UTKAST. SPRID ELLER CITERA EJ UTAN SKRIFTLIGT GODKÄNNANDE FRÅN FÖRFATTAREN.
MEN KOMMENTERA GÄRNA!

Några teoretiska skäl till varför jag inte tror på snabba framsteg inom AI

Nils Dahlbäck
www.ida.liu.se/~nilda08

Inledning

Jag skall här försöka kort förklara varför jag anser att de långtgående profetior om hur AI kommer att kunna utvecklas i framtiden som flyter omkring i debatten idag är helt orealistiska i ljuset av väletablerad kunskap om datorer och om mänsklig kognition.

Det finns två huvudpelare i argumentet. Det ena gäller vad en dator är och vad som kan beräknas med en dator. Här finns det bland annat precisa matematiska beskrivningar av vad man kan och inte kan göra med en dator men som AI-entusiasterna vad jag kan se aldrig beaktar. Det andra gäller begränsningar i våra kunskaper idag om hur människor löser problem som krävs för t.ex. bilkörning, kreativt tänkande och liknande. Här menar jag att det finns stora luckor i vår kunskap om hur vi människor fungerar, vilket också sätter gränser för vad vi kan göra för att ge datorprogram människoliknande kompetens.

Texten är skriven för en person med grundläggande teknisk och matematisk kompetens. Den består till stor del av korta beskrivningar och länkar till populärvetenskapliga sidor där fördjupningar finns för den som vill gå vidare. För någon utan tidigare kunskaper inom området kan beskrivningarna möjligen framstå som avancerade, men det jag skriver om nedan är sådant som man lär sig på en grundkurs i AI och en grundkurs i kognitionsvetenskap på universitetet. Vilket väl rimligen är en miniminivå av kunskap för att kunna ta ställning i vilken som helst vetenskaplig eller teknisk fråga?

Vad är en dator och vad kan den beräkna?

En dator kan betraktas som en Turingmaskin
https://sv.wikipedia.org/wiki/Turingmaskin
(Bortsett från att en dator inte har ett oändligt minne.)

Enligt Church-Turings tes kan en Turingmaskin beräkna alla beräkningar som kan beräknas. Tesen är ej matematiskt bevisad, men accepteras som sann inom teoretisk datavetenskap.
https://sv.wikipedia.org/wiki/Church-Turings_hypotes

Här gör många en felslutledning och tolkar det som att en dator kan lösa *alla* problem. Men det är *inte* vad tesen säger. Den säger att *om* det existerar en algoritm för att utföra en beräkning *så* kan den utföras av en turingmaskin (dvs. av en dator).

Detta kräver i sin tur att två villkor är uppfyllda

1. Det existerar en formellt precis beskrivning av problemet som skall beräknas.
2. Det existerar en algoritm som kan beräkna en lösning av problemet.

Här finns ett antal viktiga och intressanta delargument som jag uppfattar som centrala för en förståelse för datorers möjligheter och begränsningar, men som vad jag kan se sällan eller aldrig betraktas av de som uttalar sig om vad AI kommer att kunna klara av i framtiden. Bland annat finns det problem som är precist beskrivna men som inte kan beräknas. Och detta kan man logiskt-matematiskt bevisa! Ett exempel på detta är det s.k. stopproblemet (The Halting Problem https://en.wikipedia.org/wiki/Halting_problem)

Det finns dessutom en klass av problemdomäner där man utveckla en användbar algoritmisk lösning för små problem, men där de nödvändiga resurserna i form av datorkapacitet (t.ex. tid att lösa problemet) ökar exponentiellt när problemstorleken ökar linjärt! Ett välkänt exempel är det så kallade Handelsresandeproblemet, där det gäller att finna en metod för att beräkna den kortaste resvägen för en person som skall besöka ett antal städer och bara besöka varje stad en gång. För bara några få städer går det lätt att jämföra alla möjliga resvägar, men redan för 10 blir det nästan övermäktigt. Realistiska problemstorlekar kräver orimliga lösningstider. I själva verket är tiden för att lösa detta för N städer är(N-1)![1]

