Fra AI til AGI i 2025?

Hvad er det egentlig dagens AI mangler for at kunne blive til AGI, menneskelignende intelligens?

Indenfor AI er det i høj grad de store sprogmodeller, som ChatGPT, der i de seneste år har trukket de store overskrifter. Modellerne forbedres hele tiden. Og 2024 var fyldt med overskrifter om at AI nu nærmer sig et menneskelignende niveau. Så starten af 2025 virker som en god anledning til at reflektere lidt over (del 1) de seneste AI landvindinger og prøve at komme det lidt nærmere (del 2) hvad der egentlig mangler for at AI bliver til AGI, menneskelignende intelligens.

Der er ikke enighed om hvad Artificial General Intelligence præcist er. Menneskelignende intelligens kunne være et bud. Men mennesker er ikke lige intelligente på alle områder. Nogle mennesker er meget musikalske, nogle er gode til logisk tænkning og andre har stor selvindsigt osv. AI har forlængst overgået menneskelig intelligens indenfor specifikke områder, såsom at spille skak, hurtigt at kunne oversætte mellem alverdens sprog[1], forudsige 3D strukturer i proteiner[2] osv. Men det er selvfølgelig noget ganske andet at have maskiner, med generel intelligens, der kan udføre en bred vifte af opgaver på samme niveau som et menneske.

Og det er jo så her AI sprogmodellerne bliver interessante, da de jo tilsyneladende kan tale med om hvad som helst. Ja, indtil videre har svarene fra sprogmodellerne været noget "automatiske", ubevidste. Der har ikke været nogen bagvedliggende refleksion, med sammenhængende logisk tænkning og med inddragelse af egne erfaringer. Det sagt, så bliver sprogmodellerne jo faktisk hele tiden lidt bedre, og der har følgeligt ikke været en dag i 2024, som ikke har været fyldt med spekulationer om hvordan sprogmodellerne snart vil kunne tænke og ræsonnere som mennesker.

ChatGPTs skabere, OpenAI, fortæller os ikke præcist hvad de arbejder med for at introducere mere menneskelignende "reasoning" i modellerne[3], 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[4]. 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[5]. 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"[6]. 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[7].

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[8]. 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[9]. 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[10]. 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[11]. 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"[12].

"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[13]? 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?

- Dette var del 2 af "Fra AI til AGI i 2025".

Simon Laub

---

[1] https://blog.google/products/translate/google-translate-new-languages-2024/
[2] https://deepmind.google/technologies/alphafold/
[3]https://www.youtube.com/watch?v=OXwGp9YeuBg
[4]https://x.com/rao2z/status/1834354533931385203
[5]https://youtu.be/tMWMuJF-JFo?si=c0vTebJydL31ygxo
[6]https://arxiv.org/abs/2402.08115
[7]https://arxiv.org/abs/2405.04776
[8]https://arxiv.org/abs/2206.07682
[9]https://arxiv.org/abs/2307.02477
[10]https://youtu.be/y1WnHpedi2A?si=cxPp3wKF-BopG1QQ
[11]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/
[12]https://www.youtube.com/watch?v=4DsCtgtQlZU
[13]https://www.theregister.com/2015/03/19/andrew_ng_baidu_ai/