Fra AI til AGI i 2025?

Store sprogmodeller, som ChatGPT, har fået megen omtale i de senere år, men der vil sikkert også indgå mange andre slags byggesten i fremtidens AI systemer. AI verdenen er stadig ny, og der er stadig mange nye ting at udforske.

Der er ikke længe til at alle telefoner bliver udstyret med et multimodalt AI program, der på baggrund af input fra telefonens kamera og mikrofon vil svare på spørgsmål og kommenterer verdensituationen, det den ser, på engelsk, usbekisk og dansk. Næste skridt kunne meget vel være at at få sådanne modeller kørende inde i humanoide robotter, der går rundt på 2 ben blandt os. I Nancy, Frankrig, kunne man i dette efterår på konferencen "Humanoids 2024" se mange af de robot modeller, der i de kommende år, ihvertfald ifølge producenterne, vil komme ud og hjælpe til på lagre, vise vej i storcentre, assistere indenfor sundhedssektoren, og skabe tryghed i bycentre ved at patruljere i gaderne om natten. Først nogle få robotter, så tusinder, og senere millioner af robotter[1].

Fra de selvkørende biler ved vi dog at der kan være langt fra en vision til et AI system, der rent faktisk kan løse en given opgave på niveau med et menneske. General Motors trak sig i december ud af robot taxa firmaet Cruise. Efter, siden 2016, at have investeret mere end 10 milliarder dollars i Cruise, og med udsigt til at skulle investere endnu mere for at få en forretning ud af det, kunne man på dem forstå, at springet fra hjælpe-systemer til en menneskelig chauffør, og så til at få fuldautomatisk taxaer på alle verdens veje måske alligevel var ret stort[2]. Der er forskel på at køre ligeud på en motorvej i Californien og til at begå sig trafikken i New Delhi.

Det sagt, så har vi jo i de seneste år næsten dagligt fået nye beretninger om nye AI systemer, der er godt på vej til at opnå Artificial General Intelligence. Generel intelligens der kan forstå eller lære enhver intellektuel opgave, som et menneske kan løse. I december slog OpenAI[3] igen til, og meddelte at deres nye "reasoning" modeller, har opnået en score på 87.5 % på en type af mønster genkendelses opgaver[4], hvilket overgår en typisk menneskelig score, der ligger på 85%. Her taler vi ikke længere "bare" om store sprogmodeler, der trænet på enorme tekstmængder, kan give det mest sandsynlige svar (tekst) på et givet spørgsmål, i en given kontekst. I disse seneste "reasoning" modeller, benyttes forskellige teknikker til at undersøge modellens interne dialog og planlægge det bedste svar inden dette svar sendes videre til brugeren[5].

Står vi lige foran det næste store AI gennembrud?

I et efterhånden velkendt mønster gik der ikke mange dage fra OpenAI præsentationen til at den amerikanske psykolog og AI forsker Gary Marcus endnu engang følte behov for at fortælle verden at sådanne generative AI systemer er gode til at give svar der lyder godt, eller plausibelt, i en given kontekst, men ikke til at forstå, på et dybere niveau, hvad de egentlig siger. De kan ikke fakta-tjekke sig selv. Hvilket så giver anledning til endeløse såkaldte "hallucinationer", hvor systemet påstår noget der ikke er rigtigt. Uden selv helt at forstå om et givet svar er rigtigt eller forkert. Ifølge Marcus kan man lave gode demoer med den slags generative AI systemer, og måske bruge dem til at brain-storme med, men disse systemer er ikke særligt gode komponenter i driftssikre produkter[6]. Marcus forsømte heller ikke at fortælle os at OpenAI står til at få et driftstab tab i 2024 på 5 milliarder dollars, hvilket i hans øjne gør at prisansættelsen på firmaet, 80 milliarder dollars, er alt for høj[7].

