# Varför AI inte kommer att ta över världen

28-02-2024

Text: Peter Gärdenfors

Artificiell intelligens har på senare tid haft spektakulära framgångar. Kapaciteten hos språkprogram som ChatGPT, Bing AI och Bard och hos bildskapande program som Midjourney och DALL-E 2 har överraskat många. Även inom andra områden har det skett genombrott, exempelvis när det gäller att beskriva den tredimensionella strukturen hos proteiner, vilket är ett svårt problem för forskare inom biomedicin.

De snabba framgångarna för AI har lett till övertro på vad som är möjligt för AI-system att uppnå. Många AI-forskare, bland dem svenskarna Nick Boström, Max Tegmark och Olle Häggström, hävdar att AI snart kommer att utvecklas till AGI – artificiell *generell* intelligens. Ett sådant system beskrivs som att det har alla intellektuella förmågor som människor har och mer därtill. En del forskare hävdar att det finns en fara för att AGI kommer att ta över världen. De uppfattar utvecklingen av AGI som ett ingenjörsproblem och ser inga principiella hinder. Frågan är om det finns tillräckliga argument för denna åsikt.

Centralt är hur man skulle kunna avgöra om ett AI-system verkligen har generell intelligens. Att ett datorprogram är bättre än en människa inom ett specialområde – som att spela schack eller att känna igen ansikten – säger väldigt lite om generell intelligens. När det gäller människor används ofta IQ som mått – mest därför att det inte finns något bättre. Men detta mått fungerar inte för maskiner. Det skulle vara relativt lätt att konstruera ett program som får full pott på de intelligenstest som används – inte för att programmet skulle vara särskilt intelligent, utan för att testen följer begränsade matematiska, språkliga och visuella mönster. Och den ordkunskap som programmet behöver kan lätt hämtas från internet.

Märkligt nog är diskussionerna om hur man mäter AGI ganska ytliga bland AI-forskarna. Boström ger tre förslag till kriterier i boken *Superintelligens* (Fri Tanke 2017). Det starkaste kravet är vad han kallar kvalitativ superintelligens, som han definierar som ”ett system som är åtminstone lika snabbt som ett mänskligt medvetande och kvalitativt långt smartare”. Tyvärr säger detta inte särskilt mycket eftersom man måste veta vad ett mänskligt medvetande förmår, och ordet ”smartare” gör att det nästan blir en cirkeldefinition.

Häggström är mer precis. I boken *Tänkande maskiner* (Fri Tanke 2021) definierar han AGI som ett system som har ”alla de förmågor som ligger till grund för mänsklig intelligens: kort- och långtidsminne, logiskt tänkande, matematisk förmåga, geometrisk och spatial visualisering, mönsterigenkänning, induktion, planering, kreativitet, social manipulation och många andra”. Om systemets intelligens överstiger den mänskliga sägs systemet vara superintelligent. En sådan beskrivning ger ett bättre verktyg för att bedöma maskiners tankeförmåga.

En definition som kan tillämpas på såväl människor och djur som maskiner kommer från den tyske psykologen William Stern: ”Intelligens är en individs allmänna förmåga att medvetet anpassa sitt tänkande till nya krav; det är en allmän mental anpassningsförmåga till nya problem

Förutom att kunna avgöra huruvida ett AI-program är superintelligent, är ett centralt problem hur man ska *konstruera* ett sådant program. Ingen av forskarna inom området har några klara idéer om hur detta ska gå till. De antyder ofta att när väl ett program uppnår en viss nivå så kommer det att utveckla sig självt till allt mer avancerad generell intelligens.

Det finns emellertid goda skäl att tro att AGI inte kommer att bli så omvälvande. En diger samling argument för detta presenteras i den nya boken *Why machines will never rule the world*, skriven av AI-forskaren Jobst Landgrebe och filosofen Barry Smith. Deras huvudargument kan sammanfattas så här: mänsklig intelligens uppstår i ett mycket komplext system som består av hjärnans interaktion med kroppen och av kroppens interaktion med andra individer och den omgivande världen. System som är så avancerade kan inte fångas i matematiska modeller. Därför kommer de aldrig att kunna reproduceras i AI-system. Detta argument underbyggs i bokens tre delar.

I första delen går författarna igenom några av det mänskliga tänkandets egenskaper. Framför allt lyfter de fram språkets komplexitet. ChatGPT och liknande system fungerar för texter som skrivs på datorskärmar. Men människans språk handlar främst om dialoger. Ett sådant samspel är starkt beroende av kontexten: temat för diskussionen (som kan ändra sig under vägen), de talandes avsikter med dialogen, deras förväntningar på den andra, deras minne av tidigare interaktioner, omgivningen och så vidare. De språkprogram som finns kan inte hantera sådana faktorer. De har exempelvis inga avsikter med sin språkproduktion. Landgrebe och Smith argumenterar för att mänskliga dialoger är så varierande och situationsberoende att det är omöjligt att samla tillräckligt med data för att ett AI-system ska kunna lära sig att hantera dem.

Ett annat område där AI-systemen slår slint är den mänskliga inlevelseförmågan. Vi kan, i det närmaste automatiskt, tolka andras känslor, avsikter, värderingar och kunskap. Att exempelvis förstå att någon är ironisk betyder att man förstår att den som uttalar sig ironiskt inte menar vad hen säger. Det är oerhört svårt för ett AI-system att fånga de subtila signaler som leder till att man tolkar ett yttrande som ironiskt.

