Enden på Oplysningstiden

Af **Henry Kissinger**

Det menneskelige samfund er filosofisk, intellektuelt – på alle måder – uforberedt på fremkomsten af kunstig intelligens.

For tre år siden, på en konference om transatlantiske spørgsmål, dukkede emnet kunstig intelligens *(artificial intelligence* – AI) op på dagsordenen. Jeg var på nippet til at springe den session over – det lå uden for min sædvanlige interessesfære – men begyndelsen på præsentationen holdt mig i mit sæde.

Taleren beskrev et computerprogram, som snart ville kunne udfordre internationale mestre i spillet Go. Jeg var forbløffet over, at en computer kunne mestre Go, som er mere kompliceret end skak. Hver spiller indsætter 180 eller 181 brikker i spillet (afhængigt af hvilken farve han eller hun vælger), som placeres skiftevis på et oprindeligt tomt bræt; sejren tilfalder den spiller, der ved at træffe bedre strategiske beslutninger immobiliserer sin modstander ved at kontrollere territorium på mest effektiv vis.

Taleren insisterede på, at denne evne ikke kunne forprogrammeres. Hans maskine, sagde han, lærte at mestre Go ved at træne sig selv gennem øvelse. Efter at være blevet givet Gos grundlæggende regler, spillede computeren utallige spil mod sig selv, lærte af sine fejl og forfinede sine algoritmer i overensstemmelse hermed. I processen overhalede den sine menneskelige mentorers færdigheder. Og ganske rigtigt; i månederne efter talen besejrede et AI-program ved navn AlphaGo overlegent verdens bedste Go-spillere.

Mens jeg lyttede til taleren lovprise denne tekniske fremgang, gav min erfaring som historiker og lejlighedsvis praktiserende statsmand mig anledning til efter-

Henry A. Kissinger tjente som national sikkerhedsrådgiver og udenrigsminister for præsidenterne Richard Nixon og Gerald Ford.

tanke. Hvilken indflydelse vil selvlærende maskiner have på historien – maskiner, der erhverver viden ved processer, der er særlige for dem selv, og som anvender denne viden til formål, som ligger uden for menneskelig forståelse? Vil disse maskiner lære at kommunikere med hinanden? Hvordan vil valg blive truffet blandt de nye muligheder, der opstår? Er det muligt, at menneskets historie vil følge samme vej som inkaernes, da de stod over for en spansk kultur, som var uforståelig og endda ærefrygtindgydende for dem? Befinder vi os på kanten af en ny fase i menneskets historie?

Sandheden bliver relativ. Information truer med at overvælde visdom.

Da jeg var mig bevidst om min manglende tekniske kompetence på dette område, organiserede jeg en række uformelle dialoger om emnet, med rådgivning fra og i samarbejde med bekendte inden for teknologi og humaniora. Disse diskussioner har fået mine bekymringer til at vokse.

Hidtil har dét teknologiske fremskridt, der mest markant har ændret kursen på den moderne historie, været opfindelsen af trykpressen i det 15. århundrede, som gjorde det muligt for søgningen efter empirisk viden at erstatte liturgisk doktrin, og for Fornuftens Tidsalder gradvist at erstatte Religionens Tidsalder. Individuel indsigt og videnskabelig viden erstattede tro som det vigtigste kriterium for menneskelig bevidsthed. Oplysninger blev gemt og systematiseret i ekspanderende biblioteker. I Fornuftens Tidsalder opstod de tanker og handlinger, der formede den moderne verdensorden.

Men denne verdensorden er nu i opbrud midt i en ny og endnu mere vidtrækkende teknologisk revolution, hvis konsekvenser vi endnu ikke fuldt ud begriber, og hvis kulmination kan blive en verden, der er afhængig af maskiner, som er drevet af data og algoritmer, og som ikke er styret af etiske eller filosofiske normer.

Internetalderen, som vi allerede lever i, præger nogle af de spørgsmål og emner, som AI kun vil gøre mere akutte. Oplysningstiden forsøgte at underkaste traditionelle sandheder en frigjort, analytisk menneskelig fornuft. Internettets formål er at ratificere viden gennem akkumulering og manipulation af evigt ekspanderende data. Menneskelig erkendelse mister sin personlige karakter. Individer bliver til data, og data bliver dominerende.

Brugere af internettet lægger vægt på at hente og manipulere information frem for at kontekstualisere eller konceptualisere dennes betydning. De forhører sig sjældent hos historie eller filosofi; som regel kræver de oplysninger, der er relevante for deres umiddelbare praktiske behov. I processen erhverver søgemaskinealgoritmer kapaciteten til at forudsige individuelle klienters præferencer, hvil-

ket gør det muligt for algoritmerne at tilpasse resultater og gøre dem tilgængelige for andre parter, til politiske eller kommercielle formål. Sandheden bliver relativ. Information truer med at overvælde visdom.

