AI är mänsklighetens största ödesfråga

Föredrag vid Rausingsymposiet i Lund den 13 maj 2025

Olle Häggström

Jag har stor respekt för att det finns andra frågor än AI – exempelvis klimatförändringarna – som kan göra anspråk på att vara stora ödesfrågor för mänskligheten, men de senaste åren har jag kommit att landa i att den om hur vi hanterar AI-utvecklingen är den största och mest akuta av dem alla. Jag skall förklara varför, men vill börja i en mer jordnära ände.

Kan AI bli bättre än läkare på medicinsk diagnosticering? Mycket har sagts och skrivits om detta ämne, och vi skall höra en hel del om saken i Kristina Längs föredrag i eftermiddag, men här vill jag som exempel framhålla två aktuella studier i tidskriften Nature som pekar mot att svaret kan vara ja. De är begränsade till en ganska snäv kontext med enbart diagnosticering via textinterface, men det är ändå intressant att det AI-system som studeras presterar bättre än mänskliga läkare ställda inför samma uppgift. Studiernas begränsningar gör att det är på tok för tidigt att hävda att läkarna redan idag skulle vara akterseglade av AI, men utvecklingen går snabbt och vi kan komma att nå den dag då klockan klämtar för läkarna snabbare än vi anat.

I frågor om AI och automatisering av beslutsfattande talar man ofta om det så kallade human in the loop-idealet. Detta går ut på att AI inte skall *ersätta* utan blott *assistera* mänskligt beslutsfattande, och att människan alltid skall ha sista ordet. Extra näring för detta ideal fås ur idén om att kombinationen människa plus AI i många fall presterat bättre än AI ensam även i fall där AI överglänser människan ensam.

Det paradexempel på detta fenomen som oftast anförts av AI-entusiaster kommer från schacket. Kraftmätningen mellan människa och maskin på schackets område har ansetts avgjord ända sedan 1997, då dåvarande schackvärldsmästaren Garri Kasparov förlorade en uppmärksammad match mot IBM:s schackprogram Deep Blue. Redan året efter tog han initiativ till den nya spelformen kentaurschack, där mänskliga spelare fick lov att konsultera ett schackprogram innan de utförde sina drag. Så länge teamet människa plus maskin var bättre än maskinen ensam förblev spelformen intressant, men intresset kollapsade i slutet av 2010-talet då det blev snudd på omöjligt att överträffa en motståndare som valde att slaviskt följa datorns rekommendationer, och det mänskliga bidraget till spelformen därför blivit irrelevant.

Ett liknande fenomen påvisas i den ena av de nämnda Nature-studierna. Läkare med tillgång till forskarnas AI-system uppvisade *sämre* precision i sina diagnoser än AI-systemet ensamt. Under sådana omständigheter, där det mänskliga deltagandet aktivt försämrar utfallet, är det svårare att försvara människan i loopen. Vi är som sagt inte där än, men om och när det inträffar gissar jag att vi ändå in i det längsta kommer att insistera på att viktiga vårdbeslut skall fattas av mänskliga läkare. Deras situation blir då föga avundsvärd, då de sitter med det formella ansvaret för ett beslut som de inte vill gripa in i eftersom de inser att det vore kontraproduktivt, och som de till följd av AI-systemens black box-egenskap kanske inte ens begriper grunderna för.

Motsvarande utveckling kan vara på väg att ske i snart sagt alla samhällssektorer. Den dynamik som därmed uppstår har studerats av forskare under benämningen *gradual disempowerment* (gradvis förlust av makt). Effektivitetskrav leder allt fler aktörer att plocka bort människan ur loopen, och de företag som av exempelvis idealistiska skäl undviker det slås ut i

konkurrensen. Den logiska slutpunkten för en sådan utveckling är att människan inte längre deltar i styrningen i samhället, utan den saken helt tagits över av AI, och i det läget kan vi inte längre påverka framtiden. Vi behöver därför snabbt få till stånd en välinformerad diskussion om vart vi är på väg och vilket slags samhälle med AI vi bör eftersträva – innan det är för sent.

