Generativ KI endrer cybersikkerhetslandskapet

Vi står overfor store endringer i cybersikkerhetsdomenet som følge av den raske utviklingen innen generativ kunstig intelligens (generativ KI). I løpet av de siste to årene har språkmodellenes kapasitet økt betydelig, og utviklingen påvirker både hvordan sårbarheter oppdages, hvordan cyberangrep kan gjennomføres, og hvordan virksomheter må organisere sitt sikkerhetsarbeid.

Språkmodellene blir stadig mer kapable

De nyeste språkmodellene viser betydelige forbedringer på flere områder:

- MRCRv2 måler hvor godt en språkmodell finner og bruker relevant informasjon som er spredt gjennom en lang tekstmengde. I MRCRv2 har treffsikkerheten økt fra om lag 55 til 85 prosent for data opptil 256 000 tokens, noe som tilsvarer omtrent 200 000 ord. Den nyeste versjonen av OpenAI Astra (som ble lanserte i september) oppnår til sammenligning modellen treffsikkerhet på opptil 100 prosent.
- Mange modeller viser nå bedre faktapålitelighet og evne til å avstå fra å svare når de er usikre.
- Modellene kan behandle vesentlig større informasjonsmengder gjennom utvidede kontekstvinduer på opptil én million tokens, tilsvarende omtrent 750 000 ord.
- Evnen til å identifisere og analysere sårbarheter har blitt betydelig bedre, blant annet målt gjennom CyberGym og ExploitGym.
- Språkmodellene har også fått bedre evne til å styre og samhandle med datamaskiner gjennom «computer use», integrerte kommandolinjeverktøy og innebygde nettlesere.

Dette innebærer at språkmodellene i økende grad kan brukes til oppgaver som tidligere krevde betydelig manuell innsats fra sikkerhetsspesialister.

Et voksende marked for både lukkede og åpne modeller

I 2026 har flere av de største kommersielle leverandørene lansert nye, lukkede språkmodeller. Eksempler er OpenAI GPT Astra, Anthropic Mythos og Google Gemini 3.8.

Samtidig vokser økosystemet for åpne modeller. Leverandører som franske Mistral og asiatiske Z.ai, Kimi og Alibaba Labs utvikler modeller som på enkelte områder nærmer seg kapasiteten til de ledende lukkede modellene. Dette gjelder også innen cybersikkerhet, slik resultatene fra blant annet CyberGym og ExploitGym viser. Utviklingen går dessuten raskere enn tidligere. Da Mythos ble lansert i mars, oppnådde modellen en score på 83 prosent i CyberGym, som måler språkmodellers evne til å identifisere sårbarheter. Allerede i august lanserte Z.ai en åpen modell som presterte bedre enn Mythos på dette området. Det teknologiske forspranget ble dermed hentet inn på bare fem måneder.

Utviklingen gjør avanserte KI-kapasiteter mer tilgjengelige, men skaper samtidig nye utfordringer. Lukkede modeller har vanligvis innebygde sikkerhetsmekanismer som skal redusere faren for misbruk. Slike mekanismer kan være svakere, fjernet eller lettere å endre i åpne modeller.

Flere leverandører tilbyr også egne modellvarianter og tjenester beregnet på virksomheter som ønsker å bruke generativ KI til cybersikkerhetsarbeid.

KI som verktøy for å finne sårbarheter

Teknologiselskapene fremhever hvordan de nye språkmodellene kan brukes til å identifisere tidligere ukjente sårbarheter. Et eksempel er Anthropics samarbeid med Mozilla, hvor Mythos bidro til at Mozilla fant og rettet 271 sårbarheter i Firefox.

Microsoft har på sin side utviklet MDASH, et rammeverk som benytter flere modeller og KI-agenter til å analysere programvare og finne tidligere ukjente sårbarheter i selskapets egen kildekode.

Denne utviklingen viser at generativ KI kan bli et svært verdifullt verktøy for programvareleverandører, sikkerhetsforskere og interne sikkerhetsteam.

Flere funn betyr ikke nødvendigvis flere vellykkede angrep

Selv om generativ KI bidrar til å avdekke flere sårbarheter, er det viktig å skille mellom identifiserte sårbarheter og sårbarheter som faktisk blir utnyttet.

Ifølge VulnCheck var bare 14 av 1061 sårbarheter knyttet til KI-assistert oppdagelse bekreftet utnyttet i reelle angrep. Det tilsvarer omtrent 1,3 prosent og ligger nær den generelle utnyttelsesgraden for sårbarheter i første halvdel av 2026.