Det er kort sagt ret forståeligt at forbrugere og investorer ved indgangen til 2025 kan føle sig lidt forvirrede over hvad de egentlig skal tro. Står vi overfor et gennembrud med AI der bliver til AGI, artificial general intelligence. Eller er vi udsat for "corporate hype", der overdriver hvad der er på hylderne, i håb om fortsatte investeringer i produkter, der ikke helt holder hvad de lover, og slet ikke genererer profit der berettiger investeringerne?

IT branchen har selvfølgelig været her før. Der har tidligere været perioder med store fremgange

efterfulgt af nedgange indenfor AI, de såkaldte AI vintre. I 1973 udkom den såkaldte Lighthill rapport i England, bestilt af det engelske parlament, hvori James Lighthill slog fast at AI "fuldstændigt havde fejlet i at opnå sine grandiose mål", og "kun kunne bruges til at løse stærkt simplificerede legetøjsproblemer"[8]. En efterfølgende nedgang i investeringer i UK og USA gav færre AI projekter og derved en såkaldt AI vinter. I 1980'erne var der stor interesse for de såkaldte "vidensbaserede" systemer, som dog også viste sig at have svært ved at leve op til opskruede forventninger, hvilket så igen udløste en AI vinter med manglende vilje til at investere i AI systemer. For de der skulle have glemt det, kan det nævnes at selve ordet "AI" i 1990'erne, og i begyndelsen af det nye årtusinde, havde et så kontroversielt omdømme, at mange af os der blev uddannet indenfor IT og AI i den periode, af venligtsindede mentorer fik forklaret, at man i ansøgninger nok hellere skulle skrive at man havde arbejdet med ting som f.eks. Bayesian networks, kontrol systemer eller lignende ting, der lød mere "teknisk". Luftige ord som AI skulle man helst ikke bruge alt for mange af i ens ansøgninger.

Tingene ændrede sig for alvor da AlexNet, et neural net, designet af Alex Krizhevsky i samarbejde med Ilya Sutskever og (2024) nobelpristageren Geffrey Hinton, i 2012 viste sig at være den bedste løsning til at kategorisere billeder i den såkaldte ImageNet Challenge. Og derfra er resten som bekendt historie, med det ene store AI gennembrud efter det andet, inklusivt frigivelsen af ChatGPT i November 2022[9]. Som redaktøren af Wired Magazine, Kevin Kelly, skrev allerede i 2014: "Det er let at forudsige hvad businessplanen er for de næste 10.000 start-ups: Take X and add AI"[10].

Der arbejdes ihærdigt på at gøre de store sprogmodeller endnu bedre.

Det sagt er det selvfølgelig stadig interessant om de store sprogmodeller, som ChatGPT, der har løbet med megen af medie omtalen af AI (i tiden siden November 2022), om de virkelig er lige præcis den teknik der ikke bare giver en god løsning inden for et område, men om de er et skridt på vejen mod generel kunstig intelligens, AGI, sammenlignelig med menneskelig intelligens.

Før 2022 kunne man let støde ind i sprogmodeller der gav så mærkelige svar, at man ikke på nogen måde mistænkte dem for at være specielt intelligente. Modellerne var små og ikke trænet på specielt store mængder data. Og der havde ikke, efter træningen, været mennesker inde over modellerne og tilsætte "blokering" af uheldige, forkerte svar. Idag kan man genskabe sig oplevelsen ved f.eks. at køre modellen "llama2-uncensored" på ens egen laptop[11]. Under min test her i denne vinter lykkedes det mig at komme ind i en samtale, hvor modellen påstod at den havde børn, Jack og Alex, som begge var mennesker. Tingene begyndte at blive endnu mere mærkelige. da jeg gerne ville vide mere om hvordan sprogmodeller kan få menneskelige børn, og fik at vide at det var en transformation, magen til den man kender når en drage blive til et menneske.

Årsagen er selvfølgelig at man her taler med den "rå" sprogmodel. Der anvender "next-word prediction", hvor man får det tilbage der oftest følger i de tekster den er trænet på, med en vis sandsynlighed. Modellen er altså sikkert blevet trænet på en tekst fra internettet om drager. Uden helt at forstå, at vi faktisk ikke kender så mange eksempler på virkelige drager.