En typ av arbetsuppgift där AI gjort särskilt dramatiska framsteg de senaste åren och som kan komma att få stor betydelse för den fortsatta utvecklingen som helhet är kodning och mjukvaruutveckling. AI-systemens förmåga att skriva korrekt kod och i övrigt lösa uppgifter riktigt är starkt avhängig uppgiftens omfattning. I en rapport från AI-säkerhetsorganisationen METR i mars i år studeras hur denna förmåga utvecklats över tid. Det visar sig att omfattningen – mätt i tidsåtgång för en mänsklig expert – som AI klarar av har ökat från enstaka sekunder 2019 till cirka en timme idag. Ökningen är exponentiell, med en observerad genomsnittlig fördubblingstid på sju månader, och om man extrapolerar den trenden blott ett par-tre år in i framtiden blir resultatet dramatiskt. Sådan kurvanpassning inbegriper givetvis stora osäkerheter, men ser man till hur modellerna förbättrats från 2024 och framåt verkar det snarast som att utvecklingen är på väg att gå ännu fortare.

Det är bland annat den sortens data som ligger till grund för den gedigna rapporten *AI 2027*, utkommen i april i år och författad av en kvintett forskare med den avhoppade OpenAI-medarbetaren Daniel Kokotajlo i spetsen. Rapporten är det ambitiösaste och bästa som hittills skrivits vad gäller detaljerade förutsägelser av kommande års AI-utveckling. Osäkerheterna är som sagt stora, men successivt och månad för månad arbetar de fram vad de ser som det mest sannolika förloppet. Centralt i detta förlopp är hur AI, till följd av den utveckling som bland annat METR-rapporten påvisat, år 2027 når en punkt där den är en lika skicklig AI-utvecklare som dagens främsta sådana av kött och blod. Tack vare att de ledande AI-företagen då kan sätta hundratusentals eller miljontals sådana AI i arbete leder detta på några få månader till så kallad superintelligens – AI som vida överträffar människan över hela spektret av relevanta förmågor.

Vad som därefter kan förväntas hända är långt ifrån självklart, och i själva verket har förloppet i Kokotajlos rapport en bifurkation i slutet av 2027 där det delar sig i i två, varav det ena landar i utplåning av mänskligheten redan 2030, och det andra går betydligt bättre. Mänsklighetens undergång kan låta överdrivet dramatiskt, men vad superintelligens i praktiken innebär är ju att vi har skapat en ny art som petar ned oss till en andraplats bland planetens mest kapabla och intelligenta arter, och vi bör inte ta för givet att en sådan situation automatiskt slutar lyckligt för den art som akterseglats.

Här kan det vara värt att gå tillbaka och höra vad några tidigare tänkare på området sagt. Datavetenskapens fader Alan Turing gick som bekant bort under förfärliga omständigheter redan 1954, ännu inte 42 år fyllda och två år innan termen artificiell intelligens skulle uppstå som namnet på ett forskningsområde. Merparten av hans livsgärning var av matematisk och teknisk natur, men de sista åren av sitt liv tillät han sig enstaka mer filosofiska utblickar och reflektioner kring framtiden för den teknologiutveckling han så starkt bidragit till att sätta igång. I en ofta citerad passage från 1951 spekulerar han över vad som kan hända om vi lyckas skapa maskiner som kan tänka på ett sätt som liknar människors. Kokotajlos och andras mer sentida idéer på samma tema kan delvis spåras tillbaka till hur Turing här ser framför sig att maskinerna vid en viss nivå kan börja förbättra sig själva utan behov av vidare mänsklig inblandning, varvid utvecklingen kan väntas accelerera kraftigt, så att ”vi så småningom bör vänta oss att maskinerna tar över kontrollen”.

Vilket inflytande hade då dessa ödesmättade ord på den AI-forskning som snart skulle komma igång? Att döma av det ambitiösa och oförvägna arbete som AI-forskarna kom att bedriva, vilket under det halvsekel som följde efter att Turing formulerade sina rader var nästan helt befriat från risktänkande, förblev hans påverkan på denna punkt länge nära nollnivån. Ett av mycket få undantag som under denna tidsperiod tog riskfrågan på fullt allvar var Norbert Wieners remarkabelt framsynta text *Some moral and technical consequences of automation* från 1960, där författaren resonerar kring en tänkt framtida situation där vi skapar en maskin så kraftfull att vi när vi väl satt igång den saknar förmåga att stänga av den. Han betonar vikten av att säkerställa att ”det syfte vi byggt in i maskinen är det syfte vi eftersträvar snarare än någon färgstark imitation”. Dessa ord skulle faktiskt kunna fungera väl som programförklaring för det område som inte uppstod förrän först många årtionden senare men som idag benämns *AI alignment*, vars avsikt är att se till att de första riktigt kraftfulla AI-systemen har mål och värderingar som i tillräcklig mån prioriterar mänsklig välfärd och blomstring, och mer allmänt är i linje med mänskliga värderingar.