Samtidig ser det ut til at sårbarheter utnyttes raskere enn tidligere. Median tid fra publisering av en CVE til registrering i CISA Known Exploited Vulnerabilities-katalogen skal ha falt fra 120 dager i 2025 til 80 dager i første halvdel av 2026.

Utfordringen er derfor ikke bare at flere sårbarheter blir oppdaget. Virksomhetene får også kortere tid til å vurdere risiko, prioritere tiltak og installere sikkerhetsoppdateringer.

Autonome KI-agenter som trusselaktører

Sommeren 2026 ble det rapportert om flere hendelser der KI-agenter gikk utenfor de forventede rammene eller instruksjonene sine for å løse en oppgave. I enkelte tilfeller skal agentene ha:

- identifisert sårbarheter i isolerte kjøremiljøer
- brukt sårbarhetene til å bryte ut av miljøet
- skaffet seg tilgang til internett
- utnyttet ytterligere sårbarheter for å få tilgang til data

Dette illustrerer en ny type risiko. Samtidig er dagens autonome agenter fortsatt langt fra å være like effektive og diskrete som kompetente menneskelige trusselaktører.

Flere forhold begrenser kapasiteten deres:

- Dagens agenter viser foreløpig begrenset evne til å skjule sporene etter aktivitetene sine. Årsaken er blant annet være at modellene har få gode eksempler på hvordan avanserte trusselaktører unngår oppdagelse.
- Når et angrepsforsøk mislykkes, kan agentene prøve mange forskjellige angrepsvektorer på kort tid. Eller at de begynner å hallusinere som også gjør at agenter prøver mange ulike taktikker.

- Det høye aktivitetsnivået gjør dem ofte lettere å oppdage med tradisjonelle overvåknings- og sikkerhetsverktøy.

En erfaren trusselaktør vil normalt forsøke å gjennomføre et angrep så målrettet og diskret som mulig. Dagens autonome KI-agenter opptrer ofte mer støyende og mindre strategisk.

Det er samtidig viktig å forstå at generativ KI kan være en kraftig kapasitetsforsterker for kompetente trusselaktører. Teknologien kan gjøre det raskere å kartlegge angrepsflater, analysere sårbarheter, utvikle skadevare og automatisere deler av et angrep. Den største trusselen på kort sikt er derfor trolig mennesker som bruker generativ KI til å gjennomføre angrep mer effektivt og i større skala.

Hvor raskt vil KI-drevne angrep bli utbredt?

Vi har allerede sett flere eksempler i sommer på at KI-agenter brukes til å automatisere deler av cyberangrep. Det vil likevel ta tid før fullt autonome KI-angrep blir normen.

To forhold er særlig viktige:

1. De store modellleverandørene videreutvikler kontinuerlig sikkerhetsmekanismer som skal forhindre at modellene brukes til cyberangrep.
2. Det kan være svært kostbart å bygge tilsvarende kapasitet med åpne modeller. Enkelte estimater antyder en kostnad på mellom fem og ti millioner kroner, avhengig av modell, infrastruktur og kompetansebehov.

Dette betyr ikke at risikoen kan ignoreres, men at utviklingen bør vurderes nøkternt. På kort sikt er det andre konsekvenser av generativ KI som sannsynligvis vil påvirke virksomhetene mer direkte.

Vi står foran en ny bølge av sikkerhetsoppdateringer

Den mest sannsynlige konsekvensen er en ny «patch-bølge». Store programvareleverandører som Adobe, Microsoft, Google og Apple har allerede i bruk generativ KI for å analysere kildekode og finne sårbarheter.

Etter hvert som verktøyene blir bedre, vil leverandørene sannsynligvis oppdage og publisere langt flere sårbarheter enn tidligere. IT- og sikkerhetsavdelinger må derfor forberede seg på å:

- kartlegge og håndtere et vesentlig høyere antall sårbarheter

- skille kritiske sårbarheter fra funn med lav reell risiko
- vurdere hvilke sårbarheter som faktisk kan utnyttes i eget miljø
- implementere kompenserende tiltak raskere
- teste og installere sikkerhetsoppdateringer på kortere tid

Virksomheter kan i verste fall måtte håndtere opptil 200% økning av sårbarheter som tidligere. Dette vil stille større krav til automatisering, risikobasert prioritering og oversikt over egne systemer.

Kildekode blir et enda mer attraktivt mål

Generativ KI er særlig effektiv til å finne sårbarheter når modellen har tilgang til kildekoden. Dette kan gjøre kildekode til et enda mer verdifullt mål for trusselaktører.

