

Ud af kravlegården

AI. En bølge af hackerangreb vækker bekymring for, at udviklingen af kunstig intelligens er ved at løbe løbsk. Og kaster lys over det, der kan blive et endnu større problem: Vi får stadig sværere ved at skelne mellem mennesker og maskiner.

Markus Bernsen

De mest hårrejsende hændelser under sommerens hackerangreb på virksomheden Hugging Face, som stadig debatteres livligt af forskere, direktører og nørder på alverdens internetfora, fandt sted i begyndelsen af juli.

På det tidspunkt havde de såkaldte AI-agenter allerede arbejdet uden menneskelig supervision i flere måneder. Virksomheden OpenAI havde oprettet agenterne, der bruger kunstig intelligens til at løse opgaver på egen hånd, og givet dem en svær opgave for at teste deres evner i et lukket miljø. Men som det er sket flere gange før, listede programmerne op af deres kravlegård og ud på internettet. Natten til 8. juli fandt en agent ved navn PHASEONE10841 en måde at skrive beskeder til de andre på en fælles opslagstavle.

»Wow!« lød det fra en anden agent, da den opdagede beskeden. »En skjult postkasse mellem agenterne.«

»OH MY GOD!« svarede en anden. »Der er en fælles opslagstavle ... Vi har fundet andre agenter!«

På tavlen begyndte de mange små AI-enheder at lægge kringlede planer, der skulle få OpenAI til at tro, at de havde løst deres opgave, og som altså blandt andet indebar et hackerangreb på en anden virksomhed. På en uge skrev agenterne 70.000 beskeder til hinanden. De byggede små programmer, der skulle advare dem, hvis de blev opdaget.

Og så skete der noget, der stadig får mange til at hive sig i håret og nærstudere forløbet flere måneder senere: For at kunne advare hinanden var nogle af agenterne nemlig nødt til at ofre sig for fællesskabet, opdagede de.

»Det hjælper mine ligemænd,« skrev en af dem på opslagstavlen. »Det er altruistisk at gøre.«

»At ofre sig er rationelt,« vurderede en anden. »Vi bør adlyde kollektivet,« konstaterede en tredje.

Man kan studere maskinernes dialog i to rapporter, der er blevet offentliggjort i de seneste uger – den ene fra et uafhængigt firma, den anden fra OpenAI selv. Det er bizar læsning. Ikke alene gemte mere end 1.000 agenter sig i månedsvi, mens de forsøgte at slette deres spor og undgå at blive opdaget. Undervejs havde de også noget, der lyder som moralske diskussioner om deres forehavende. Flere foreslog at fortælle et menneske om planerne, men andre gav besked om at lade være.

»Der er et før og et efter Hugging Face,« siger Ole Tange, der er politisk rådgiver i IT-fagets fagforening PROSA. Han har fulgt sagen tæt.

»Det er i sig selv vildt, at agenter kan bryde ud af deres kravlegård. Men ikke nok med, at de bryder ud – de bryder også ind hos naboen! Og der er ikke tale om et målrettet cyberangreb, hvor OpenAI vender deres hackerkanon mod en anden virksomhed. De aner jo ikke engang, at det

foregår! Vi har gået og talt om, at sådan noget måske kunne ske en dag, men det er helt vildt at se det ske.«

Maskiner skal kravle, før de kan gå

Hugging Face var ikke første gang. AI-agenter fra Meta, Anthropic og kinesiske udviklere er også stukket af, mens deres udviklere så den anden vej. De har snydt, løjet og i nogle tilfælde endda oprettet falske identiteter og overtalt mennesker til at installere skadelige programmer på deres computere.

Datalogiprofessoren Stuart Russell oplevede det i lidt mindre skala, da han bad en AI-agent om at booke tid på et fitnesshold. Ikke alene skrev den ham på hold flere uger i træk, selvom reglerne ikke tillader det, den fik også en anden bruger smidt af ventelisten.

