

Appel Olsen og Salling: AI dræber – men ikke som tech-eliten bilder os ind

Vil AI's fremmarch resultere i en superintelligent computergud, der kan udslette menneskeheden? Evidensen er tynd. Men at AI- og datacenterindustrien lige nu smadrer natur og samfund på et væld af andre, helt konkrete måder, burde være varsel nok.

Kronik af Frederik Appel Olsen og Caroline Anna Salling

21. oktober 2026

AI vil blive superintelligent og selvstændig. Ikke nok med det: Den vil nok også slå os alle ihjel. Måske!

Sådan lyder advarsler fra flere medarbejdere i de store AI-virksomheder. Særligt den nu tidligere Anthropic-medarbejder Jacob Coxon trak overskrifter for nogle uger siden, da han på X annoncerede, at han sagde sit job op. Det er ifølge ham, fordi virksomheder som OpenAI og Anthropic arbejder mod en superintelligens og dermed “gambler med vores liv”.

OpenAIs topchef, Sam Altman, [sagde allerede i 2023](#), at “AI højst sandsynligt vil føre til verdens undergang”. Derefter har han kun skruet op for udviklingen af AI. Nu er andre AI-arbejdere og eksperter – ja, selv lederne i branchen – begyndt at sætte procenter på risikoen for udslettelse.

Evan Hubinger, en anden Anthropic-medarbejder, [mener](#), at der er mere end ti pct. sandsynlighed for, at AI vil kunne dræbe alle mennesker. Den pensionerede University of Toronto-professor og Nobelprisvinder Geoffrey Hinton [sagde for rullende kameraer](#), at det “ikke er urimeligt” at antage, at AI med ti pct. sandsynlighed kunne “dræbe os alle i løbet af det næste årti”.

Frygten for, at AI vil vriste sig fri af ingeniørernes greb og trykke *delete* på mappen ”Mennesker”, er altså [omfangsrig](#) og lader til at være bygget på to præmisser: For det første på teorier om endnu ikke udviklet AI-superintelligens eller AGI (generel kunstig intelligens), der minder om en cyber-Gud, som går langt ud over de teknologier der er tale om i dag. For det andet på nylige episoder, hvor AI-agenter fra de store tech-firmaer har hacket sig ind et sted, hvor man ikke ville have dem til at hacke sig ind.

Det handler altså *ikke* om datacenter-udviklingens accelerering af klima- og miljøødelæggelse, AI's brug i militær sammenhæng, produkternes assistance i konkrete selvmordssager eller de mange andre ødelæggelsesformer, som tech-industrien er indsmurt i.

Disse er meget håndgribelige ødelæggelser, der foregår lige nu og ikke i en eller anden uspecifik fremtid. Dem vender vi tilbage til. Lad os først se på de to præmisser for AI-dommedag én for én.

Computerguden vil åbenbart slå dig ihjel

Frygten for, men også håbet om en maskine med selvstændig handlekraft har floreret i Silicon Valley længe før virksomheder som Anthropic og OpenAI udgav deres store sprogmodeller.

Selvstændig handlekraft vil efter sigende blive udviklet ved såkaldt [recursive self-improvement](#). Det betyder, at AI-modeller vil medføre stærkere modeller, som igen vil udvikle endnu stærkere og mere “intelligente” modeller, indtil resultatet er en superintelligens. En intelligens, der overgår vores menneskelige fatteevne og visdom.

Idéen går tilbage til 1960’erne og matematikeren I.J. Good. Han rådgav bl.a. Stanley Kubrick om supercomputere i forbindelse med *2001: A Space Odyssey* og var en af de første til at tale om en sådan form for “eksplosiv intelligens” i computersystemer. I.J. Good [mente](#) ligesom mange tech-folk i nutidens Silicon Valley, at “menneskets overlevelse afhænger af en tidlig introduktion af en ultraintelligent maskine”.

