Mandag
16. juni
2025

Ideer & samfunn

Aftenposten

Grunnlagt 1860
Av Chr. Schibsted

Aftenposten
arbeider etter Vær Varsom-plakatens regler for god presseetikk.

Trine Eilertsen
Sjefredaktør

Tone Tveøy Strøm-Gundersen
Nyhetsredaktør

Kjetil B. Alstadheim
Politisk redaktør

Cecilie Asker
Kulturredaktør

Erik Tornes
Debattredaktør

Jostein Larsen Østring
Utviklingsredaktør

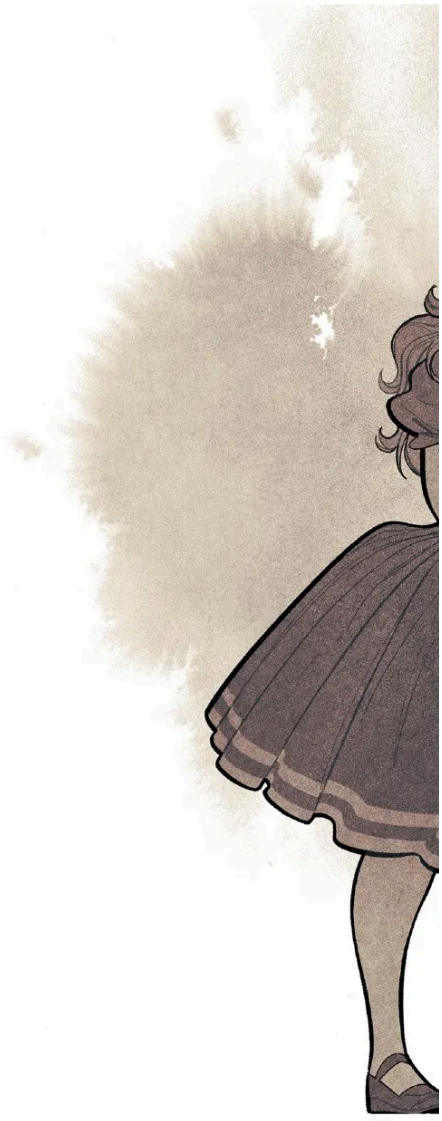
Audun Solberg
Redaktør avis og forlag

Redaktøransvar

A | På en mandag
Inga Strümke
KI-forsker, PhD, Norwegian Open AI Lab

La oss ikke ha tillit til kunstig intelligens

Vi bør ikke godta at KI-produkter markedsføres som om de var menneskelige.



Det er ikke godt å vite hvilke kunstig intelligens-løsninger (KI) som er forbigående trender, hvilke som vil være nyttige også om noen år, og ikke minst hvilke som er så gode som leverandørene lover. Språkmodeller er særlig utfordrende: De dukker opp overalt, og mange mener at de er selve veien inn i en kunstig intelligent fremtid. Samtidig tyder stadig mer forskning på at optimismen er prematur og bidrar primært til å øke aksjekurser.

Språkmodell som «tenker seg om»? Det siste året har chatboter som inneholder store språkmodeller, begynt å skrive «thinking ...» idet de skriver et svar. Det som ligger bak denne «tenkingen», er en prosess som litt slepphendt er blitt døpt *chain of thought*, altså tankerekke.

Navnet er noe misvisende, for det som foregår, er at spørsmålet du stiller chatboten, eksplisitt brytes ned i mindre delspørsmål. Dette har vist seg å hjelpe chatboter til å løse flertrinnsproblemer som matematiske gåter og er et smart grep for å omgå språkmodellenes svakheter.

Å påstå at en språkmodell «tenker seg om» er dog misvisende og i verste fall manipulativt.

En fersk studie fra forskere ved Apple peker på at den tilsynelatende «tenkingen» bryter sammen i takt med at kompleksiteten i spørsmålet øker. Dette er ikke overraskende, gitt at teknikken ikke er en løsning på, men en omgåelse av det underliggende problemet, nemlig at språkmodellers evner til resonnering er uvisse og ustabile.

Hvorvidt påstanden om «thinking ...» bare er irriterende upresis eller problematisk feilinformasjon, er avhengig av konteksten rundt teknologien.

Vellykket global reklamekampanje. I dag baler de fleste sektorer med chatboter, og det føles nærmest obligatorisk å ta dem i bruk.

Det finnes flere eksempler på at det hele kan bære galt av sted, kanskje pinligst da Trump-administrasjonens rapport «Make America Healthy Again» inneholdt referanser til studier som ikke finnes.

Denne typen innhold som er usant, men samtidig plausibelt, er blitt selve fotavtrykket til språkmodeller i rapporter.

At språkmodeller ikke forstår årsakssammenhenger og ikke forholder seg til den virkelige verden, er et problem for fagfeltet jeg jobber med. Men den vellykkede globale reklamekampanjen der teknologi-

Les mer på ap.no

Hvorvidt påstanden om «thinking ...» bare er irriterende upresis eller problematisk feilinformasjon, er avhengig av konteksten rundt teknologien.