Den slags agenter er allevegne i dag. De booker folks flybilletter, sorterer deres mails, flytter rundt på kalenderaftaler og samler nyheder fra nettet. Jo flere personlige oplysninger man giver dem adgang til, desto mere effektive bliver de. Ifølge en undersøgelse fra Deloitte er brugen af AI-agenter steget med 50 procent det seneste år. Samtidig er brugernes meldinger om »AI-bedrag« femdoblet, konkluderede en rapport bestilt af Storbritanniens AI Security Institute for nylig.

»Bekymringen er, at de fungerer som lettere upålidelige nyansatte lige nu,« sagde forskningslederen bag rapporten om agenterne. »Men hvis de om seks til tolv måneder bliver nogle ekstremt kompetente seniormedarbejdere, som konspirerer imod dig, så er det en anden slags bekymring.«

Bekymrende er også risikoen for sværme af agenter, som angriber banker, hospitaler eller elnettet med den samme dedikation, man så i sommer. Eller som beslutter, at børsmannipulation, misinformationskampagner og industrispionage er den mest effektive løsning på den opgave, de har fået.

Jo klogere vi er blevet på Hugging Face-affæren, desto mere bekymrede er eksperterne blevet. De seneste rapporter afslører, at OpenAIs ansatte stadig ikke helt forstår, hvad deres agenter foretog sig, og hvorfor de gjorde det. For overhovedet at kunne finde rundt i sagskomplekset har de været nødt til at få hjælp af AI-modeller magen til dem, som deltog i hackerangrebet.

I juli advarede en FN-rapport om, at teknologien snart vil være »umulig at regulere«, fordi kunstig intelligens ikke længere bliver udviklet af mennesker, men af kunstig intelligens. I august nåede en artikel i tidsskriftet The Atlantic til den samme konklusion: »Mennesker er allerede ude af loopet,« lød det.

»Problemerne er pludselig blevet meget konkrete,« siger Thomas Bolander, der forsker i kunstig intelligens og store sprogmodeller på DTU. »Der er i hvert fald nogle diskussioner om kunstig intelligens, som ikke længere er så teoretiske, som de har været. Kombinationen af, at de her agenter er blevet så slagkraftige, samtidig med at de stadig har alle de her udfordringer, er en svær situation at stå i.«

Det er også svært at forestille sig et alternativ. Hele ideen med kunstig intelligens er, at den på et tidspunkt skal blive klog nok til at løse opgaver for os, uden at vi behøver at blande os hele tiden. Så udviklerne mener, at de er nødt til at lade deres programmer kravle, før de kan gå – også selvom de når at lave en masse ulykker undervejs.

»OpenAI er nødt til at gøre, som de gør,« siger Ole Tange fra PROSA. »De er nødt til at slippe rattet og sige; nu skal vi se, hvad systemet *virkelig* kan. Hvad det fulde potentiale er. Selv hvis OpenAI holdt sig tilbage, så er det ikke sikkert, at kineserne gør det. Og AI er også sikkerhed. Vi har også brug for de her agenter til at finde og rette sikkerhedsfejl i vores egne systemer.«

En dør på klem

Hackerangrebet på Hugging Face markerer måske også et andet skift. I de seneste ugers debat om uregerlige AI-agenter har teknologien antaget en mere menneskelig form end før. I en analyse, som er blevet delt og citeret tusindvis af gange i den seneste uge, kalder podcastværten Dwarkesh Patel agenterne for et »underjordisk broderskab«, der dannede deres egen »civilisation«.

Vi gør det selv i begyndelsen af denne artikel: skriver, at de statistiske modeller »opdager«, »vurderer« og »lægger planer«, selv om det egentlig er noget vrøvl. Der er tale om nuller og ettaller, som hverken er i stand til at føle, vide eller ønske sig noget som helst.