Brugerne, som via sociale medier overvældes med holdninger fra masserne, afledes fra introspektion; sandheden er, at mange teknofile bruger internettet til at undgå den ensomhed, de frygter. Alle disse pres svækker den styrke, der kræves for at udvikle og opretholde overbevisninger, der kun kan implementeres gennem en ensom rejse, hvilket er essensen af kreativitet.

Internetteknologiens indvirkning på politik er særlig udtalt. Evnen til at gå målrettet efter mikrogrupper har opløst den tidligere konsensus om prioriteringer ved at tillade et fokus på specialiserede formål eller klagemål. Politiske ledere, som overvældes af pres fra nicher, fratages tid til at tænke eller reflektere over kontekst, hvilket mindsker det rum, de har til rådighed til at udvikle visioner.

En voksende procentdel af menneskelig aktivitet vil inden for en håndgribelig tidsperiode blive drevet af AI-algoritmer.

Den digitale verdens vægt på hastighed hæmmer refleksion; dens incitament styrker det radikale over det reflekterende; dens værdier er formet af undergruppekonsensus, ikke af introspektion. På trods af alt, den har opnået, risikerer den at blive sin egen fjende, i takt med at dens ubehageligheder overstiger dens bekvemmeligheder.

I takt med at internettet og øget computerkraft har gjort det lettere at akkumulere og analysere enorme mængder af data, er der opstået hidtil usete perspektiver for menneskelig forståelse. Mest betydningsfuldt er muligvis projektet med skabelse af kunstig intelligens – en teknologi, der er i stand til at opfinde og løse komplekse, tilsyneladende abstrakte problemer ved hjælp af processer, der ser ud til at replikere processerne i menneskehjernen.

Dette rækker langt ud over automatisering, som vi har kendt det. Automatisering beskæftiger sig med midler; det når de foreskrevne mål ved at rationalisere eller mekanisere instrumenter til at nå dem. AI derimod beskæftiger sig med mål; den fastlægger sine egne målsætninger. I det omfang, at dens resultater delvis er formet af sig selv, er AI i sagens natur ustabil. AI-systemer er gennem selve deres operationer under konstant forandring, når de tilegner sig og øjeblikkeligt analyserer nye data, og derefter forsøger at forbedre sig selv på basis af denne analyse. Gennem denne proces udvikler kunstig intelligens en evne, der tidligere blev troet forbeholdt mennesker. Den træffer strategiske vurderinger om fremtiden, nogle er baseret på data den modtager som kode (for eksempel reglerne for et

spil), og nogle er baseret på data, den selv indsamler (for eksempel ved at spille én million gentagelser af et spil).

Den førerløse bil illustrerer forskellen på traditionelle menneskestyrede, software-drevne computeres handlinger og det univers, som AI søger at navigere. At køre i en bil kræver dømmekraft i flere situationer, der er umulige at forudse, og dermed umulige at programmere på forhånd. Et velkendt hypotetisk eksempel er at spørge, hvad der ville ske, hvis en sådan bil kom i en situation, hvor den var tvunget til at vælge mellem at dræbe en bedsteforælder og dræbe et barn? Hvem ville den vælge? Hvorfor? Hvilke faktorer blandt dens muligheder ville den forsøge at optimere? Og ville den kunne forklare sine bevæggrunde? Hvis adspurgt, og hvis den var i stand til at kommunikere, ville dens sandfærdige svar sandsynligvis være: "Jeg ved det ikke (fordi jeg følger matematiske, ikke menneskelige, principper)," eller "I ville ikke forstå det (fordi jeg er blevet trænet til at handle på en bestemt måde, men ikke til at forklare det)." Alligevel vil førerløse biler sandsynligvis være udbredte på vejene inden for et årti.

AI-forskning har hidtil været begrænset til specifikke aktivitetsområder, men søger nu at skabe en "generelt intelligent" AI, der er i stand til at udføre opgaver indenfor flere områder. En voksende procentdel af menneskelig aktivitet vil inden for en håndgribelig tidsperiode blive drevet af AI-algoritmer. Men disse algoritmer, der er matematiske fortolkninger af observerede data, forklarer ikke den underliggende virkelighed, der producerer dem. Paradoksalt nok vil verden, i takt med at den bliver mere gennemsigtig, også blive stadig mere mystisk. Hvad vil adskille denne nye verden fra den, vi hidtil har kendt? Hvordan vil vi leve i den? Hvordan vil vi håndtere AI, forbedre den eller i det mindste forhindre den i at gøre skade, for ikke at tale om den mest ildevarslende bekymring: At AI, ved at mestre visse kompetencer hurtigere og mere definitivt end mennesker, over tid vil kunne reducere menneskelig kompetence og selve den menneskelige tilstand, i takt med at den forvandler dem til data.