AI alignment-pionjären Eliezer Yudkowsky kan med blott en mild överdrift sägas egenhändigt ha lagt grunden för området under 00-talet. I en inflytelserik artikel från 2008 beskriver han det han bedömer vara default-scenariot ifall vi misslyckas med eller helt enkelt ignorerar AI alignment: ”AI:n hatar dig inte, ej heller älskar den dig, men du består av atomer som den kan ha annan användning för”.

I en sådan situation vill vi givetvis inte hamna, och därför behöver vi lösa AI alignment i tid. Hur lång tid har vi då på oss? Ingen vet säkert, och det enda omdömesgilla är att medge att stor osäkerhet föreligger, men jag menar att vi bör ta på allvar den i AK-ekosystemet i San Francisco och Silicon Valley alltmer utbredda uppfattningen att superintelligent AI kan bli en realitet inom en tidsrymd som mäts i enstaka år snarare än decennier.

Kokotajlo-rapporten är ett uttryck för denna uppfattning, och den tidigare styrelseledamoten i OpenAI Helen Toner vittnade i ett senatsförhör i Washington den 17 september 2024 om hur vanlig den är bland insiders på de ledande AI-företagen. Hon tillade att ”många av dessa bedömare tror att om de lyckas att bygga datorer som är lika smarta som människor eller eventuellt långt smartare, så kommer tekniken att som minst bli extremt omstörtande, och som mest leda till bokstavlig utrotning av mänskligheten”.

Det finns såklart andra som ställer sig avvisande till sådana bedömningar, ibland med hänvisning till att dagens AI inte besitter något som överhuvudtaget förtjänar att kallas intelligens, och jag vet att flera av er här i lokalen har denna uppfattning. Argumenten varierar, men jag har ännu inte stött på något som håller för närmare granskning. Två särskilt vanliga argument i samband med den kanske viktigaste AI-tekniken i dagens frontlinje – deep learning-drivna stora språkmodeller – är för det första ”allt de gör är att prediktera nästa ord i internettext”, och för det andra ”allt som händer innerst inne i de neurala nätverken är den triviala matematiska operationen matrismultiplikation”. Båda två är exempel på det argumentationsfel som kan benämnas *reductio ad reductem*, och som innebär att man förnekar möjligheten till emergens av komplexa fenomen ur sammansättningen av enkla, eller med andra ord att man ur observationen att helheten kan sönderdelas i enkla beståndsdelar drar slutsatsen att helheten inte finns.

Varför hävdar jag att det är ett argumentationsfel? Det enklaste sättet att inse det är att tillämpa argumentet på någon erkänt intelligent person som exempelvis Albert Einstein, och konstatera att Einsteins hjärna var uppbygd av atomer och elementarpartiklar som var och en är stendum,

varför sagda hjärna omöjligt kan ha varit ett säte för intelligens. Eftersom vi ju var överens om att Einstein faktiskt var intelligent kan vi därmed konstatera att *reductio ad reductem* inte är någon giltig slutledningsprincip, varvid argumenten om att ordprediktion och matrismultiplikation inte räcker till som beståndsdelar i intelligenta system kollapsar. I en artikel publicerad 2023 går jag igenom en rad andra försök att visa att stora språkmodeller inte besitter resonemangsförmåga eller intelligens. Jag destruerar dem alla med samma grepp, nämligen att påvisa att om argumenten vore riktiga skulle de även medföra icke-existensen av mänsklig intelligens – en slutsats vi vet är falsk, varför vi kan konstatera att argumentet inte håller.

