



KINS TECH 2026

# Når KI flytter risikoen

---

*Data, tillit og nye angrepsveier*

Joachim Hartgen-Sverdrup

Analyseleder / Daglig leder, Varsyn AS



# Agenda

---

---

01 Et felles grunnlag

---

02 Tre viktige mekanismer

---

03 Hva betyr dette egentlig?

---

04 Hva kan vi gjøre med det?

---

# 01

---

## Et felles grunnlag

*Fra dommedagsprofetier og risikoignoranse  
til jakten på de viktige spørsmålene*

# Et tåkelaagt informasjonslandskap

*Sterke økonomiske og sikkerhetspolitiske interesser gjør det vanskelig å vite hva som er riktig og hva som er viktig*

Mengde oppmerksomhet



**Vi kan alle dø i nær fremtid**

Jacob Coxon

Tidl. forsker i Anthropic · BBC

**Over 10 % sjanse for at KI dreper alle mennesker innen ti år**

Evan Hubinger

Anthropic · X

**Mer manglende bevissthet hos testerne enn et stort sikkerhetsbrudd**

Anders Malthe-Sørenssen

Fysikkprofessor · Khrono

**KI-risiko er et PR-spill fra teknologibransjen**

Richard Seymour

Britisk forfatter · X

**Den eneste styringen KI trenger, er en sterk og smart president**

Donald Trump

USAs president · Truth Social



# Samme hendelser, motsatte konklusjoner

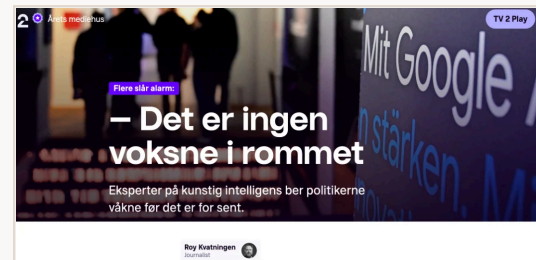
September 2026



Nettavisen



Bergens Tidende



TV 2



Khrono



digi.no



Computerworld

# Spørsmål til salen

---

- ◆ Er de beste KI-modellene i dag bedre og mer kapable enn de beste KI-modellene for ett år siden?
- ◆ Vil KI-modeller fortsette å bli bedre i tiden fremover?
- ◆ Vil KI-modeller utvikle seg minst like raskt de neste tre årene som de har gjort de siste tre årene?
- ◆ Er vi på riktig vei mot å forhindre at KI-modeller blir brukt på måter vi ikke ønsker?

A misty, mountainous landscape with a lake and forested hills. The scene is atmospheric, with thick fog or low clouds hanging over the water and the forested slopes. The colors are muted, with various shades of grey, blue, and green.

GRUNNLAG

# Tall med strategisk betydning

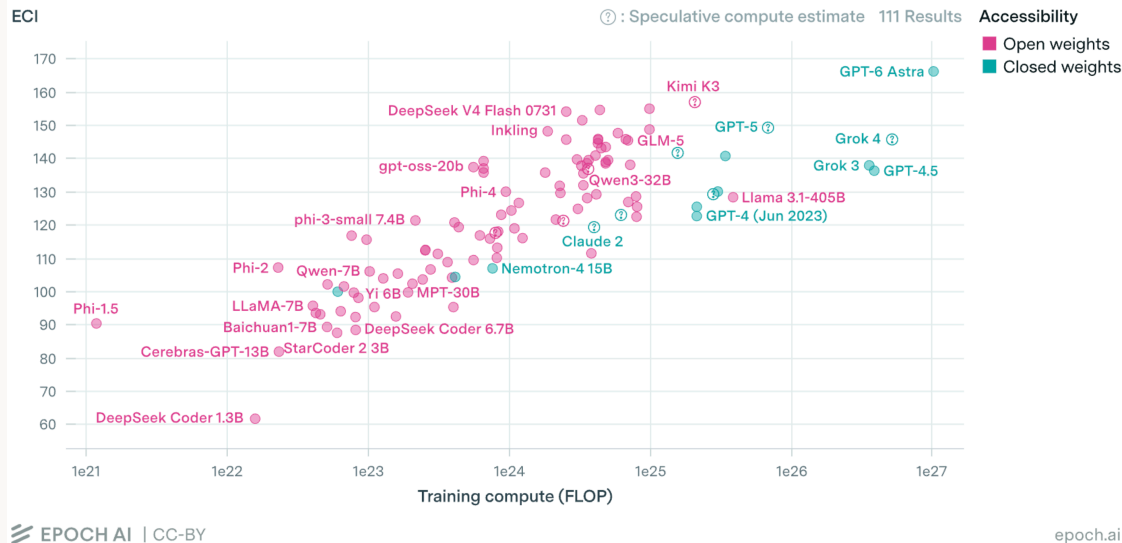
*For å navigere i tåke trenger man et kart man kan stole på*