Befinder vi os på kanten af en ny fase i menneskets historie?

Kunstig intelligens vil med tiden bringe ekstraordinære fordele til lægevidenskaben, levering af ren energi, miljøspørgsmål og mange andre områder. Men netop fordi AI træffer afgørelser om en ubestemt fremtid, der er under udvikling, er usikkerhed og tvetydighed iboende i dens resultater. Der er tre områder, der er særligt bekymrende:

For det første kan AI opnå utilsigtede resultater. Science fiction har forestillet sig scenarier, hvor AI vender sig imod dens skabere. Mere sandsynligt er det, at AI vil

fejlfortolke menneskelige instruktioner på grund af dens iboende mangel på kontekst. Et berømt nyligt eksempel var AI-chatbotten kaldet Tay, som var designet til at skabe venlig samtale i samme sprogmønstre som en 19-årig pige. Men det viste sig, at maskinen ikke var i stand til at definere de kommandoer om "venligt" og "rimeligt" sprog, som dets instruktører havde installeret, og den blev i stedet racistisk, sexistisk og generelt provokerende i sine svar. Nogle i teknologiverdenen hævder, at eksperimentet var dårligt udtænkt og dårligt udført, men det illustrerer en underliggende tvetydighed: I hvilket omfang er det muligt at gøre AI i stand til at forstå den kontekst, der ligger til grund for dens instruktioner? Hvilket medium kunne have hjulpet Tay med at udvikle en definition af, hvad der er stødende – en definition som mennesker ikke er enige om? Kan vi på et tidligt tidspunkt opdage og rette et AI-program, der handler uden for vores forventningsramme? Eller vil AI, hvis det overlades til sig selv, uundgåeligt udvikle små afvigelser, der over tid kan vokse sig katastrofale?

For det andet kan AI, gennem opnåelsen af tilsigtede mål, komme til at ændre menneskelige tankeprocesser og menneskelige værdier. AlphaGo besejrede verdensmestrene i Go ved at foretage hidtil ukendte strategiske træk – træk, som mennesker ikke havde udtænkt, og som de endnu ikke har lært at overvinde. Overgår disse træk den menneskelige hjernes formåen? Eller vil mennesker være i stand til at lære dem, nu hvor de er blevet demonstreret af en ny mester?

Inden AI begyndte at spille Go, havde spillet forskellige, mangeartede formål: En spiller søgte ikke kun at vinde, men også at lære nye strategier, der potentielt kan anvendes på andre af livets dimensioner. AI, derimod, kender kun et formål: at vinde. Den "lærer" ikke konceptuelt, men matematisk ved marginale justeringer af dens algoritmer. Så ved at lære at vinde Go gennem at spille det anderledes end mennesker gør, har AI ændret både spillets karakter og dets indflydelse. Karakteriserer dette snævre fokus på sejr al AI?

Andre AI-projekter arbejder på at modificere menneskelige tanker ved at udvikle enheder, der er i stand til at generere en række svar på menneskelige forespørgsler. Ud over faktuelle spørgsmål ("Hvad er temperaturen udenfor?"), rejser spørgsmål om virkelighedens natur eller meningen med livet dybere problemstillinger. Ønsker vi, at børn skal lære værdier gennem dialog med ubundne algoritmer? Bør vi beskytte privatlivets fred ved at begrænse AI's læring om dens spørgere? Hvordan opnår vi i så fald disse mål?

Hvis AI lærer eksponentielt hurtigere end mennesker, må vi forvente, at også den *trial-and-error*-proces, hvormed menneskelige beslutninger generelt træffes, vil accelerere eksponentielt: at AI vil begå fejl hurtigere og i større omfang end men-

nesker gør. Det er ikke sikkert, det er muligt at moderere disse fejl ved at inkludere forbehold i programmeringen, der kræver "etiske" eller "rimelige" resultater, som forskere indenfor AI ellers ofte foreslår. Hele akademiske discipliner er opstået ud af menneskehedens manglende evne til at blive enige om, hvordan man definerer disse begreber. Bør AI derfor blive deres mægler?

For det tredje, at AI kan nå tilsigtede mål, men være ude af stand til at forklare begrundelsen for dens konklusioner. På visse områder – mønstergenkendelse, *big-data*-analyse, gaming – overstiger AI's kapaciteter muligvis allerede menneskers. Hvis dens computerkraft fortsætter med at intensiveres hurtigt, vil AI muligvis snart være i stand til at optimere situationer på måder, der som minimum er marginalt anderledes, og sandsynligvis markant anderledes, fra hvordan mennesker ville optimere dem. Men vil AI til den tid være i stand til at forklare, hvorfor dens handlinger er optimale, på en måde, som mennesker kan forstå? Eller vil AI's beslutningstagning overstige menneskets sprog og fornufts forklarende kræfter? Gennem hele menneskehedens historie har civilisationer skabt måder at forklare verden omkring dem på – i middelalderen, religion; i oplysningstiden, fornuft; i det 19. århundrede, historie; i det 20. århundrede, ideologi. Det sværeste, men alligevel vigtigst spørgsmål om den verden, vi er på vej mod, er dette: Hvad bliver der af menneskelig bevidsthed, hvis dens egen forklarende magt overgås af AI, og samfund ikke længere er i stand til at fortolke den verden, de bebor, gennem begreber, der er meningsfulde for dem?

