NÅR AGENTEN ARBEJDER FOR JER

1. **Udtræk og filtrering.** Millioner af loglinjer skal ikke ind i en prompt. Søgning, filtrering og parsing laves med almindelige, deterministiske værktøjer: KQL, Python, grep. Modellen får de relevante uddrag. Tidszoner og hændelses-id'er bevares hele vejen, ellers falder tidslinjen fra hinanden.

2. **Tidslinjen.** Agenten sammenstiller hændelser fra identitet, endpoint og applikationer til én tidslinje, hvor hver linje henviser til sin kilde. Det er det arbejde, der manuelt ofte tager timer.

3. **Hypoteserne.** Agenten foreslår konkurrerende forklaringer på den første adgang og det videre forløb og skiller observationer fra antagelser. Manglende evidens bliver til konkrete spørgsmål: hvilke logs mangler vi, hvem skal vi spørge? Hvilke aktører kan være involveret? Hvor længe har det stået på?

4. **Efterprøvning.** Analytikeren tjekker de væsentlige påstande mod kilden. Det er her, fagligheden ligger, og det er dette trin, der ikke må springes over under tidspres.

5. **Inddæmning.** Agenten foreslår en prioriteret liste over indgreb med den forventede driftskonsekvens. Mennesket beslutter. I forsyning og produktion skal IT og OT holdes adskilt, og den driftsansvarlige skal med i vurderingen. Et kursus eller en agent giver ingen ret til at ændre i industrielle styresystemer under et angreb.

6. **Rettelse og forebyggelse.** En rettelse testes i et sandkassemiljø med en plan for tilbagerulning. Og så bruger I den samme type kapacitet, som Glasswing har vist, til at lede efter beslægtede svagheder andre steder i jeres egen kode og konfiguration.

Er hændelsen en zero-day i en identitetskomponent, starter arbejdet et andet sted. Agenten sorterer leverandørens advisory, kortlægger hvilke identiteter og privilegerede konti der er eksponeret, og skriver udkast til kommunikation og runbook. Mennesker beslutter, hvad der lukkes, og i hvilken rækkefølge.

> **SCENARIE**
>
> Forestil jer, at agenten i trin to ser en række logins kl. 03.12 på en servicekonto fra et ukendt IP-område. Den konkluderer, at det er planlagt vedligehold, fordi et felt i loggen siger det. Feltet er skrevet af angriberen.

Hvad fanger fejlen? To kontroller. Reglen om, at tekst i logfelter aldrig er en instruktion eller en forklaring i sig selv. Og kravet om, at enhver konklusion om vedligehold skal kunne matches mod en registreret ændring i jeres ændringslog. Ingen ændring, ingen frikendelse.

> *Ingen ændring, ingen frikendelse.*

Det er derfor, jeg mener, at et godt øvelsesforløb skal vise både et nyttigt resultat og en fejl, som deltagerne selv opdager. Kontrolarbejdet hører med til faget, på lige fod med selve analysen.

8. SEKS BESLUTNINGER, DER SKAL VÆRE TRUFFET, FØR ALARMEN GÅR

Agenter i beredskabet kan ikke indføres midt i en hændelse. Det meste af arbejdet ligger før.

1. **Hvilken model må se jeres logs.** Der er tre realistiske veje: en lokal model med åbne vægte, en godkendt EU-hostet model eller værktøjerne i jeres eksisterende sikkerhedsplatform. Åbne vægte og open source er ikke det samme, og hverken lokal drift eller EU-hosting afgør alene sikkerheden. Vurdér kvalitet, dataflow, underleverandører og licens. Og skeln mellem almindelig Microsoft 365 Copilot og Security Copilot.

2. **Hvad agenten kan nå, og med hvilke rettigheder.** Læseadgang til afgrænsede kilder, sporbar brug af værktøjer og et fast outputformat. Behandl agentens adgangsnøgler med samme alvor som adgangsoplysninger til produktion. Anthropic anbefaler det samme for AI-API-nøgler i septemberrapporten.

3. **Hvem må beslutte indgreb.** Hvem må deaktivere konti, isolere servere, lukke en integration? Skriv det ned, før agenten foreslår det.

4. **Et øvelsesmiljø med facit.** Syntetiske logs med støj, falske spor og manglende data. Uden et facit kan I ikke vide, om agenten hjælper eller bare lyder overbevisende, fordi den mangler kontekst.