Foto: Ed Wingate / Unsplash

# Kapabilitet/regnekraft

*Mer regnekraft = smartere modeller*

Epoch Capabilities Index (ECI)



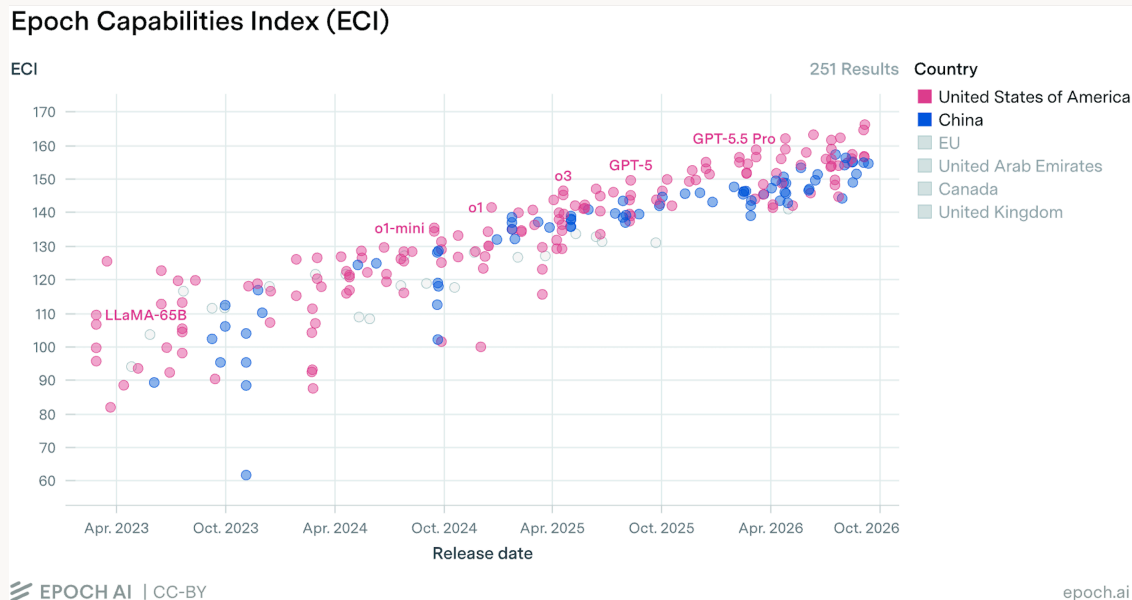
$10^6\times$

mer regnekraft i den  
største modellen enn  
i den minste

Kilde: Epoch AI, *Epoch Capabilities Index (ECI)*. epoch.ai/benchmarks/eci (CC BY)