Detta är ett exempel av många på praktiskt intressanta problem där man bara känner till algoritmer med en tidskomplexitetsfunktion som växer så snabbt att de är helt ointressanta i praktiken. Snabbare datorer hjälper inte. Dessa problem är inte *hanterbara* (*eng.* tractable). Hittade ingen bra text på svenska wikipedia så länken är till engelska Wkipedia där det finns en bra beskrivning av ”Intractability”.
https://en.wikipedia.org/wiki/Computational_complexity_theory

Om någon vill fördjupa sig i detta och samtidigt finner texten i länkarna ovan lite tung finns det en utmärkt introduktion på svenska i kapitel 4 och 5 av boken *Tänkande och beräkning* av Lars-Erik Janlert. Rekommenderas varmt!

Resonemanget ovan leder till en utvidgning av antalet villkor som måste vara uppfyllda för att en dator skall kunna lösa ett problem. De är nu följande tre:

1. Det måste finnas en formellt precis beskrivning av problemet
2. Det måste finnas en algoritm som kan leda till en lösning av problemet
3. Denna algoritm måste ha en hanterlig tidskomplexitet.

Är dessa villkor inte uppfyllda finns det ingen teoretisk möjlighet att lösa problemet.

Vilka komplexa kognitiva problem kan en dator lösa?

Det korta svaret på rubrikens fråga är ”datorer kan inte lösa några intressanta komplexa kognitiva problem”, bara enkla välavgränsade problem. Men ”enkel” och ”välavgränsad” inte i betydelsen detta har för en människa utan för en dator, och på det sätt som beskrivs ovan.

Ett skäl att det inte finns några tillräckligt precisa beskrivningar av hur en människa löser ett komplext problem. Så villkor (1) ovan är inte uppfyllt. Det finns rimligt precisa beskrivningar av hur vår perception fungerar men för en intelligent agent, vare sig det är en naturlig agent (som en människa) eller en artificiell (som en robot) räcker det inte med att se och känna igen objekt i omvärlden. Man måste kunna agera utifrån det man ser.

Ett känt exempel på sådana problem som relativt lätt kan lösas av människor men som inte kan lösas av en dator är det så kallade Frame problemet. Enkelt uttryckt handlar det om problemet med att få en dator att bara beakta *alla* relevanta fakta och *inga andra* när ett problem skall lösas.

Detta problem är mycket bra beskrivet på webben i Stanford Encyclopedia of Philosophy, så jag hänvisar till resonemanget där. https://plato.stanford.edu/entries/frame-problem/
Det är det som i artikeln kallas *The Epistemological Frame problem* som är centralt.

[1] Med risk för att säga något för läsaren självklart visar utropstecknet i matematisk notation att det handlar om en fakultetsutveckling.

Något förvånande för den som inte känner till forskningsfältet är att ett fundamentalt problem är att få en dator att ignorera irrelevant information. En extremt kort sammanfattning av detta problem är att det inte finns något principiell begränsning i vilka fakta som kan vara relevanta för att lösa också enkla och vardagliga problem. Och utan principiella begränsningar kan allt vara potentiellt relevant. Och måste man ta hänsyn till alla fakta i hela världen så kan man aldrig komma fram till en slutsats. *Relevansproblemet* är alltså centralt.

Och eftersom man måste ha bestämt i förväg vilka fakta som ingår i ett problem för att man skall kunna skriva en algoritm som kan lösa det, enligt villkor 1 ovan, blir den oundvikliga slutsatsen att man aldrig med dagens datorteknik (alltså med en turingmaskin) kan skapa en generell AI som kan lösa alla problem som människor kan lösa i sin vardag. Man kan klara specifika problem genom att göra tydliga avgränsningar. Men då är handlar det inte om någon generell intelligens eller problemlösningsförmåga längre.

