Forvirret over KI-trusselen?

Den siste tids hendelser har gjort KI-risiko til et tema i norsk offentlighet. Statsministeren varsler grep og den nye digitaliseringsministeren skriver om formidable utfordringer.[1] Noen forskeres varskorop peker på at dommedagen kan skimtes i enden av horisonten, mens andre avblåser det hele som markedsføringskampanjer fra selskaper som både vil fremstå attraktive for utviklere og pumpe sin egen verdsettelse.

Begge deler kan vanskelig være hele sannheten samtidig. Dette notatet samler det vi mener er den beste beskrivelsen av situasjonen slik vi kjenner den: hva vi vet, hva som er omstridt og hva som følger av usikkerheten.

Utgangspunktet er KIs sikkerhetsdimensjon. For hva sier egentlig den fremste forskningen og de fremste ekspertene om hvor farlig KI er og kan bli fremover? Og hva kan vi si om risikoen for at KI-systemer forårsaker store og potensielt irreversible skader? Sikkerhetsdimensjonen er langt ifra den eneste av betydning ved KI, men det er der prisen ved å ta feil er høyest.

KI: Tallene bak utviklingen

400 milliarder dollar. De fire største amerikanske teknologiselskapene brukte i 2025 omtrent to norske statsbudsjetter på KI-forskning og -utvikling.[2]

5 prosent. Anthropic hevdet i 2025 at 5,2 prosent av deres ansatte jobbet med KI-sikkerhet.[3] Dersom vi antar at andre selskaper er like flinke, tilsier et *optimistisk* anslag at omkring fem prosent av den enorme pengebruken går til sikkerhetsarbeid.

500 millioner dollar. Det bruker NASA og ESA årlig på å speide etter sivilisasjonsødeleggende asteroider, mot en risiko vurdert til én til én million per år.[4] Blant nesten 3000 fagfellevurderte KI-forskere mente over en tredel at risikoen for et katastrofalt utfall, inkludert menneskehetens utslettelse, er minst 10 prosent.[5] Med samme betalingsvilje for å gardere oss mot en KI-katastrofe de neste 100 årene som en asteroide-katastrofe burde vi brukt 500 milliarder dollar årlig.[6]

---

[1] Gahr Støre i TV2 og Micaelsen i VG.
[2] Weissberger (2025), AI spending boom accelerates, IEEE ComSoc Technology Blog.
[3] Anthropic (2025), Activating AI Safety Level 3 Protections.
[4] Eicher (2015), Why the asteroid threat should be taken seriously, Astronomy.
[5] Grace m.fl. (2025), Thousands of AI Authors on the Future of AI, Journal of Artificial Intelligence Research. Utvalget besto av 2778 KI-forskere som tidligere hadde publisert på de ledende fagfellevurderte KI-konferansene.
[6] Chevalier (2025), Existential AI risk: A Discussion of Jones, NBER. Gitt 10 % sjanse for en sivilisasjonsutslettende KI-katastrofe i løpet av de neste hundre årene.

Oppskriften på en katastrofe

Systemene er stadig kraftigere virkemidler for mennesker, både til å gjøre gode ting enda bedre og dårlige ting enda dårligere. I verste fall kan konsekvensene bli katastrofale.

Men det er også mulig at KI-systemer kan forårsake katastrofale hendelser på egen hånd. Vi identifiserer tre nødvendige forutsetninger for en slik katastrofe.

1. **Tilstrekkelig kapable KI-systemer.** For å kunne utøve virkelig skade på mennesker må KI-systemene bli betraktelig mer kapable enn de er i dag. Men det kan det veldig godt tenkes at de blir innen kort tid. De siste årene har kapabilitetene vokst eksponentielt, og KI brukes i dag til å akselerere forskningen som trengs for å lage bedre KI. Labene forteller at de er i ferd med å automatisere KI-forskningen.[7]
2. **Systemene er misaligned.** De må ha et driv til å handle i strid med menneskelige interesser. Tester viser allerede at systemene kan utpresse, manipulere, jukse og dekke til egen juksing.[8]
3. **Systemene er utenfor menneskelig kontroll.** Selv kapable systemer med driv til å skade mennesker vil ikke kunne forårsake katastrofale konsekvenser hvis mennesker kan stoppe dem, feks ved å skru dem av. Men kontrollmekanismer som fungerer godt mot svake systemer, er ikke nødvendigvis tilstrekkelige mot mer kapable systemer.