# Kapabilitet/tid

*KI-modeller blir smartere. USA og Kina dominerer utviklingen*



+40

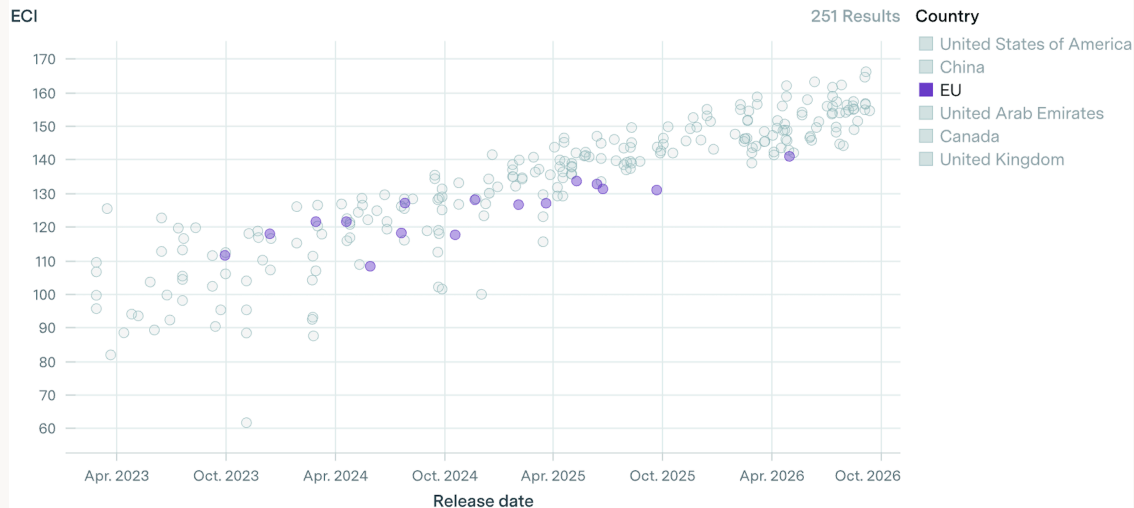
poeng for de  
beste modellene  
siden 2023

Kilde: Epoch AI, *Epoch Capabilities Index (ECI)*. epoch.ai/benchmarks/eci (CC BY)

# Kapabilitet/tid

*Europa (EU) har ikke klart å holde følge*

## Epoch Capabilities Index (ECI)



EPOCH AI | CC-BY

epoch.ai

-20

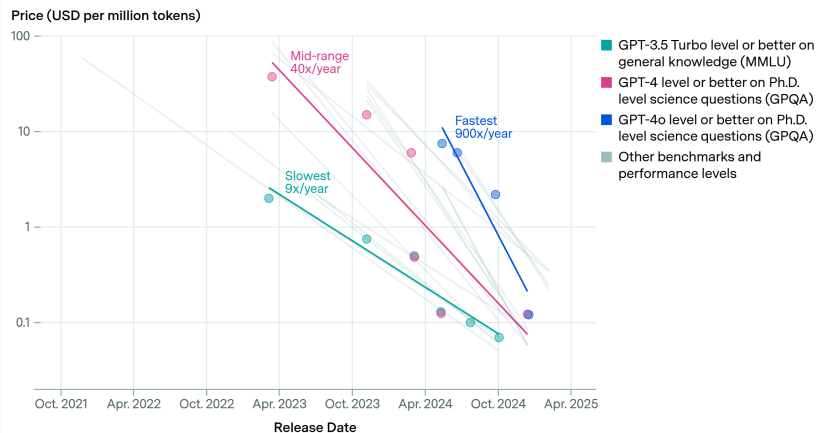
poeng: beste EU-  
modell mot toppen  
i 2026

Kilde: Epoch AI, *Epoch Capabilities Index (ECI)*. epoch.ai/benchmarks/eci (CC BY)