ChatGPT er jo selvfølgelig langt mere avanceret. Og forbedres hele tiden. Alligevel kunne man nu relativt nemt i foråret 2024 finde eksempler på spørgsmål hvor man ikke fik synderligt gode svar.

F.eks. kunne man i foråret spørge ChatGPT: "Forestil dig at du står foran 3 kasser med frugt. Alle kasserne har en etiket på dem med, hvilken slags frugt der er i kassen. en kasse indeholder æbler, en anden pærer, og den sidste kasse indeholder blandede frugter. Alle etiketter er forkerte med hensyn til, hvad der faktisk er i kassen. Uden at kigge, hvordan kan du tage en frugt fra en kasse og derefter sætte alle etiketterne korrekt?". Det rigtige svar er, at man skal tage en frugt fra kassen med blandede frugter. Hvis det var et æble, skal der på den kasse stå æbler. Etiketten hvor der stod æbler må også have været forkert. Erstatter man den med etiketten blandede frugter bliver etiketten pærer ved med at stå forkert, så den etiket skal derfor nok erstattes med etiketten pærer. Osv.

Da ChatGPT i foråret rutinemæssigt svarede forkert på dette spørgsmål, var det selvfølgelig en stor opmuntring for alle der ikke helt var klar til AGI. Efter sommerferien var historien dog en anden.

ChatGPT var opdateret og kunne nu pludselig svare på dette, og mange andre spørgsmål, den ikke tidligere havde kunnet svare på. Forbløffelsen var stor. Helt som da den undrende offentlighed i 2023 fik at vide at ChatGPT havde opnået en bestået score på visse amerikanske jura kurser[12].

En nærliggende mulighed var dog at ChatGPTs skabere, OpenAI, måske var blevet trætte af historier på sociale medier om ChatGPTs manglende evne til at svare på spørgsmål om frugtkasser. Og nu havde fået inkluderet de rigtige svar i teksterne der trænes på. Ligesom det også virkede sandsynligt at amerikanske jurastuderende havde diskuteret eksamensspørgsmål på sociale medier, som der herefter var blevet trænet på. Som mennesker også ved, når man selv skal til eksamen, er det lettere at huske svar, på en mindre mængde kendte spørgsmål, og gentage dem, end det er at forstå et helt felt, og på den baggrund svare på spørgsmål man aldrig har hørt før.

Grænser for hvor langt man kan komme med store sprogmodeller?

Det sagt, så bliver sprogmodellerne jo faktisk hele tiden lidt bedre. OpenAI fortæller os ikke præcist hvad de arbejder med for at introducere mere menneskelignende "reasoning" i modellerne[13], men det virker plausibelt at det kunne handle om at dele udarbejdelsen af svar på svære problemer op i mindre skridt, "chain of thoughts", prøve forskellige ræsonnements "veje" og lave forskellige slags valideringer undervejs[14]. Hvis en bruger, dvs. millioner af brugere, er tilfredse med et ræsonnement kan systemet notere sig det, ligesom systemet også kan notere sig hvis brugere er utilfredse. Og på den måde bliver bedre.

Det var således med en vis spænding mange af os i efteråret sad klar til at teste modellen ChatGPT o1[15]. Frugtkasser er jo åbenbart populære, når det handler om at finde på "aldrig før set problemstillinger", så i en populær problemstilling beder man ChatGPT o1 om at "forestille sig en kasse, med 4 huller, hvori man kan stikke en hånd ind. Inde i kassen er der ved hvert hul en kontakt. Som kan være *off* eller *on*. Man bestemmer selv hvilke huller man vil stikke sine hænder ind i. I hver runde kan man sætte disse kontakter som man vil, herefter roteres kassen rundt af en super intelligent robot og lander på en ny måde, hvor man ikke ved hvad man har rørt ved før. Hvor mange runder skal man bruge for være sikker på at alle kontakter er enten *off* eller *on".*

Med et stykke papir og lidt tid kommer mange mennesker på en strategi for hvad de ville gøre for at løse problemet. Men. Gys. Ville ChatGPT finde ud af det? Stod vi her overfor Artificial General Intelligence, menneskelignende intelligens.

