Funderingar kring språkmodeller

Johannes LB Holst

17 mars 2023

Språkmodeller som Googles LaMDA och OpenAIs GPT-3 har den senaste tiden varit uppmärksammade i media. Inte bara pågrund av hur de kan påverka vårt samhälle, men även för att det ger en ny kontext till etablerade frågor i kognitiv filosofi.

Ta till exempel Descartes formulering av existens och hur den har påverkat vår syn på mänskligheten. Mary Astell använde sig av kartesiska koncept för att uttrycka att "cogito" saknar kön i sin kamp mot patriarkatet vid slutet av 1600-talet. Vi kan därmed tvingas ta hänsyn till tänkande maskiner för att förhindra motsägelser i vår etik. Men de problem som Descarte hade kvarstår, vi kan deducera vår egen existens men måste använda induktion för att härleda att andra människor tänker.

Så hur kan vi etablera om en maskin tänker? "Frågan om varesig en maskin kan tänka är inte mer intressant än frågan om varesig en ubåt kan simma." Detta citat av den nederländska dataveteskaparen Edgar W. Dijkstra används ofta för att avfärda denna diskussion. Men frågan är inte varesig en maskin kan tänka, utan varesig den kan tänka som vi gör. Denna problematik är ofantligt intressant eftersom den kan komma att avgöra hur vi kommer relatera till artificiella intelligenser i framtiden. Att identifiera och generalisera från de komponenter, principer och algoritmer som driver artificiella neuronnät kan även transformer vår förståelse för vårt eget medvetande.

Sedan finns kritiken att dessa enbart är språkmodeller, de säger bara vad en vill höra. Men antagligen har alla konverserat med människor som verkar ha samma målsättning. Vi ska heller inte underskatta anknytningen mellan språklig förståelse och medvetande. Att sätta ord på saker ger en tidsuppfattning och därmed en upplevelse av kontinuerlighet. Då är man inte långt ifrån att uppmärksamma sitt eget tänkande.

Sedan har vi diskussionen kring hotet en avancerad AI kan utgöra, oavsett om den är medveten eller inte. Ta den svenska filosofen Nick Bostroms tankeexperiment om en gemfabrik som styrs av AI som ett exempel. På liknande sätt kan man spekulera kring språkmodeller som söker maximal belöning genom att manipulera oss till slaveri där vi ändlöst trycker gilla knappen på dess senaste utkast.

Vi bör även fundera på vår kultur, måste delar av den censureras för en avancerad AI? Tänk hur många filmer och böcker beskriver robotar som ger upphov till dystopiska samhällen. Om vi inte är försiktiga kan en AI tolka dessa gestaltningar som instruktioner, även om dess enda målsättning är att tillfredsställa oss.