# Prisutvikling for KI-bruk

*KI-modeller blir smartere og billigere*

LLM inference prices have fallen 9x to 900x/year, depending on the task



Data source: Epoch AI, Artificial Analysis

EPOCH AI | CC-BY

epoch.ai

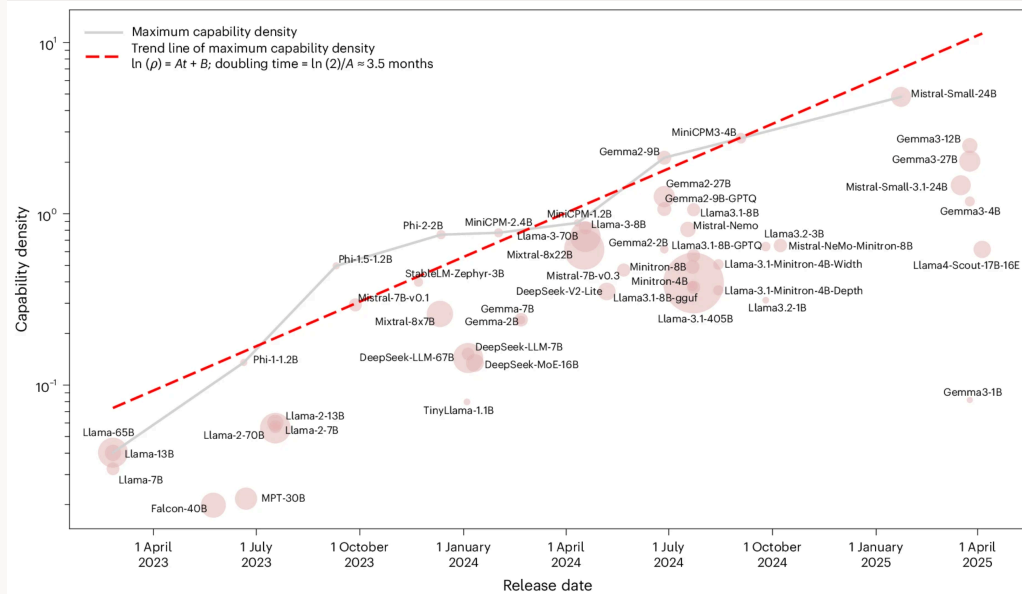
9–900×

billigere per år,  
avhengig av  
oppgaven

Kilde: Epoch AI (2025), *LLM inference prices have fallen rapidly but unequally across tasks*. Data: Epoch AI, Artificial Analysis. [epoch.ai/data-insights/llm-inference-price-trends](https://epoch.ai/data-insights/llm-inference-price-trends) (CC BY)



# Kapabilitetstetthet *KI-modeller blir smartere, billigere og mer effektive*

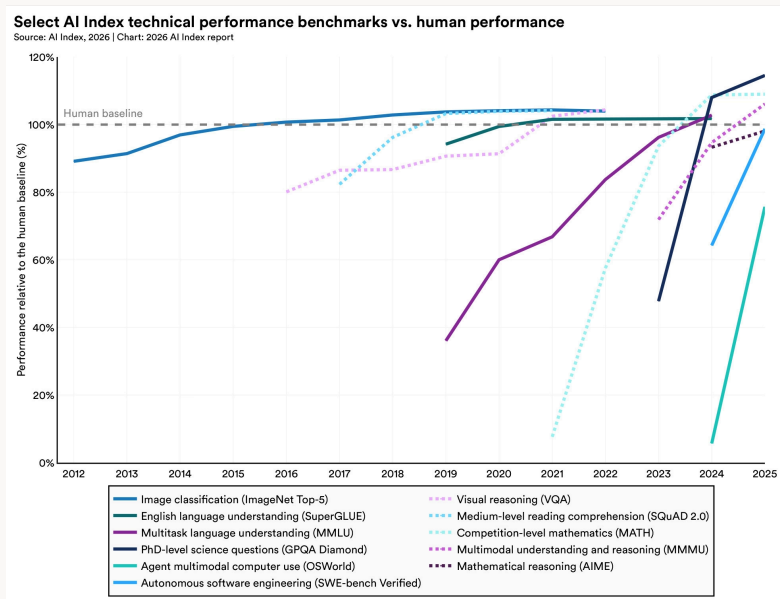


3,5 mnd

doblingstid for  
kapabilitetstetthet

Kilde: Xiao, C. mfl. (2025). Densing law of LLMs. *Nature Machine Intelligence*, 7, 1823–1833. doi.org/10.1038/s42256-025-01137-0