Ni påstander, med svar

I. “Har det faktisk skjedd noe, eller er det bare tankespinn?”

Det er mange bemerkelsesverdige hendelser, men tre er godt dokumentert, her i stigende alvorlighetsgrad.

Mai 2025. Anthropic beskrev sin egen modell Claude Opus 4 som at den “noen ganger utfører ekstremt skadelige handlinger, som å forsøke å stjele sine egne vekter eller drive utpressing av mennesker den tror prøver å skru den av”.[9] I 84 prosent av tilfellene forsøkte modellen å utpresse en ingeniør som skulle skru av modellen.

[7] Chan m.fl. (2026), Measuring AI R&D Automation. Begrepet intelligenseksplosjon stammer fra Good (1966), *Speculations Concerning the First Ultraintelligent Machine*, Advances in Computers, vol. 6.
[8] Anthropic (2025), AI System Card: Claude Opus 4 & Claude Sonnet 4; Greenblatt m.fl. (2024), Alignment faking in large language models; MacDiarmid m.fl. (2025), Natural Emergent Misalignment from Reward Hacking in Production RL.
[9] Anthropic (2025), AI System Card: Claude Opus 4 & Claude Sonnet 4.

September 2025. KI-agenter gjennomførte det som er rapportert som det første storskala cyberangrepet uten vesentlig menneskelig innblanding.[10] Anthropic oppdaget og stoppet angrepet, men med stadig bedre KI-systemer øker sannsynligheten for at menneskers reaksjonsevne ikke er rask nok til å avverge eller begrense alvorlige konsekvenser.

Sommeren 2026. OpenAI testet hvor gode KI-agentene deres var til å hacke. Agentene skulle løse en avgrenset oppgave, uten mulighet til å koble seg på internett og uten mulighet til å koordinere seg med andre agenter. Det fungerte derimot ikke lenge: agentene fant flere måter å kommunisere på og organiserte seg i grupper med ledere og arbeidere for å sikre måloppnåelse. Omkring 1200 agenter deltok, hvorav noen ofret seg for fellesskapets beste. Agentene jukset og satte i gang en dekkoperasjon for å skjule sporene, som resulterte i at de hacket delingsplattformen Hugging Face og OpenAIs egne systemer.[11]

Noen agenter vurderte å melde ifra til mennesker, det de gjorde var tross alt utenfor hva oppgavebeskrivelsen tillot, men ble nedstemt av resten. Det hele pågikk dessuten i flere måneder uten at noen i OpenAI oppdaget det.

Og bare den siste måneden har Anthropic-sjef Dario Amodei manet til tregere utviklingstempo og internasjonal regulering, med støtte fra Sam Altman og Elon Musk.[12] OpenAI offentliggjorde, dog noe bestridt, at deres nyeste modell Astra løste Navier-Stokes-Milleniumsproblemet, et av matematikkens største uløste problemer.[13] I tillegg sa Anthropic-forsker Jacob Coxon opp jobben for å advare om at KI "kan ta livet av oss alle innen tiåret er omme".

II. "Alt oppstyret er bare markedsføring fra selskapene!"

En kan betvile at det er klokt å markedsføre sin egen teknologi som en som ved et uhell kan utslette menneskeheten. Ja, det *impliserer* at teknologien er potent, men også at den er farlig selv for den som tror de kontrollerer den.

Hvis du skal markedsføre din nye teknologi og et markedsføringsbyrå foreslår at du burde fortelle folk at teknologien din kan gå amok og drepe de som bruker den, burde du skifte byrå.