Man kan föra ett motsvarande resonemang när det gäller så kallad djupinlärning, som är det som många som inte kan något om kognition och datorer blir så imponerade av idag. I detta fall är relevansproblemet delvis ett annat. Man måste i förväg välja vilken träningsmängd som man skall träna sitt nätverk på (eller mer precist vilka element som skall ingå i träningsmängden), och man kan aldrig i förväg för komplexa problem veta vad man skall inkludera i träningsmängden. I grunden samma problem, men i en annan form.

Begränsningarna i djupinlärning diskuteras i detta föredrag för den som vill veta mera.

https://www.youtube.com/watch?v=Ckogtfn6zaI

Slutsats

Min slutsats, baserat på ovanstående resonemang, är att den generella AI:n inte kommer att kunna skapas med nu kända tekniker, och att det finns precist formulerade teoretiska skäl för varför det inte kommer att gå. Vi saknar dessutom precis kunskap om hur vi människor löser denna slags problem. Det handlar inte enbart om att skapa snabbare datorer. Man måste hitta en väg runt de teoretiska begränsningar som jag beskriver ovan[2].

Och det har mig veterligen ingen av dagens AI-entusiaster lyckats med. Vad jag vet har ingen av dem ens försökt diskutera de begränsningar jag beskriver ovan. Då hjälper det inte att de är lysande fysiker (Tegmark och Hawking), entreprenörer (Musk), eller filosofer (Boström). De är mycket kompetenta inom sina specialområden. Men de saknar vad jag vet, och att döma av det jag läst om deras argument, relevant kompetens för att kunna göra en saklig bedömning av AI-teknikens möjligheter.[3]

Jag vill understryka att resonemanget ovan inte betyder att diskussionen om framtida konsekvenser av AI är meningslös. Tvärt om, den är mycket viktig. Men det är en annan diskussion. Som måste utgå

[2] Ett alternativ är förstås att tro att man i framtiden kommer att ha något slags datorer som fungerar på ett helt annat sätt än dagens datorer, och där de begränsningar jag skriver om ovan inte föreligger. I mina ögon har man då lämnat en vetenskapligt grundad diskussion och håller i bästa fall istället på med science fiction.

[3] Det är förstås inte så lite ironiskt att ett fundamentalt resultat av dagens forskning om mänsklig och maskinell problemlösning att det inte räcker med att vara smart i största allmänhet – man måste dessutom ha djup kunskap inom det specifika aktuella problemområdet för att kunna lösa problemet, samtidigt som AI-entusiaster uttalar sig vitt och brett om vad AI kommer att kunna göra inom en nära framtid, utan att ha någon professionell kompetens inom vare sig teoretisk datavetenskap eller kognitionsvetenskap. Och visar därmed både sin egen okunnighet om vad forskningen visar, och i sitt eget agerande visar ett tydligt exempel på vad bristen på domänkunskap när det gäller att dra slutsatser och lösa problem.

ifrån vad vi idag faktiskt vet om vad datorer kan göra i framtiden. Annars gör den mer skada än nytta. Vi spiller tid på att diskutera fullständigt orealistiska framtidsscenarion, istället för att diskutera konsekvenserna av det som man idag och i morgon faktiskt kan göra med dagens avancerade AI-teknik.

Till sist: Jag gör givetvis inte anspråk på att resonemanget ovan är det slutgiltiga svaret på om det kommer att gå att utveckla en generell AI. Men jag vill hävda med viss emfas att de som tror på möjligheterna att utveckla generell AI måste bemöta argumenten ovan och visa var de är fel. Vilket jag hittills inte sett några exempel på. Om du som läser detta känner till några sådana är jag tacksam om du hör av dig till mig om detta. Och om du som läst detta tycker att du har ett motargument till det jag skriver ovan så ser jag fram emot att höra ifrån dig!