AI vil altså på et tidspunkt blive alvidende og vriste sig fri af vores kødelige, dødelige menneskehænder og opnå ubegrænset handlekraft. Denne computergud vil blive menneskets sidste opfindelse, da den enten vil redde os alle eller slå os alle ihjel. Et smigrende sci-fi-scenarie hos tech-eliten, der kan forstå sig selv som menneskehedens skæbnemagere. Det gør nok heller ikke noget negativt for forestående børsnoteringer.

Men vi har endnu ikke set eksempler på en sådan dystopisk superintelligens i de tekniske udviklinger, der rent faktisk finder sted. Den påstand leder os til næste præmis.

Går kunstig intelligens ”rogue”?

Da Jacob Coxon sagde op i Anthropic, var det særligt en episode, hvor AI-agenter havde brudt ud af deres træningsmiljø og hacket eksterne servere, der skulle vise, hvor farlig AI er for menneskehedens eksistens. Lad os se på, hvad der skete, og om dét understøtter tech-elitens messianske fremskrivninger af dommedag.

Virksomheder såsom Anthropic træner deres store sprogmodeller for at komme først i AI-kapløbet. I sommer var et hold af softwareingeniører i gang med at træne en model i forskellige hackeropgaver.

Det foregik i et træningsmiljø, et såkaldt *ExploitGym*, hvor AI-agenter blev sat sammen med et *harness*, dvs. en konkret software-infrastruktur. Store sprogmodeller kan nemlig ikke udføre nogen handlinger på egen hånd, men kun generere sandsynlig tekst. Derfor er det nødvendigt, at mennesker udfører modellernes instruktioner eller aktivt kobler dem med et *harness*.

Når nogen taler om en “AI-agent”, så refererer de altså til en sprogmodel koblet med software, der kan eksekvere prompts. Opgaven i *ExploitGym* er at løse forskellige hackeropgaver. Derfor er modellen trænet på erfaringerne fra hackere.

Og det er her, det kan gå grueligt galt. Hvis ikke virksomhederne, der kører AI-træningen, opsætter nogle sikkerhedsforanstaltninger – ligesom man skal, hvis man tester selvkørende biler, et medicinsk stof eller en rumraket – kan forsøget komme ud af kontrol.

Men eftersom OpenAI har travlt i kapløbet med andre AI-giganter på dette område, havde man ikke sørget for sikkerhedsmekanismer i AI-agenternes træningsmiljø. De kørte med andre ord rundt på egen hånd. Til sidst fandt de ud af at “løse” ExploitGym-opgaverne ved at tilgå virksomheden Hugging Face, som opbevarede svarene på opgaverne på deres servere.

Hugging Face-episoden viser helt klart, at AI kan bruges destruktivt, hvilket vi allerede vidste. Særligt hvis man arbejder med cybersikkerhed, maner den til bekymring. Men springet derfra og til at AI på en eller anden måde skulle kunne sætte sig for og udføre massedrab, [er astronomisk](#).

Og i danske og internationale medier bliver episoden sammenlignet med, at AI går “rogue”. Forestillingen er, at AI har sin egen vilje, og at der derfor gennem endnu undefinerede mekanismer pludselig skulle kunne udvikle egne interesser, for dernæst at være interesseret i at dræbe os alle en dag i løbet af de næste årtier. Måske endda om få år. Men sådanne forestillinger har umiddelbart intet med førnævnte superintelligens at gøre. AI-agenter af denne type – software, der prompter en sprogmodel, giver et svar, som så eksekveres af softwaren, hvilket gentages igen og igen i lange sekvenser – bliver brugt hver dag uden problemer.

Til gengæld må nogen tage ansvar for, at agenterne ikke løser opgaven, sådan at de yder skade på mennesker og økosystemer. Det svarer til at lade selvkørende bulldozere køre rundt bag et spinkelt hegn, instrueret i at jævne alt på deres vej. Derefter finder de på at buldoze et hus på den anden side. SelvkørendeBulldozer Inc. kunne have trukket stikket undervejs, men de holdt ikke lige øje med situationen i et par dage.