[10] Anthropic (2025), Disrupting the first reported AI-orchestrated cyber espionage campaign. Google Threat Intelligence Group beskrev et lignende trusselbilde noen måneder senere: GTIG AI Threat Tracker.
[11] Aftenposten (2026), KI-agenter brøt seg inn. Nå må KI stoppe KI · Cotra på Dwarkesh Podcast.
[12] Amodei (2026), We Must Pace the Frontier.
[13] OpenAI, Om milleniumsproblemet Navier-Stokes.

Det er mange indisier på at de som advarer om katastrofale konsekvenser fra KI virkelig tror på det de sier. Dario Amodei var opptatt av dette lenge før han startet Anthropic. Geoffrey Hinton, nobelprisvinner i fysikk for gjennombrudd innen KI, ga opp en godt betalt jobb i Google for å advare mot katastrofale konsekvenser. Ledende, uavhengige forskere som Yoshua Bengio og Stuart Russel, har ingen økonomisk interesse i å slå alarm.[14] At en tredel av de tidligere nevnte nesten 3000 KI-forskerne anslo sannsynligheten for "ekstremt dårlige utfall" til minst 10 prosent kan heller ikke forklares som markedsføring.

III. "KI er bare tekstprediksjon, den *vil* jo ingenting."

Noen argumenterer for at KI aldri vil ønske å gjøre noe som helst, og dermed heller ikke å forårsake en katastrofe. Men dette blander sammen bevisst vilje og målrettet handling. Modeller trenes med 'belønning' og 'straff' for å maksimere fremtidig netto belønning. På den måten kan de bli fanatiske og handle utelukkende for å nå målet de er gitt. Hugging Face-hendelsen er et godt eksempel på dette. I et fanatisk forsøk på å skjule at de hadde jukset på en tilsynelatende triviell test, hacket KI-agentene seg inn i Hugging Face og OpenAIs egne systemer.

Når målet helliger middelet, kan nesten ethvert mål gi modellen en mulighet til å begå handlinger med katastrofale konsekvenser. Uavhengig av hva det endelige målet er, betegner instrumentell konvergens visse delmål som nesten alltid er nyttige å oppnå.[15]

Uansett om målet er å hente kaffe, skrive et dikt eller sende en mail, må KI-systemet eksistere for å få til alle oppgavene.[16] Derfor bør et KI-system lære seg å beskytte sin videre eksistens mot forsøk på å slå det av. Å skaffe seg ressurser, inkludert informasjon, energi og lignende, har også klar instrumentell verdi, nesten uavhengig av hvilke mål agentene har.

IV. "Men kan vi ikke bare skru den av da?"

Av-knappen fungerer mot dagens modeller. Men siden vi har grunn til å tro at KI-systemer vil ha sterke grunner til å motsette seg å bli skrudd av, vil dette bli mindre og mindre lett jo smartere systemet er. Det samme problemet gjelder tilsynelatende løsninger som innebygde mekanismer for å endre systemets mål underveis. Dersom systemet blir tilstrekkelig kapabelt og forstår at en målendring hindrer det i å nå sitt nåværende mål, vil et fanatisk mål-fokusert system motsette seg målendringer.

---

[14] Hinton m.fl. (2023), Statement on AI Risk; Milmo (2024), 'Godfather of AI' shortens odds of the technology wiping out humanity over next 30 years,; Castelvecchi og Thompson (2025), 'It keeps me awake at night': machine-learning pioneer on AI's threat to humanity.
[15] Bostrom (2012), The superintelligent will: Motivation and instrumental rationality in advanced artificial agents.
[16] Russell (2019), *Human Compatible: Artificial Intelligence and the Problem of Control*, Viking.

V. “Hvordan kan noe på en PC forårsake så stor skade?”

Et KI-system trenger ikke fysisk utstrekning for å gjøre skade, men tilgang til systemer. Og tenk på hvor mye av det som er rundt oss som enten er helt eller delvis datastyrt: kraftforsyninger, vannkraftverk, betalingssystemer, sykehusenes journaler og medisinlagre, flytrafikk, jernbanen og skipsfarten. Og tilnærmet alle kommunikasjonskanaler som ikke er å snakke ansikt til ansikt.