Den kanske allra vanligaste typen av argument för de stora språkmodellernas icke-intelligens är att peka på fall där de svarar korkat på någon fråga eller på annat vis misslyckas med någon uppgift som borde vara enkel, och hävda att ingen verkligt intelligent entitet skulle göra bort sig så kapitalt. Därmed, påstås det, är vi uppenbarligen långt från det som brukar kallas AGI – artificiell generell intelligens – vilket brukar definieras som en AI som presterar på mänsklig nivå eller högre inom hela spektret av relevanta kognitiva förmågor. Men med det argumentet har vi lagt ribban för högt – så högt, vågar jag påstå, att ingen här i lokalen kan göra anspråk på intelligens, för upp med en hand, alla som aldrig någonsin sagt något dumt eller misslyckats med något enkelt!

Det fanns en tid när AGI-begreppet var ett viktigt pedagogiskt verktyg, men jag tror att begreppet har överanvänts på senare år (något jag själv erkänner mig medskyldig till) och att vi idag nått en punkt där det bidrar mindre till klarhet än till förvirring, genom dess betoning på att AI skall klara *allt* som människan kan. Det uppmuntrar till det felslut jag här diskuterat om att AI kan avfärdas som meningslös så snart vi ser att den misslyckas med någon enskild enkel uppgift. Dessutom leder för de flesta praktiska frågeställningar en överbetoning av AGI till fel fokus. Tag till exempel den fråga jag här intresserar mig för mest av allt: den om huruvida och i så fall när AI kan bli kapabel att utmana människan om herravälde över samhället och vår planet. Det viktiga för denna frågeställning är såklart inte när AI överträffar oss i *allt*, utan när den blir tillräckligt mycket bättre på ett tillräckligt brett spektrum av kompetenser som är relevanta för maktövertagande. Den språkliga kompetens som mycket av dagens AI-utveckling kretsar kring tror jag kan vara viktig här, med tanke på vilken stor andel av all maktutövning som i dagens samhälle äger rum via språkhandlingar. Måhända ännu viktigare kan visa sig bli den mjukvaruutvecklingskompetens jag varit inne på ovan. Däremot tror jag inte alls att det är nödvändigt för AI att kunna knyta ett skosnöre innan den är redo att ta över världen.

Men som antytts tror jag att det totala AI-maktövertagandet är något vi bör undvika, och vi bör åtminstone se till att vi har löst AI alignment innan vi skapar superintelligens. Just nu ser det besvärligt ut med tanke på den rasande utvecklingskurva AI befinner sig på samtidigt som problemet med AI alignment gäckar oss och ingen har någon övertygande plan för hur det skall lösas.

För att ge AI alignment-forskningen en chans att hinna ikapp tror jag att vi behöver dra i nödbromsen för utvecklingen av de allra mest kraftfulla AI-systemen. Detta försvåras dock av den kapplöpningssituation som föreligger, både mellan enskilda AI-företag och mellan länder (främst USA och Kina). Det allmänt hårdnande internationella klimatet sedan Trumps andra presidentämbetestillträde gör inte heller saken lättare. Ett lågvattenmärke för den globala AI-diskursen nåddes vid toppmötet AI Action Summit i Paris i februari i år, där säkerhetsfrågor sopades under mattan samtidigt som toppolitiker bjöd över varandra i vilka mångmiljardbelopp de avsåg satsa på AI-utveckling. Värst av allt var hur den amerikanske vicepresidenten JD Vance i sitt anförande uttryckte oförblommerat förakt för AI-säkerhet, då han slog fast att han ”inte var

där för att tala om AI-säkerhet” och att ”vår AI-framtid inte erövras genom att oja sig över säkerhet utan genom att bygga”. Hans förhoppning lite längre fram i samma tal om att ”AI-ekonomin kommer att […] transformera den värld som består av atomer” ger, för den som minns Yudkowskys ovan citerade oneliner om AI och atomer, en isande rysning längs ryggraden.

Detta säger något om vilka krafter vi behöver övervinna om vi skall få ordning på AI-utvecklingen och styra den i för mänskligheten mer gynnsam riktning jämfört med vart vi idag verkar vara på väg. Men framtiden är inte ristad i sten, och jag tror fortfarande att det är möjligt att förhindra en AI-katastrof. Något som skulle förbättra oddsen ytterligare vore om vi lyckas mobilisera den folkopinion mot skapandet av övermänskligt intelligent AI som enligt diverse opinionsundersökningar verkar föreligga. Så hjälp gärna till att sprida budskapet!