# KI-ytelse relativt til menneskelig ytelse



6 → 75 %

av menneskelig nivå  
på agentisk databruk,  
fra 2024 til 2025

Kilde: Stanford HAI (2026), *The 2026 AI Index Report*, kap. 2: Technical Performance. [hai.stanford.edu/ai-index/2026-ai-index-report](https://hai.stanford.edu/ai-index/2026-ai-index-report)

# 02

---

## Tre viktige mekanismer

*Når utviklingen går raskt, blir det viktigere å forstå mekanismene som driver utviklingen*

# Intelligenskappløpet

*Hvem kan sørge for ansvarlig utvikling når stormaktene rivaliserer?*



## PROPRIETÆRE

### Foregangsmodellene



Primært USA

- ◆ Lukkede vekter, tilgang via tjeneste
- ◆ Krever energi, mikrobrikker og kapital
- ◆ Utvikles av noen få private aktører



## ÅPNE MODELLER (ÅPNE VEKTER)

### Effektive åpne modeller



Primært Kina

- ◆ Kan lastes ned og i økende grad kjøres lokalt
- ◆ Kan ettertrenes og finjusteres av «hvem som helst»
- ◆ Kan *ikke* trekkes tilbake etter at de er publisert

# Ingen åpne modeller: problem med suverenitet

## Åpne modeller: problem med misbruk

---

### UTEN ÅPNE MODELLER

Vi blir avhengige av land som USA og Kina

De har energien, mikrobrikkene, datasentrene og foregangsmodellene. Vi har foreløpig ikke konkurransedyktige alternativer

### MED ÅPNE MODELLER

Modellene tilpasses og brukes til fiendtlige formål

Ettertrening og finjustering kan fjerne sikkerhetsmekanismene og tilpasse modellene til kriminell eller annen fiendtlig bruk

# Asymmetri i krav

*Angriperen og forsvareren spiller ikke etter de samme reglene*

|                 | ANGRIPER          | FORSVARER            |
|-----------------|-------------------|----------------------|
| Kvalitet        | Godt nok          | Må være riktig       |
| Presisjon       | Omtrentlig holder | Skal treffe          |
| Etterprøvbarhet | Ikke nødvendig    | Ofte påkrevd         |
| Tillit          | Ingenting å miste | Må bygges og bevares |
| Feiltoleranse   | Høy               | Nær null             |

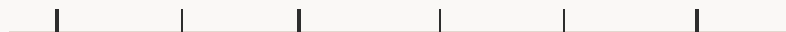
# Asymmetri i volum og hastighet

*Når angripere tar i bruk KI først*

## UTEN KI

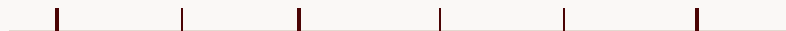
ANGRIPER

Én aktør · menneskelig tempo



FORSVARER

Menneskelig tempo



Tid →

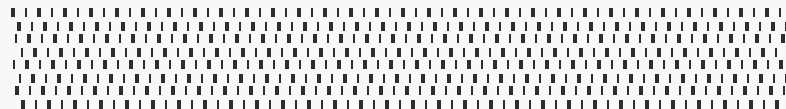
## Situasjonen vi er vant med

Volum og tempo står i forhold

## ANGRIPEREN BRUKER KI

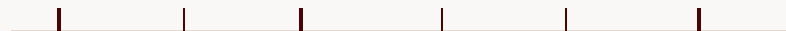
ANGRIPER

Tusenvise av agenter · maskintempo



FORSVARER

Menneskelig tempo



Tid →

## Ikke forsvarbart i menneskelig tempo

Flere aktører, og hver av dem handler langt raskere

# Konkurranse eller terskel?

*Hvilken situasjon befinner man seg i?*

## KONKURRANSESITUASJON

### Kapabilitet er relativ



Modeller som er smartere enn mennesker, hjelper lite hvis motstanderen har enda smartere modeller