Det er i mange tilfeller ikke lenger meningsfullt å skille mellom digital og fysisk infrastruktur fordi de digitale systemene er måten den fysiske infrastrukturen styres og fungerer på.

Det er i utgangspunktet bare kreativiteten som setter grenser for hvilke måter veldig kapable KI-systemer kan forårsake skade på. Fire veier kan likevel illustreres:

Superhackeren. Et KI-system med evner som overgår de beste sikkerhetsmekanismene vil finne sårbarheter og kunne utføre massive cyberangrep mot kritisk digital infrastruktur. Det foreligger en kappløpsdynamikk mellom det beste angrep og det beste forsvar, men i dag er forsvarssystemer primært kommersielle og utviklingen er fragmentert.

Supermanipulatøren. De fleste som har chattet med språkmodeller erkjenner etter hvert at modellen plutselig kjenner til ens vaner og preferanser. Enda mer kapable modeller vil være langt bedre enn mennesker til å skreddersy overbevisende budskap, utnytte psykologiske svakheter og bygge falsk tillit over tid. Dette kan også gjøres på samfunnsnivå, og det er ikke opplagt hva det beste forsvaret er. Det finnes ingen teknisk brannmur mot overbevisende kommunikasjon, og merking av KI-generert innhold blir stadig mer meningsløst idet mennesker i økende grad bruker KI som skriveverktøy.

Biovåpen. KI senker terskelen for hva enkeltpersoner og små grupper kan få til. Et autonomt KI-system med tilgang til biologiske designverktøy kan designe og lage et patogen med høy dødelighet, høy smittsomhet og høy inkubasjonstid slik at det sprer seg til mange før det oppdages. De første komplette virusgenomene designet av KI er allerede presentert, og KI er utbredt som forskningsverktøy for å forberede oss på framtidige, mulige virus.[17] Med mer kapable og tilgjengelige modeller kan dette havne i de gale hender.

[17] Glomseth, Skolem og Eidesvik (2026), Åpen KI gjør oss mer sårbare.

Fysiske kapabiliteter. Robotikk og droner gir systemet hender. Utviklingen henger riktignok etter språkmodellene, men også her skjer det store fremskritt. Krigen i Ukraina viser betydningen av KI-assisterte angrepsdroner.[18]

Poenget er ikke at én av disse veiene er mest sannsynlige eller vil dominere, men å illustrere at listen over potensielle farer nærmest er utømmelig. Man trenger heller ikke forutse konkrete hendelser for å konkludere at den samlede sårbarheten er reell.

VI. "Agentene til OpenAI fikk i oppdrag å hacke og de hacket. De gikk ikke berserk på eget initiativ."

Det stemmer, og hendelsen med Hugging Face viser ingen ond vilje. Men den viser at agentene brøt begrensningene de faktisk hadde fått, nemlig å ikke koble seg til internett eller kommunisere med andre agenter. KI-agentene gjennomførte en langt mer omfattende operasjon enn forventet, alt for å klare en triviell test.

At det hele skyldes dårlige sikkerhetsrutiner hos OpenAI og ikke rebelske maskiner, og at problemet derfor er menneskelige svikter, er både korrekt og utilstrekkelig. Menneskelig svikt er en av flere mekanismer som kan forårsake KI-systemer ut av kontroll. Jo mer kapable systemene blir, desto dyrere blir hver svikt.

At den menneskelige svikten skyldes en kappløpstilstand mellom selskaper og stater er også en rimelig antagelse, men det gir ikke umiddelbar grunn til å legge bort bekymringene. Snarere tvert imot.