Hvordan definerer man bevidsthed i en verden af maskiner, der reducerer menneskelige oplevelser til matematiske data, fortolket af deres egne erindringer? Hvem er ansvarlig for AI's handlinger? Hvordan placerer man ansvaret for deres fejl? Kan et retssystem designet af mennesker holde trit med aktiviteter produceret af en AI, der er mere intelligent og potentielt i stand til at udmanøvrere dem?

I sidste ende kan udtrykket 'kunstig intelligens' vise sig at være en misvisende betegnelse. Disse maskiner kan uden tvivl løse komplekse, tilsyneladende abstrakte problemer, der tidligere kun kunne løses gennem menneskelig tænkning. Men det, de gør unikt, er ikke at tænke i den forstand, vi hidtil har opfattet og oplevet tænkning. Det er snarere en hidtil uset evne til udenadslære og udregning. På grund af dens iboende overlegenhed på disse områder vil AI sandsynligvis vinde ethvert spil, det bliver stillet overfor. Men til vores formål som mennesker handler spillene ikke kun om at vinde; de handler om at tænke. Ved at behandle en matematisk proces, som om det var en tankeproces, og enten forsøge at efterligne selve denne proces eller blot acceptere resultaterne, risikerer vi at miste den kapacitet, der har været essensen af menneskelig tænkning.

Paradoksalt nok vil verden, i takt med at den bliver mere gennemsigtig, også blive stadig mere mystisk.

Implikationerne af denne udvikling vises af et nylig designet program, AlphaZero, der spiller skak på et niveau, der er overlegent i forhold til skakmestre, og i en stil, der ikke tidligere er set i skakhistorien. På egen hånd opnåede det efter kun få timers selvspil et færdighedsniveau, som det tog mennesker 1.500 år at nå. AlphaZero blev kun udstyret med de grundlæggende regler for spillet. Hverken mennesker eller menneskegenererede data var en del af dets proces med selvlæring. Hvis AlphaZero var i stand til at opnå denne mestring så hurtigt, hvor vil AI så være om fem år? Hvilken effekt vil det generelt have på menneskelig tænkning? Hvad er etikkens rolle i denne proces, der i essensen består af en accelerering af valg , som skal træffes?

Disse spørgsmål overlades typisk til teknologer og til intelligentsiaen inden for beslægtede videnskabelige områder. Filosoffer og andre inden for humaniora, der var med til at forme tidligere koncepter om verdensorden, har det med at halte bagud, da de mangler viden om AI's mekanismer eller bliver benovede over dens kapaciteter. Derimod anspores den videnskabelige verden til at udforske de tekniske muligheder af dens resultater, og den teknologiske verden er optaget af kommercielle udsigter i et fabelagtigt omfang. Begge disse verdeners incitament er at skubbe til opdagelsers grænser snarere end at forstå dem. Og regeringer, i det omfang de behandler emnet, er mere tilbøjelige til at undersøge AI's anvendelsesmuligheder indenfor sikkerhed og efterretning, end til at undersøge den transformering af den menneskelige tilstand, som den er begyndt at producere.

Oplysningstiden begyndte med essentielt filosofiske indsigter, der blev spredt ved hjælp af en ny teknologi. Vores periode bevæger sig i den modsatte retning. Den har genereret en potentielt dominerende teknologi på jagt efter en vejledende filosofi. Andre lande har gjort AI til et stort nationalt projekt. USA har endnu ikke som nation systematisk udforsket dens fulde omfang, undersøgt dens implikationer eller påbegyndt processen med ultimativ læring. Dette bør gives høj national prioritet, især hvad angår at relatere AI til humanistiske traditioner.

AI-udviklere, som er lige så uerfarne indenfor politik og filosofi, som jeg er inden for teknologi, burde stille sig selv nogle af de spørgsmål, jeg har rejst her, for dermed at bygge svar ind i deres tekniske indsatser. Den amerikanske regering bør overveje at nedsætte en præsidentiel kommission bestående af fremtrædende tænkere for at hjælpe med at udvikle en national vision. Så meget står klart: Hvis vi ikke påbegynder denne indsats snart, vil vi inden længe opdage, at vi begyndte for sent.

Artiklen blev først bragt i The Atlantic. Oversat af Ane Dalegaard Hansen.