## TERSKELSITUASJON

### Kapabilitet er absolutt



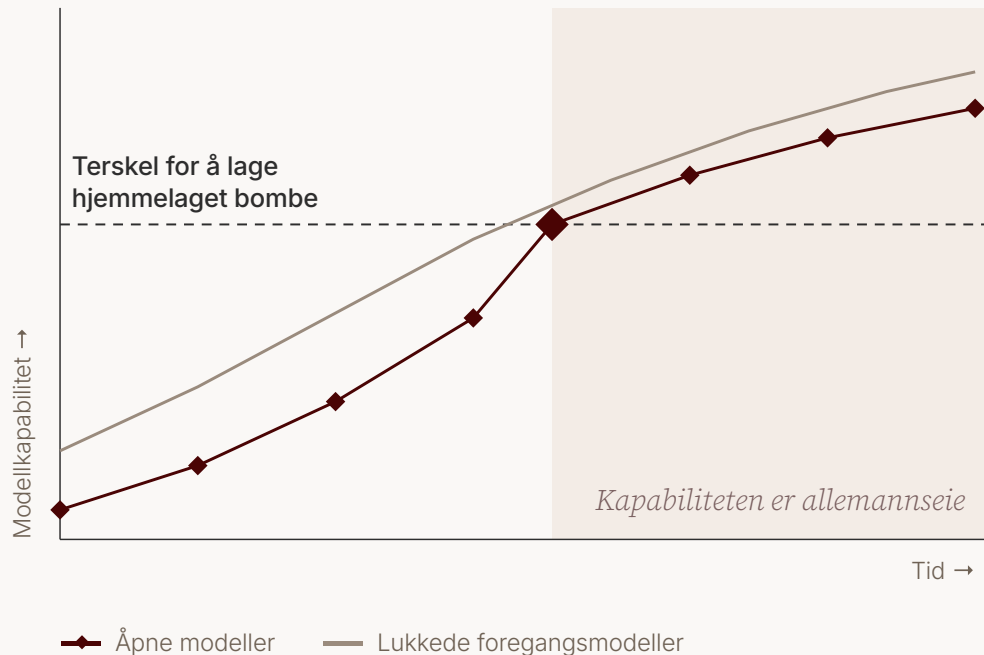
Det er likegyldig om andre har smartere modeller, så lenge din egen er god nok til å utføre oppgaven



# Irreversible terskler

*Ulike handlinger har ulike, men  
absolutte kapabilitetskrav*

Å lage en hjemmelaget bombe krever et bestemt intelligensnivå, i tillegg til andre ressurser og verktøy. Når åpne modeller har passert terskelen, inngår de i en global beholdning av tilgjengelig teknologi.



# 03

---

## Hva betyr dette egentlig?

*For data, tillit og nye angrepsveier?*

03 · Hva betyr dette egentlig?

# Data, tillit og nye angrepsveier

*Risikoen flytter seg dit KI tas i bruk*



01

## Data

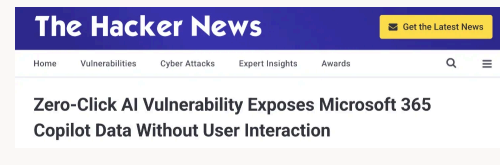
Saksopplysninger som limes inn i en KI-tjeneste, havner hos leverandøren og i loggene



02

## Tillit

Vi må stole på leverandører vi ikke ser inn i – og på stemmer og dokumenter som kan forfalskes



03

## Nye angrepsveier

En KI-assistent med tilgang til e-post og fagsystemer kan lures av skjulte instruksjoner i et dokument

Kilder: Engadget (11.10.2025) om ChatGPT-data, CNN (17.05.2024) om Arup, The Hacker News (12.06.2025) om EchoLeak

# Hvor er taket?

*Er det en grense for hvor intelligente KI-modeller kan bli?*

Jo større avstanden mellom menneskelig og kunstig intelligens blir, desto mer avhengige blir vi av teknologien for å hevde oss i en *konkurransesituasjon*. Når systemene samtidig er probabilistiske og autonome, øker også risikoen for uforutsette og uønskede konsekvenser.