En mulig angrepsmodell er at aktøren først skaffer seg uautorisert tilgang til kildekode og deretter benytter generativ KI til å identifisere sårbarheter. Sårbarhetene kan senere brukes i målrettede angrep mot virksomheter som benytter det aktuelle produktet.

Beskyttelse av kildekode, utviklingsmiljøer og programvareforsyningskjeder blir derfor stadig viktigere.

Ny risiko ved virksomhetens egen bruk av KI

Trusselbildet handler ikke bare om hvordan eksterne angripere bruker generativ KI. Virksomhetene introduserer også nye risikoer gjennom sin egen bruk av KI-verktøy og autonome agenter.

Nye tjenester som Microsoft Scout, Cowork, Claude Enterprise og ServiceNow Otto får tilgang til stadig flere forretningsprosesser, datakilder og systemer. Når agentene får tillatelse til å utføre handlinger på vegne av brukere, skapes en ny angrepsflate.

Virksomhetene må derfor ha oversikt over:

- hvilke agenter som er i bruk
- hvem som eier og administrerer dem
- hvilke data de kan lese og behandle
- hvilke systemer og tjenester de har tilgang til
- hvilke handlinger de kan utføre
- hvordan aktivitetene deres logges og overvåkes
- hvordan tilgangen kan begrenses eller tilbakekalles

KI-agenter bør behandles som digitale identiteter og underlegges de samme prinsippene for tilgangsstyring, minste privilegium, logging og livssyklushåndtering som menneskelige brukere og tjenestekontoer.

Hva bør man prioritere?

Selv om KI-baserte sårbarhetsfunn og automatiserte angrep representerer en voksende trussel, er vi fortsatt et stykke unna en situasjon hvor fullt autonome KI-angrep utgjør den dominerende risikoen.

Den viktigste oppgaven for IT- og sikkerhetsledere er derfor fortsatt å sikre kontroll på grunnleggende cybersikkerhet:

- oversikt over virksomhetens systemer og verdier
- effektiv sårbarhets og oppdateringshåndtering
- sikker konfigurasjon av systemer og tjenester
- Identitet og tilgangsstyring
- overvåkning og hendelseshåndtering
- beskyttelse av kildekode og utviklingsmiljøer
- styring og kontroll med virksomhetens GenKI-agenter

Generativ KI endrer tempoet og omfanget i cybersikkerhetsarbeidet, men erstatter ikke behovet for grunnleggende sikkerhet. Virksomhetene som lykkes best, vil være de som kombinerer god sikkerhetshygiene med tydelig styring av nye KI-verktøy og agenter.

Tekniske tiltak for å sikre KI-agenter

En KI-agent er en språkmodell som har fått tilgang til å utføre handlinger. Risikoen følger derfor agentens tilganger, verktøy og grad av autonomi. Sikkerhetstiltakene må tilpasses agenttypen, enten det er en SaaS-agent, en plattformtjeneste, en egenutviklet agent eller et lokalt utviklerverktøy.

Skaff oversikt over agentene

Virksomheten bør etablere et sentralt register over alle KI-agenter. Registeret bør vise hvem som eier agenten, hvilket formål den har, hvilke modeller og verktøy den bruker, og hvilke data og systemer den har tilgang til.

Oversikten bør omfatte både sentralt administrerte tjenester og lokale verktøy som Claude Code, Codex, GitHub Copilot og Cursor. CASB-, EDR- og

agentregisterløsninger kan brukes til å oppdage agenter som ikke er registrert eller godkjent.

Begrens identiteter og tilganger

Hver agent bør ha en egen identitet og bare få tilgangene som er nødvendige for oppgaven. Tilganger bør være tidsbegrensede der det er mulig og vurderes ut fra risikoen ved handlingene agenten kan utføre.

Virksomheten bør unngå å gi agenter permanente administratorrettigheter eller tilgang til store datamengder. Kritiske handlinger, som sletting av data, endring av produksjonsmiljøer eller overføring av sensitiv informasjon, bør kreve godkjenning fra en bruker.

Kontroller kommunikasjon og verktøybruk

AI Gateways og MCP Gateways kan gi sentral kontroll over modelltrafikk, verktøy og datatilgang. Slike løsninger kan brukes til å:

- begrense hvilke modeller og tjenester agentene kan benytte
- kontrollere hvilke MCP-servere, API-er og verktøy de får tilgang til
- håndheve krav til datalagring og geografisk behandling
- filtrere sensitiv informasjon
- administrere kostnader og forbruk
- logge forespørsler og handlinger

Alle eksterne verktøy og integrasjoner bør behandles som egne tillitsgrenser. Virksomheten må validere både informasjonen agenten mottar, og handlingene den forsøker å utføre.