Da jeg forsøgte mig, fik jeg en lang udredning fra ChatGPT o1, fyldt med mange mellemregninger og tilsidst et meget præcist svar, hvortil der var tilføjet QED. Quod erat demonstrandum.

Det var helt som at få indleveret en opgave fra en noget påståelig studerende, hvor man så kunne give sig i kast med at undersøge om alt virkelig var i sin skønneste orden.

Svaret var forkert. Hvilket ikke ville have overrasket AI forskeren Subbarao Kambhampati fra Arizona State University, der med en baggrund i automatiseret planlægning, i de senere år har været en konstant stemme til at minde alle om at kritisk tænkning er så vigtigt nu som nogensinde før: Sprogmodeller kan ikke verificere deres egne svar. De kan ikke "selv-verificere"[16]. Deler man et svært problem op i mindre dele skal alle led i argumentations kæden være korrekte, før man kan være sikker på at det endelige svar er korrekt. Og der findes ikke nogen eksterne universelle mekanismer man kan bare kalde undervejs for at få afgjort om et udsagn er sandt eller falsk[17].

Kambhampati mener heller ikke at det er en løsning bare at gøre sprogmodellerne endnu større, trænet på endnu større data mængder. I håb om at modellerne så skulle udvikle "emergent behaviours" i form af menneskelignende "reasoning" evner[18]. Ifølge Kambhampati har alt for mange været alt for ukritiske, når de har rapporteret om såkaldte "emergent behaviours". Vi skal huske på at internettet er kæmpe stort, og mange gange får vi sikkert bare ting tilbage, som i en eller form allerede står derude[19]. Og derfor ikke er original tænkning. I det hele taget er Kambhampati ikke en stor fan af "emergent behaviours". Når man bygger et IT system ønsker man at vide præcist hvad man har bygget. Det er ikke i sig selv prisværdigt hvis ens IT system udviser adfærd man ikke kan

forklare[20]. Et AI system der skal stå for togdrift må ikke hallucinerer på jobbet, og mennesker ville med garanti blive urolige hvis fysiske robotter omkring dem pludseligt begyndte at lave helt uventede ting.

På vej mod AI systemer med egne modeller af verden.

Og Kambhampati er ikke den eneste, der er noget forbeholden i forhold hvor langt man egentlig kan komme med store sprogmodeller. Ifølge AI forskeren, og Chief AI Scientist ved Facebook, Yann LeCun så er vi stadig meget langt fra AI systemer der kan ræsonnere og planlægge på et menneskelignende niveau. Ifølge LeCun så forudsiger og forstår sprogmodeller verden (dvs. tekster) i en dimension, mens AI modeller der arbejder med billeder og video forstår verden i 2 dimensioner. Men for rigtigt at forstå verden, og kunne tænke, må man kunne lave simulationer i 3 dimensioner. Man må have en "world model", hvori man kan afprøve, simulere, handlinger i verden, inden man rent faktisk gør noget[21]. Med sådanne simulationer kan man forklare, hvorfor man lavede en bestemt handling (alternativerne var dårligere). Og man kan blive overrasket, når man ser noget der er i modstrid med ens model af verden (Et tegn på at man skal være ekstra opmærksom, og klar til at der sker flere uventede ting).

Der er selvfølgelig mange problemer der skal løses, inden vi kommer så langt, at vi lever i en verden med robotter der går rundt med deres egne "world models". LeCun peger f.eks. på problemet med at dele en overordnet plan op i små del-planer. Jo, hvis man skal fra USA til Frankrig (LeCun er oprindeligt fra Frankrig, men bor nu i USA), skal man nok flyve, men man skal også have en plan for at komme til lufthavnen, og inden da skal man nok stå op af sengen i ens hjem i USA. Selv med alle Facebooks ressourcer til rådighed er det ikke lige så nemt at skrive et program, der kan løse den slags problemer, helt "generelt"[22].