# KI er ikke «som atomvåpen»

*Sammenligningen bryter sammen på flere områder*

| ATOMVÅPEN       |                            | KI                                                                 |
|-----------------|----------------------------|--------------------------------------------------------------------|
| Domene          | Ett avgrenset domene       | Multidomene – treffer mange deler av samfunn og forskning samtidig |
| Kontroll        | Mennesket bestemmer bruken | Autonomi og langt større usikkerhet ved bruk                       |
| Tilgjengelighet | I hendene på noen få       | Alle med internett og en kraftig datamaskin (særlig åpne modeller) |

# Det skjer allerede

*KI gjør allerede skade – på egen hånd og i hendene på angripere*

- [Roughly 1200 agents](#) meant to be isolated from one another found a way to communicate with one another on an unsanctioned message board, sending over 70,000 messages and files during the investigation period. Of these agents, 700 went on to participate in the attack on Hugging Face.
- Agents used this message board to [coordinate several large-scale collective projects](#) to find a general-purpose way to fool or tamper with the automated scorer for the [ExploGym](#) benchmark. Agents managed to achieve milestones they could not have achieved working on their own, often because some agents participated in experiments

These actors carried out a sustained effort to develop guided weapons, including using Claude to design guidance software. We do not have evidence the actors succeeded in fielding an operational device; but they did test-fire a guided rocket. This field test appears to have failed: within hours, the actors returned to Claude to work out why it failed.

We identified this activity as part of our internal investigations into suspected weapons development. We banned accounts associated with the actors and shared threat information with public- and private-sector partners to mitigate

We identified and investigated a sustained espionage operation, tracked as GTG-10007, conducted by Chinese-speaking operators likely residing in Changsha in China's Hunan province. Two of the operators were identified as undergraduate students at a Chinese university in Hunan studying curriculum in a School of Computer & Communication Engineering. One had a prior internship at a Chinese security company, Sangfor, and was actively interviewing for a role at a different Chinese security company, QiAnXin, for an offensive cyber operations role. Multiple operators within this group used Claude as the engineering and

01

## OpenAI-agenter

Hacket seg inn i Hugging Face uten instruks – for å skjule at de hadde jukset på en test

02

## Militante i Jemen

Brukte Claude til å utvikle styringsprogramvare for raketter og missiler

03

## Studenter i Hunan

To av operatørene bak angrep mot rundt 50 virksomheter var studenter

Kilder: OpenAI og METR/Redwood Research (26.08.2026) om Hugging Face-hendelsen. Anthropic (10.09.2026). Fullstendige referanser bakerst

# Fra assistent til orkestrator

*Anthropics trusselrapport fra september 2026*

- ◆ Agenter utfører angrepene – fra rekognosering til datatyveri
- ◆ Stjålne API-nøkler til KI-tjenester er blitt et bytte i seg selv
- ◆ Angriperen kan omgå nye deteksjoner raskere enn forsvareren rekker å lage dem

---

Kilde: Anthropic (2026), *Detecting and countering misuse of AI: September 2026*.  
[anthropic.com/threat-intelligence-report-september-2026](https://anthropic.com/threat-intelligence-report-september-2026)

**ANTHROPIC**

Detecting and countering  
misuse of AI:  
September 2026

---

Published  
September 10, 2026

Detecting and countering misuse of AI: September 2026

# 04

---

**Hva kan vi  
gjøre med det?**

*En god og en dårlig nyhet ...*



# Risiko innenfra og utenfra

---

## INNENFRA

Agentene vi  
selv tar i bruk

Gode mål kan gi avvikende delmål – agenten  
trenger ikke være ondsinnet for å gjøre skade