Genom så kallad *affective computing* försöker man få AI-system att förstå människors känslor. Som input använder man människornas språk, ansiktsuttryck och kroppsspråk. Än så länge har forskarna inte kommit särskilt långt inom detta område. Man vill också att AI-systemens värderingar ska överensstämma med människornas. Även detta är besvärligt eftersom det inte är klart vad det innebär att ett system har en värdering.

I den andra, mest tekniska delen av boken argumenterar Landgrebe och Smith för att det övervägande antalet naturliga system är så komplexa att det inte går att skapa modeller för dem. Därmed kan de inte hanteras av några datorsystem. Även de enklaste formerna av liv är så komplicerade att de inte går att simulera med en dator. En enda biologisk cell innehåller runt hundra miljarder atomer som bildar hundra tusen olika RNA-molekyler. Levande system är också självorganiserande, och de upprätthåller sig själva genom att hämta energi från omgivningen. Djurens nervsystem, kanske framför allt människans, är de mest avancerade biologiska system som finns. De modeller av neuroner som används i AI-system är radikala förenklingar av riktiga biologiska celler.

När man jämför den mänskliga intelligensen med AI är ett vanligt argument att jämföra antalet neuroner som finns i den mänskliga hjärnan med antalet artificiella neuroner som AI-systemen använder sig av. Som stöd för systemens intelligens har man exempelvis påpekat att AI-systemet GPT-4 har motsvarigheten till hundra miljarder neuroner, medan den mänskliga hjärnan har åttio–hundra miljarder neuroner. Jämförelsen håller inte, eftersom hjärnan består av så mycket mer än bara neuroner. Signalsubstans er som dopamin, adrenalin och oxytocin spelar en stor roll för hjärnans processer, och dessa har inga motsvarigheter i AI-systemen. En ny teori hävdar dessutom att de magnetfält som uppstår genom de elektriska strömmarna i neuronerna också påverkar processer över hela hjärnan. Ett sådant fenomen går inte att fånga i de artificiella neuronnätverk som AI-systemen använder.

Hjärnan är inte en maskin. Även om vi skulle kunna mäta hjärnans molekylära egenskaper

alla påståenden om att man skulle kunna ”ladda upp” en människas hjärna till en dator.

Ett tekniskt begrepp som är centralt för Landgrebe och Smiths argumentation är *ergodicitet*. Lite förenklat är ett system ergodiskt om de data man kan samla in om systemet i det långa loppet blir representativa för systemets beteende. Landgrebe och Smith argumenterar för att datorer (och alla andra system som utför beräkningar) bara kan modellera ergodiska system. De hävdar också att de flesta naturliga system inte är ergodiska. Det vill säga att hur länge vi än studerar ett sådant system, kommer vi aldrig att kunna förutsäga dess beteende. Tomas Tranströmer har en träffande metafor för detta: ”Varje abstrakt bild av världen är lika omöjlig som ritningen till en storm.” De artificiella neuronnät som används inom så kallad *deep learning* bygger på statistiska mönster, och de kan därmed inte hantera situationer som faller utanför de ramar som ges av träningsdata.

Landgrebe och Smith har förmodligen rätt i att de flesta naturliga system inte är ergodiska, men de kan inte bevisa detta. Det är tänkbart att beteendet hos enkla biologiska system, exempelvis insekter, låter sig beskrivas ergodiskt, även om de enskilda cellerna är icke-ergodiska. Beteendet på makronivån kan kanske beskrivas uttömmande även om mikronivån är oöverskådlig.

I den tredje delen av boken beskriver de AI-systemens begränsningar inom flera områden och argumenterar för att det inte finns något som kommer i närheten av AGI. I synnerhet gäller detta mänskligt språk och mänskligt beteende. ChatGPT och andra språkmodeller hittar komplicerade mönster i sekvenser av ord, men de förstår inte ordens mening. Systemen härmar språkmönstren, men de tolkar dem inte. Dessutom blir svaren alltmer platta om man fortsätter att chatta med dem, eftersom systemet inte kan hålla reda på dialogens kontext.

En annan begränsning är att språkmodellerna är textbaserade. Systemen kan inte se den de talar med. Mänsklig kommunikation bygger på så mycket mer än text: en dialog påverkas också av tonfall, icke-språkliga ljud, blickar, ansiktsmimik, gester och så vidare.

Lika svårt blir det att konstruera system som uppvisar något som liknar människans inlevelseförmåga. Vi kan inte bygga maskiner som har avsikter och vilja, eftersom vi vet för lite om hur de uppstår hos människor. Utan att ge argument postulerar flera AI-forskare att detta är möjligt.

Även om Landgrebe och Smith anser att AGI aldrig kommer att uppnås, är de inte motståndare till AI. De ger exempel inom flera områden på hur AI kan komma att utvecklas. Ett område de lyfter fram är sjukdomsdiagnoser. Men eftersom människokroppen inte är ett ergodiskt system måste AI-systemens svar alltid tolkas av läkare för att täcka de fall som algoritmerna inte kan fånga. Ett obehagligare område, som säkerligen kommer att växa, är militära system. Sådana tillämpningar är extra farliga, eftersom de inte uppvisar AGI och alltså inte kan anpassa sig till nya situationer.

Följ oss på sociala medier

INSTAGRAM FACEBOOK

INFO@FRITANKE.SE

ANNONSERA I SANS

JOBBA HOS OSS

PRESS

VANLIGA FRÅGOR

INTEGRITETSPOLICY

IN ENGLISH