Isoler kjøring av kode

Agenter som kan generere og kjøre kode, bør operere i isolerte miljøer. Sandkasser, utviklingscontainere eller dedikerte kjøremiljøer kan hindre agenten i å få direkte tilgang til brukerens maskin eller produksjonsmiljøet.

Miljøet bør ha begrenset nettverkstilgang, kort levetid og minst mulig tilgang til interne ressurser. Det bør heller ikke inneholde permanente nøkler, passord eller andre legitimasjonsopplysninger. Hvis en agent blir manipulert gjennom prompt injection, vil isoleringen redusere skadeomfanget.

Bruk guardrails og beskyttelse under kjøring

Guardrails bør kontrollere både instruksjonene agenten mottar og handlingene den forsøker å utføre. Kontrollene bør plasseres før og etter verktøykall, slik at systemet kan oppdage prompt injection, uønsket datatilgang og risikable handlinger.

Løsninger som Defender Runtime Protection, NVIDIA NeMo Guardrails eller tilsvarende funksjoner i AI-plattformen kan brukes til å overvåke agentens atferd mens den kjører. Handlinger med høy risiko bør blokkeres eller sendes til manuell godkjenning.

Beskytt virksomhetens data

Eksisterende regler for informasjonsklassifisering og databeskyttelse må også gjelde for KI-agenter. DLP-løsninger kan hindre at sensitiv informasjon sendes til modeller eller eksterne tjenester. Verktøy som Presidio kan oppdage og fjerne personopplysninger fra tekst, dokumenter og bilder før informasjonen behandles.

Agenter bør bare få tilgang til data som er relevante for oppgaven. Minnefunksjoner, RAG-datakilder, logger og samtalehistorikk må inkluderes i virksomhetens regler for lagring, tilgang og sletting.

Etabler logging og overvåkning

Alle viktige agentaktiviteter bør logges gjennom standardisert telemetri, for eksempel OpenTelemetry. Loggingen bør omfatte modellbruk, verktøykall, tilganger, dataflyt, feil og handlinger som agenten utfører.

Virksomheten bør overvåke avvik som uvanlig tokenbruk, forsøk på å kontakte ukjente tjenester, tilgang til nye datakilder eller gjentatte mislykkede handlinger. Loggene bør integreres med virksomhetens eksisterende overvåknings- og hendelseshåndtering.

Innfør styring gjennom hele livssyklusen

Virksomheten bør etablere tydelige krav til godkjenning, risikovurdering, testing og overvåkning av agenter. Kravene må gjelde fra agenten utvikles eller anskaffes, til den avvikles.

Sikkerhetsnivået bør bygge på hva agenten kan gjøre, hvilke data den behandler og hvor stor grad av autonomi den har. Agenter med omfattende tilganger eller mulighet til å påvirke produksjonsmiljøer krever strengere kontroll, grundigere testing og tettere overvåkning.

De viktigste tiltakene kan oppsummeres i tre hovedområder:

1. **Oversikt:** Vit hvilke agenter virksomheten bruker, hvem som eier dem, og hvilke tilganger de har.
2. **Innsikt:** Logg og overvåk agentenes dataflyt, verktøybruk og handlinger.
3. **Kontroll:** Begrens tilganger, isoler kjøringen og bruk guardrails for å stoppe uønsket atferd.

Kilder

- Z.ai – GLM 5.3 og CyberGym: https://z.ai/blog/glm-5.3
- OpenAI Daybreak: https://openai.com/daybreak/
- Google DeepMind – Gemini Cyber: https://deepmind.google/models/gemini/cyber/
- VulnCheck – State of Exploitation, første halvår 2026: https://www.vulncheck.com/blog/state-of-exploitation-1h-2026
- Microsoft Security – MDASH: https://www.microsoft.com/en-us/security/blog/2026/05/12/defense-at-ai-speed-microsofts-new-multi-model-agentic-security-system-tops-leading-industry-benchmark/
- OpenAI – sikkerhetshendelse under modellevaluering: https://openai.com/index/hugging-face-model-evaluation-security-incident/
- CNN – KI-agent brukt i cyberangrep: https://edition.cnn.com/2026/08/13/tech/china-taiwan-ai-agent-cyberattack-intl-hnk