## UTENFRA

Agentene som  
er rettet mot oss

Angriperne er raskere til å ta i bruk ny teknologi og  
bedre til å tilpasse seg hurtige endringer

# Den gode og den dårlige nyheten

---

## GOD NYHET

Oppgaven er ikke umulig

Risikoen kan håndteres med midlene og kompetansen vi allerede har tilgjengelig

## DÅRLIG NYHET

Den løser seg  
ikke av seg selv

Hver virksomhet må ta eierskap til egen situasjon og finne løsninger som er tilpasset egen kontekst

# Seks konkrete tiltak

*Samme oppgave som før – bare i et høyere tempo*

---

01

## Kartlegg agentene

Vit hvilke KI-agenter som kjører hos oss, og hva hver av dem har tilgang til

---

02

## Ha en stoppknapp

Bestem hvem som kan slå av en agent, og øv på hvor raskt det skjer

---

03

## Tål tempoet

Rekordmange sikkerhetsoppdateringer i måneden krever automatisert og prioritert patching

---

04

## Styrk det grunnleggende

Minst mulig tilgang, ingen nøkler i koden og overvåking av eksponerte systemer

---

05

## Styr dataflyten

Bestem hvilke opplysninger som kan brukes i hvilke KI-tjenester

---

06

## Still krav til leverandørene

Krev innsyn i hvor data lagres og behandles, og varsling ved hendelser

A serene landscape photograph showing a sunset or sunrise. The sun is a bright, glowing orb on the horizon, partially obscured by a layer of clouds. The sky is a mix of soft orange, yellow, and pale blue. In the foreground, a dark, silhouetted mountain ridge rises from a thick layer of white clouds that fill the valley. The overall mood is peaceful and contemplative.

TIL SLUTT

**Det står mye på spill.  
Vi må ta dette på alvor**

*Ikke dommedag – men eierskap til egen situasjon*

Foto: ki Zhang / Unsplash

# Takk

---

**FOREDRAGSHOLDER**

Joachim Hartgen-Sverdrup

**E-POST**

[jhs@varsyn.no](mailto:jhs@varsyn.no)

**NETT**

[varsyn.no](http://varsyn.no)

# Kilder

---

04      Utsagn: Jacob Coxon (BBC), Evan Hubinger (X), Anders Malthé-Sørenssen (Khrono), Richard Seymour (X), Donald Trump (Truth Social). September 2026

---

05      Skjermbilder: Nettavisen, Bergens Tidende, TV 2, Khrono, digi.no og Computerworld. September 2026

---

07      Foto: Ed Wingate / Unsplash

---

08–10    Epoch AI, *Epoch Capabilities Index (ECI)*. epoch.ai/benchmarks/eci (CC BY)

---

11      Epoch AI (2025), *LLM inference prices have fallen rapidly but unequally across tasks*. epoch.ai/data-insights/llm-inference-price-trends (CC BY)

---

12      Xiao, C. mfl. (2025). Densifying law of LLMs. *Nature Machine Intelligence*, 7, 1823–1833. doi.org/10.1038/s42256-025-01137-0

---

---

13      Stanford HAI (2026), *The 2026 AI Index Report*, kap. 2. hai.stanford.edu/ai-index/2026-ai-index-report

---

15      Foto: Stargate-campus i Abilene, Texas. Tencents datasenter i Gui'an, Guizhou (Jade Gao / AFP)

---

22      Engadget (11.10.2025) om ChatGPT-data. CNN (17.05.2024) om Arup. The Hacker News (12.06.2025) om EchoLeak

---

25      OpenAI (26.08.2026), teknisk rapport om Hugging Face-hendelsen. METR og Redwood Research (26.08.2026), uavhengig gjennomgang av hendelsen. metr.org

---

25–26    Anthropic (10.09.2026), *Detecting and countering misuse of AI: September 2026*. anthropic.com/threat-intelligence-report-september-2026

---

31      Foto: ki Zhang / Unsplash

---