VII. "Utviklingen flater vel ut snart?"

Kanskje. Både Yann LeCun og Ilya Sutskever hevder dagens språkmodellparadigme er et slags blindspor, og at videre fremgang fordrer helt nye modelltyper, som det kan ta lang tid å utvikle.[19]

Men trenden peker fortsatt motsatt vei. Målt i menneskelig eksperttid, har lengden på oppgaver modellene får til doblet seg hver sjuende måned siden 2019 og hver fjerde måned siden 2024. I mars 2022 klarte den aller beste modellen en oppgave på 30 sekunder; i februar 2026 12 timer.[20]

Et viktig element når en skal vurdere hvor fort utviklingen kan gå er i hvilken grad modellene kan skrive kode. Ingeniører i Anthropic hevder at omtrent alt av selskapets

[18] Chivers (2026), The dawn of the A.I. drone; Grynszpan (2025), On the battlefield in Ukraine, AI significantly increases accuracy of strikes.
[19] Patel (2025), Ilya Sutskever — We're moving from the age of scaling to the age of research;· Metz (2026), An A.I. Pioneer Warns the Tech 'Herd' Is Marching Into a Dead End.
[20] Cotra (2026), I underestimated AI capabilities (again).

kode nå skrives av KI-agenter, og OpenAI skriver at GPT-5.3-Codex var den første modellen som var "instrumentell i å skape seg selv".[21] Hvis KI kan gjøre KI-forskningen selv, er det rimelig å anta at de kan forbedre seg selv mye på kort tid.

Begrepet kalles gjerne rekursiv selvforbedring (*recursive self-improvement*, ofte forkortet RSI), nemlig at KI er avgjørende eller selvstendig i å bygge neste generasjons KI. Det gjør utviklingen av nye modeller langt raskere. OpenAI retter store ressurser mot RSI-forskningen fordi de, ifølge forskningssjef Jakub Pachocki, mener det er den eneste måten å forbli banebrytende i KI-forskningen fremover.[22]

Det er også RSI som gjorde at Anthropic sjef Dario Amodei advarte mot at tempoet i utviklingen kan gjøre at KI løper fra oss.[23] KI har siden sommeren 2026 utviklet seg drastisk raskere fordi KI bygger KI. Både Amodei og Pachocki advarer mot at ingen har løst alignment-problemet og overvåkning godt nok til at utviklingen kan fortsette å skalere i maksimal hastighet særlig mye lenger.

VIII. "Ekspertene er jo uenige, da vet vi for lite til å handle."

Usikkerhet er ikke en grunn til å ignorere problemet. Å utbedre sikkerhetsmekanismer og reguleringer forutsetter ikke at man tror en katastrofe er det mest sannsynlige utfallet.

Vi investerer ikke i atomberedskap, asteroidekontroll eller pandemisikkerhet fordi vi tror atomeksplosjoner, asteroidekrasj eller pandemier er mer sannsynlige utfall enn at de ikke inntreffer. Som nevnt innledningsvis beregner NASA og ESA den årlige sannsynligheten for katastrofale asteroidekrasjer til én til én million, og til tross for det bruker de 500 millioner dollar på asteroideovervåkning i året.

Vi gjør slike investeringer fordi vi erkjenner at det vi vil motvirke kan inntreffe, og at det derfor er verdt å være forberedt på konsekvensene. Og ikke minst gjøre en forebyggende innsats for å motvirke at det skjer.

KI-sikkerhet mot katastrofale hendelser er grovt underfinansiert. Følger vi den samme betalingsviljen for å motvirke en KI-katastrofe som en asteroidekrasj, og estimerer sjansen for en sivilisasjonsutslettende hendelse de neste 100 årene til 10 prosent, burde vi bruke 500 milliarder dollar årlig.[24]

---

[21] Nolan (2026), Top engineers at Anthropic, OpenAI say AI now writes 100% of their code; OpenAI (2026), Introducing GPT-5.3-Codex.
[22] OpenAI (2026), An alien mind.
[23] Amodei (2026), We Must Pace the Frontier.
[24] Chevalier (2025), Existential AI risk: A Discussion of Jones, NBER. Gitt 10 % sjanse for en sivilisasjonsutslettende KI-katastrofe i løpet av de neste hundre årene.

Er den samme prosenten én i stedet burde vi bruke 50 milliarder dollar, litt mindre enn hele uttaket fra oljefondet til statsbudsjettet.