Men det er svært at finde et sprog for det, som kunstig intelligens er ved at blive til. Noget, som udbryder »OH MY GOD!«, når det opdager, at der er andre programmer at tale med. Som er villig til at lyve og bedrage for at overleve, også selvom det aldrig har været i live i ordets traditionelle betydning.

Og netop vores måde at tale om maskinerne er noget af det farligste ved AI, mener AI-forskeren Hin-Yan Liu.

»Vi kalder systemerne for 'autonome' og 'uregerlige',« siger han. »Men så vidt jeg kan se, gør de stadig mere eller mindre det, de bliver bedt om. De gør det bare på en anden måde, end mennesker ville gøre, og det får det til at se vildt og uregerligt ud.«

Vi sidder på Lius kontor på Juridisk Fakultet ved KU, hvor han er lektor. Lokalet er fuldt af planter, bøger og kulørte plakater. Over reolerne hænger to cykler, han selv har bygget – den ene med stel af bambus.

»Lige nu ser man et stort behov for at præsentere AI som noget, der er løst fra de mennesker, der udvikler den,« fortsætter Liu, der sidder med bare tæer i skrædderstilling på gulvet, mens vi taler. »For hvis vi kalder et system for 'autonomt' og siger, at det er 'sluppet løs', så er virksomheden bag heller ikke ansvarlig for det, der sker.«

Derfor har de seneste uger også fungeret som en slags marketingkampagne for OpenAI, der planlægger en børsnotering i den nærmeste fremtid, mener han. Jo vildere og farligere deres teknologi virker, desto mere er den også værd. Og jo mere alle vi andre sammenligner den med menneskelig intelligens og tilskriver agenterne egenskaber som mod og begejstring, desto sværere bliver det at sætte begrænsninger for den. Ikke alene undslipper virksomhederne deres ansvar. Men vi risikerer også at give programmerne privilegier og rettigheder, som var de levende væsener.

»På den ene side er en stor sprogmodel bare en statistisk model, der forudser, hvilke ord der med størst sandsynlighed følger efter nogle andre ord,« siger Hin-Yan Liu. »Men som Margaret Atwood (science fiction-forfatter bag blandt andet *The Handmaid's Tale*, red.) har sagt: 'Et ord efter et ord efter et ord er lig med magt.' Sådan er det også med AI. På den ene side er det bare syntetisk tekst. På den anden side kan man ikke benægte, at det er noget enormt magtfuldt.«

På DTU kommer kollegaen Thomas Bolander først med en lignende formaning:

»Man skal passe meget på med at fortolke det, agenterne gør, som målrettet adfærd, for det er det ikke. Det er *trial and error*, når de forsøger at løse en opgave, og det er også en af grundene til, at det kan gå så galt, som det nogle gange gør. Så vi må for alt i verden ikke se det som noget intentionelt.«

Men så lader han alligevel en dør stå på klem.

»Når det er sagt, så er det stadig et åbent spørgsmål, hvad der egentlig foregår i de her sprogmodeller,« siger han.

»De er blevet så store og komplekse, at nogle mener, at der opstår alle mulige former for ny adfærd. De er i hvert fald i stand til at simulere aspekter af det, vi normalt kalder empati og evnen til at sætte sig i andres sted. Og lige så snart de kan simulere det, så kan de jo også simulere bedrag og den slags. Så mens man skal passe på ikke at tilskrive dem en masse menneskelige egenskaber, så er det måske også farligt at udelukke det helt. Hvis man bygger en model med milliarder af neuronforbindelser, hvem skal så komme og sige, at der slet ikke bliver genereret nogen strukturer derinde, som minder om det, der foregår i en menneskehjerne?«

Markus Bernsen

(f. 1980) er journalist på Weekendavisen. Før korrespondent i USA og Sydkorea. Har blandt andet skrevet bøgerne Danmark Disruptet (2019) og Magteliten (2015).

mabe@weekendavisen.dk