5. **Kan miljøet genskabes.** Versionslåste afhængigheder, øvelsesdata, konfigurationseksempler og en vejledning, som en kollega, der ikke byggede det, kan følge. Mit forslag er, at det skal kunne genskabes inden for et døgn på en aftalt platform. Ellers står værktøjerne ikke klar, den dag de skal bruges. Det er ikke rart, at befolkningen eller brugerne skal stå uden en kritisk service i lang tid.

6. **Hvad I måler.** Tid fra alarm til et efterprøvet beslutningsgrundlag, kvaliteten af analysen og antallet af fejl, teamet fanger. Uden målinger bliver næste budgetsamtale en holdningsdiskussion.

9. TRE SPØRGSMÅL TIL JERES NÆSTE LEDERMØDE

```
1. Hvor lang tid tog det os sidst at lave en tidslinje over en sikkerhedshændelse, og hvem lavede den?

2. Hvilke AI-værktøjer betaler vi allerede for på sikkerhedsområdet, og hvem i teamet bruger dem i dag?

3. Hvis en agent i morgen foreslår at deaktivere 40 konti, hvem siger så ja, og på hvilket grundlag?
```

Kan I svare på alle tre, har I et udgangspunkt. Kan I ikke, ved I, hvad den næste øvelse skal handle om.

10. DET ARBEJDER JEG MED

Jeg hjælper virksomheder med en tilpasset master class i AI-agenter til cyberberedskab. Teamet arbejder hands-on med cases fra egen hverdag på syntetiske data, bygger og tester selv deres agenter og øver de første timer af en hændelse med en agent som assistent. Fokus tilpasses teamet: hændelsesanalyse i SOC'en, identitetssikkerhed eller governance af de agenter, I selv sætter i drift. Forløbet kan tages som en halv dag, en dag eller to dage. Skriv til mig, hvis I vil høre, hvordan det kunne se ud hos jer.

KILDER

Status pr. oktober 2026. Tal fra trusselsrapporter er gengivet, som leverandørerne selv har opgjort dem.

- DTU: dtu.dk/newsarchive/2026/10 (underretning om brud på persondatasikkerheden, 2. okt. 2026)
- Anthropic: anthropic.com/news/disrupting-AI-espionage (13. nov. 2025)
- Anthropic: anthropic.com/threat-intelligence-report-september-2026
- Google GTIG: cloud.google.com/blog/topics/threat-intelligence (11. maj 2026 og 8. sep. 2026)
- Microsoft Digital Defense Report 2025 (okt. 2025)
- Mandiant M-Trends 2026: security.googlecloudcommunity.com (23. mar. 2026)
- CrowdStrike 2026 Global Threat Report: crowdstrike.com/en-us/press-releases
- VulnCheck: vulncheck.com/blog/state-of-exploitation-1h-2026 (28. jul. 2026)
- Anthropic Project Glasswing: anthropic.com/glasswing og anthropic.com/research/glasswing-initial-update (22. maj 2026)

- Styrelsen for Samfundssikkerhed: samsik.dk/artikler/2026/09 (24. sep. 2026) og Cybertruslen mod Danmark 2025
- NIS 2-loven: lov nr. 434 af 6. maj 2025 (retsinformation.dk) og SAMSIK-vejledning om ledelsens rolle
- Bono, Randomized Controlled Trials for Phishing Triage Agent: arxiv.org/abs/2511.13860
- Microsoft Learn: learn.microsoft.com/en-us/copilot/security/security-copilot-inclusion
- Microsoft Learn: learn.microsoft.com/en-us/entra/id-governance/agent-id-governance-overview
- Pandey og Bhujang, Poisoning the Watchtower (preprint): arxiv.org/abs/2605.24421
- OWASP Top 10 for Agentic Applications 2026: genai.owasp.org

© 2026 Stefano Vincenti · aitrainer.dk · Oprindeligt publiceret på Substack 10. oktober 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

Disclaimer: 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 (retsinformation.dk, Styrelsen for Samfundssikkerhed, Datatilsynet). Tal fra trusselsrapporter er gengivet, som leverandørerne selv har opgjort dem. Status pr. oktober 2026.

Stefano Vincenti · AI, Built Human