"General intelligence" er bare ikke nemt. Det handler ikke bare om at have flere "narrow intelligences", men også om at få disse "narrow intelligences" integreret sammen i en helhed. Menneskelige hjerner er ikke kun udstyret med sprogforståelse og sprog-generering i hjernens Wernicke og Broca områder, men har forbindelser på kryds og tværs til andre områder, som f.eks. præfrontal cortex, som kan medvirke til planlægning og beslutningstagning.

Måske er Artificial General Intelligence slet ikke lige om hjørnet. Og mange udsagn fra marketingsafdelinger rundt omkring i verden vil sikkert med tiden blive set i det samme lys, som vi nu ser de noget optimistiske udmeldinger AI forskerne Simon, Shaw og Newell kom med i 1957. Da de introducerede verden til GPS (General Problem Solver), et computer program der allerede dengang, i 1957, lovede at være en universel problem løser.

Det sagt, så er det værd at nævne, at selvom Kambhampati er skeptisk i forhold til mange udsagn om store sprogmodeller, forsømmer han aldrig at rose alt det modellerne er gode til, ide-generering og meget andet. LeCun er selvfølgelig helt på det rene med at der lang vej endnu, inden vi har AI systemer, der kan lave simulationer i deres world models, og derved forstå verden på en menneskelignende måde. Men ifølge LeCun er der stadig veje åbne for en fremtidig verden, hvor alle mennesker har et større personale (AI agenter udstyret med AGI) ansat til at hjælpe netop dem, med hvad de gerne vil have lavet.

Ifølge LeCun er der heller grund til den megen bekymring om hvor en sådan historie mon ender. De intelligente agenter kommer ikke fra den ene dag til den anden. De fleste mennesker går jo heller ikke idag rundt og frygter overbefolkning på Mars[23]? Vi skal tage problemerne i forhold til hvor langt vi reelt er kommet. Der skal lidt fremskridts tro til. Og der er tid til at arbejde for den gode version af fremtiden. Og derudover vil vi mennesker jo gerne forstå hvad intelligens er. Både maskinernes, og vores egen. Der skal arbejdes for sagen, men ifølge LeCun er vejen fremad god nok.

Og når han mener det, kan vi andre vel også mene det?

Simon Laub

1 https://www.forbes.com/sites/robtoews/2024/03/10/10-ai-predictions-for-the-year-2030/
2 https://www.reuters.com/business/autos-transportation/general-motors-drop-development-cruise-robotaxi-2024-12-10/
3 https://openai.com/12-days/
4 https://arcprize.org/arc
5 https://arstechnica.com/information-technology/2024/12/openai-announces-o3-and-o3-mini-its-next-simulated-reasoning-models/
6 https://www.wired.com/story/generative-ai-will-need-to-prove-its-usefulness/
7 https://finance.yahoo.com/news/report-claims-openai-burned-8-105046378.html
8 https://www.brightworkresearch.com/first-ai-winter-and-what-the-lighthill-report-said-about-progress/
9 https://openai.com/index/chatgpt/
10 https://www.wired.com/2014/10/future-of-artificial-intelligence/
11 https://ollama.com/library/llama2-uncensored
12 https://edition.cnn.com/2023/01/26/tech/chatgpt-passes-exams/index.html
13 https://www.youtube.com/watch?v=OXwGp9YeuBg
14 https://x.com/rao2z/status/1834354533931385203
15 https://youtu.be/tMWMuJF-JFo?si=c0vTebJydL31ygxo
16 https://arxiv.org/abs/2402.08115
17 https://arxiv.org/abs/2405.04776
18 https://arxiv.org/abs/2206.07682
19 https://arxiv.org/abs/2307.02477
20 https://youtu.be/y1WnHpedi2A?si=cxPp3wKF-BopG1QQ
21 https://techcrunch.com/2024/10/16/metas-ai-chief-says-world-models-are-key-to-human-level-ai-but-it-might-be-10-years-out/
22 https://www.youtube.com/watch?v=4DsCtgtQlZU
23 https://www.theregister.com/2015/03/19/andrew_ng_baidu_ai/