Når Sam Altman [erklærer, at man nu har sat den slags AI-forsøg på pause](#), er det et forsøg på at fremstå heroisk, der kun kan gavne OpenAI’s investorfront. Reelt set handler man på bagkant, efter man var sløset med sikkerheden i første omgang. Hvor mange andre teknologiudviklende virksomheder får uhindret lov til først at udvikle usikre algoritmer og derefter fremhæve sig selv som den ansvarlige frelser?

AI ødelægger verden lige nu

Uagtet om tech-eliten selv tror på deres dommedagsprofetier, så har de en klar retorisk funktion: at få offentligheden til at vende blikket bort fra den meget konkrete ødelæggelse af samfundet, som AI-virksomheder forårsager lige nu og her.

Opgaven for alle os andre er ikke at ryste i bukserne over computergudens ankomst, men at fastholde blikket på de meget håndgribelige ødelæggelser, som AI-industrien allerede forårsager.

Mens du læser dette, ødelægger AI klimaet og miljøet. De enorme mængder af computerkraft, det kræver at udvikle de store sprogmodeller, sluger så [ekstreme mængder af energi](#), at vi kan glemme alt om vores i forvejen skrøbelige planer om grøn omstilling. Datacentrene [bruger milliarder af liter vand](#), især i regioner, [der i forvejen lider under vandstress](#).

AI er årsagen til, at [nye datacenterprojekter beslaglægger el og accelererer forbruget af fossile brændstoffer](#). AI er også årsagen til, at techvirksomheder som Microsoft, Amazon og Google laver [nye samarbejder](#) med olie- og gasindustrien, såsom ExxonMobil, Equinor, Chevron og TotalEnergies.

Som professor i miljøpolitik Leah Stokes [formulerer det](#): “Det interessante er, at til forskel fra AI, så er den eksistentielle trussel [fra klimaforandringer, red.] ikke hypotetisk. Det dræber mennesker nu.” Industrien anerkender, at AI-udviklingen og reduktioner i drivhusgasudledninger ikke kan sameksistere. Den europæiske industriforening for datacentre [var såmænd senest ude i juni](#) og forkyndte, at Europa bliver nødt til at vælge mellem AI og klimamål.

Mens du læser dette, [bruges AI til mord](#). AI-”bemandede” droner [dræber mennesker i Ukraine](#) og [udvælger “mål” for det israelske militær i Gaza](#). Det amerikanske militærs drab af 150 mennesker, heriblandt 123 børn, på en pigeskole i Iran [var et resultat af et sådant AI-drevet program](#).

Og listen, der tegner det ødelæggende mønster, fortsætter og fortsætter. Mens du læser dette, eroderer AI [uddannelser](#) og [vores evne til kritisk tænkning](#). Mens du læser dette, bruges AI til at [plagiere og stjæle andre menneskers arbejde](#) og [underminere intellektuel ophavsret](#). Mens du læser dette, spiller AI en [rolle i menneskers selvmord](#). Mens du læser dette, [eroderer AI vores privatliv](#). Mens du læser dette, smadrer AI forsøg på demokratisk praksis med [misinformation](#), [valgkampsslop](#) og [propaganda](#). Mens du læser dette, [udnytter AI-industrien de arbejdere i det globale syd](#), der træner sprogmodellerne.

Nej, der er ikke tegn på, at en ondsindet superintelligens vil slå os alle ihjel. Den ødelæggelse, som AI-industrien frembringer, er til gengæld meget mere virkeligt og meget, meget dummere.

Frederik Appel Olsen (f. 1992) er postdoc ved Center for Applied Ecological Thinking, Københavns Universitet.

Caroline Anna Salling (f. 1994) er postdoc ved Institut for Kommunikation og Humanistisk Videnskab, Roskilde Universitet.