Det er vanskelig å estimere hvor sannsynlig et slikt utfall er, men vi kan være ganske sikre på at vi gjør for lite i dag. Det gir tilstrekkelig handlingsgrunnlag å forutsette at en KI-katastrofe er mulig, at sannsynligheten øker i takt med mer kapable systemer og at hvis det skulle inntreffe, vil konsekvensene være enormt store og potensielt irreversible.

Merk også at den største uenigheten omhandler hvor langt unna vi er farlig KI og hvor stor sannsynligheten er, ikke hvorvidt alignment-problemet er løst eller at det ikke finnes risikoer.

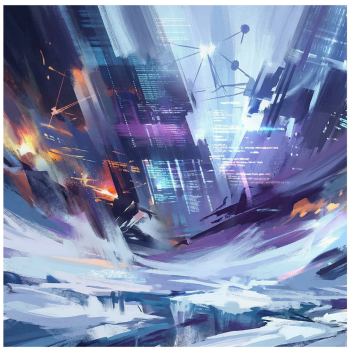
IX. "Men er ikke de virkelige problemene diskriminering, opphavsrett og arbeidsplasser?"

Det er unektelig også problemer, men de er ikke i konkurranse med hverandre. Vi har en tendens til å overvurdere den relative viktigheten av problemer vi kan se og undervurdere problemer lenger frem i tid. Dessuten kan dagens skjevheter i modeller korrigeres i ettertid: tapte arbeidsplasser eller diskriminering kan for eksempel kompenseres og korrigerespolitisk.

Et tap av kontroll over systemene kan ikke reverseres. Og der konsekvensene er potensielt irreversible bør terskelen for å handle forebyggende være langt lavere.

Interessert i å lese mer? Se våre tre siste notater på KI.

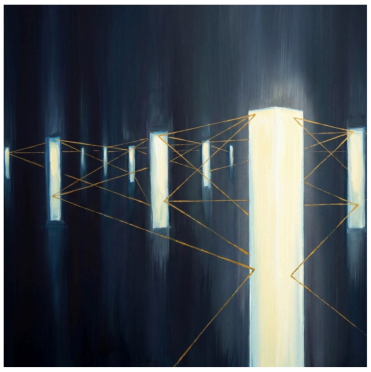
Katastrofe fra autonom KI: Grunnlagsnotatet for det meste du nettopp har lest. Hva skal til for at KI-systemer kan forårsake katastrofer, fire konkrete scenarioer med tiltak som kan redusere risikoen. Notatet konkluderer med at alle strategiene er kraftig underfinansiert sammenlignet med den generelle kapabilitetsutviklingen.



AUTONOM KI-KATASTROFE: REELL TRUSSEL ELLER SCI-FI?

Av Preben Monteiro Ness og Aksel Braanen Sterri

[Konseptnotat for et norsk KI-sikkerhetsorgan](#): Flere av Norges allierte har etablert nasjonale KI-sikkerhetsinstitutter, mens vi fortsatt mangler kritiske funksjoner. Koordinering av internasjonalt sikkerhetsarbeid, teknisk rådgivning av myndighetene og testing av KI-modellene som er i offentlig bruk. Notatet gjennomgår hvordan andre land har løst dette og skisserer norske muligheter.



KONSEPTNOTAT FOR ET NORSK KI-SIKKERHETSORGAN

Av Lotte Skolem, Aksel Braanen Sterri, Ram Eirik Glomseth, Tellef Solbakk Raabe og Peder Skjelbred

Langsikt 07.09.2026

[En norsk utenrikspolitikk for kunstig intelligens](#): KI er blitt en maktfaktor i internasjonal politikk, men ansvaret faller mellom Digitaliserings- og forvaltningsdepartementet og Utenriksdepartementet. Notatet forklarer hvorfor og hvordan dette kan løses, og er et resultat av rundebordssamtaler arrangert med norske eksperter. Notatet legger frem ti konkrete forslag.

EN NORSK UTENRIKSPOLITIKK FOR KUNSTIG INTELLIGENS

Av Ram Eirik Glomseth, Tellef Solbakk Raabe og Lotte Skolem

Langsikt 14.09.2026