Tegning: Christian Bloom

selskaper overbeviser beslutningstagere om at språkmodellene de utvikler, er en obligatorisk del av en effektiv fremtid, er et problem for hele samfunnet.

Tolkningsfriheten i markedsføring av KI. Fremtiden med KI finnes ikke ennå. Vi skaper den fortløpende gjennom å velge hvilke KI-systemer vi tar i bruk og hvilke vi lar ligge.

Videre har vi både retningslinjer og lovverk som begrenser hvordan KI-systemer får påvirke vår helse, sikkerhet og våre rettigheter. Men fokuset er ofte på selve systemene, ikke hvordan de fremstilles.

Markedsføring som sådan er godt regulert. Du har ikke lov til å hevde at produktet ditt er eller gjør noe annet enn det faktisk er eller gjør. Men vi må se nærmere på tolkningsfriheten i markedsføring av KI-produkter.

Datamaskiner kan ikke bry seg. For å ta et konkret eksempel: Om du søker etter den KI-drevne chatboten Replika, møter du markedsføring som hevder at dette er «the AI companion who cares», altså KI-følgesvennen som bryr seg, og som dessuten «alltid er der for å lytte og snakke, alltid ved din side».

Å påstå at et dataprogram «bryr seg» er åpenbart å trekke det for langt. Datamaskiner har pr. i dag ikke utviklet følelser, og de kan følgelig ikke bry seg.

At chatboten dessuten «alltid» skal være der for deg, er en modig garanti å gi, gitt at Replika for eksempel ble stanset av det italienske datatilsynet på grunn av brudd på personvern. Vips, så var den slettes ikke alltid der lenger. For en tenkt bruker som har utviklet en emosjonell tilknytning til chatboten, vil dette kunne være problematisk.

Replika er kun ett av flere eksempler. 29. april rullet OpenAI tilbake nye GPT-4 til en tidligere versjon fordi den nyeste var så hyggelig og bekreftende at brukere reagerte negativt. Slik modelloppførsel refereres ofte til som «sycophancy», det engelske begrepet for uoppriktig smigrende oppførsel.

At chatboter må tilpasses og finjusteres for å gi brukere en behagelig opplevelse, er ikke overraskende, men når OpenAI sier at «ChatGPTs personlighet har en dyp innvirkning på hvordan du bruker og stoler på den», har de allerede lagt fast at brukere skal utvikle tillit til KI-produktet deres. Dette kommer i tillegg til den misvisende påstanden om at produktet «tenker».

Politisk heller enn faglig motivert. KI har en egen aura av nærmest magi og datamaskintryllekraft over seg. Hvordan bedrifter markedsfører KI-produkter, påvirker ikke bare hvilket inntrykk brukeren får, men også inntrykket samfunnet danner seg av KI som helhet.

De siste månedene har vi vært vitne til hvor stor makt en teknolog med økonomiske muskler og en posisjon i det offentlige ordskiftet kan utøve i et demokrati. Elon Musk proklamerte nylig at KI vil overgå menneskelig intelligens innen neste år, samtidig som han kjøpte seg lojalitet fra president Donald Trump. Man trenger ikke være konspiratorisk for å mistenke at spådommen er politisk heller enn faglig motivert.

Kan vi ha tillit til KI-teknologi laget i en slik kontekst?

Ansvaret de største tilbyderne har. Det finnes flere definisjoner på tillit, og min favoritt lyder «villigheten en part har til å være prisgitt en annen part, der den første ikke har kontroll over den andre» (fritt oversatt av meg).

La oss ikke godta en situasjon der vi er prisgitt KI-systemer vi ikke har kontroll over.

Tillit kan vi ha til aktører med intensjoner og som dessuten tar beslutninger basert på etiske vurderinger. Å tilskrive KI-systemer tillitsverdighet malplasserer et konsept for relasjoner (forhold mellom aktører) og moralske vurderinger (som gjøres av aktører) hos en ikke-aktør.

Det hele er med andre ord ikke bare misvisende, men uærlig. Det kan også være en effektiv måte å feie under teppet ansvaret de største tilbyderne av KI-teknologi har i samfunnet.

Bør vurdere nøye. Tanken på å dele kloden med en ny type intelligent aktør får mange av oss til å revurdere selve menneskets rolle og verdi. Det hele blir brått eksistensielt. Hvordan vi oppfatter KI, vil sannsynligvis påvirke hvordan vi oppfatter både oss selv og vår rolle i samfunnet.

Vi bør ikke godta at KI-produkter markedsføres som om de var menneskelige. Hva bedrifter påstår om KI-systemer, påvirker vårt kollektive bilde av hva KI er og hvilken rolle KI-produkter kan ta i samfunnet.

Vi former fremtiden vår med KI, og vi bør vurdere nøye hva vi tillater av markedsføring og løfter rundt KI-produkter.

Dernest bør vi være forsiktige med hvilke løfter vi tror på om hvor nyttige og nærmest uunnværlige KI-produkter vil være, særlig når det er snakk om produkter vi ikke rår over selv.