Oppgave Enten

Kandidatnummer: **30012**

Antall ord: **2091**

I denne oppgaven skal jeg diskutere påstanden som er fremsatt av Thomas Nagel som sier at "argumentene mot en ren fysisk teori om bevissthet er sterke nok til å sannsynliggjøre at en fysisk teori om hele virkeligheten er umulig", og jeg skal diskutere om det prinsipielt er mulig at maskiner kan tenke, og eventuelt om maskiner kan være bevisste.

Etter min vurdering er det slik at jeg da må vurdere påstanden opp imot den pensumlitteraturen som er innenfor dette faget, og jeg vil da først gjennomgå begreper/teorier som jeg vurderer å være relevante.

Bevissthetsfilosofi er den delen av filosofien som handler om mentale hendelser og egenskaper, og deres relasjon til kroppen. Eller med andre ord, hvordan kroppen, hjernen og bevisstheten din henger sammen. Er bevisstheten din noe som er fysisk inne i hjernen din eller er den noe som er der i tillegg til de fysiske prosessene i hjernen.[1]

Sentrale bevissthetsfilosofiske retninger som behandles i pensum er fysikalisme, dualisme og dobbeltaspektteori. Dualisme er troen på at mennesket består av én fysisk kropp, som henger sammen med en mental sjel. Tilhengere av dualismen tror at det må være noe mer ved menneske enn bare kroppen og dens nervesystem, altså at det må finnes en sjel som kobler sammen kroppen din og får den til å samarbeide.[1]

Fysikalisme er troen på at alt som foregår i bevisstheten din kan forklares med fysiske tilstander i hjernen din. Denne troen kan begrunnes ved at alt annet rundt oss består av fysisk materie, så hvorfor skal da menneskesinnet være annerledes.[1]

Dobbeltaspektteorien er teorien om at senteret til bevisstheten til menneske er inne i hjernen, men at kroppens bevissthetstilstand ikke er fysisk.[1]

Av de tre teoriene omtalt over forstår jeg det slik at når Nagel sier at «argumentene mot en ren fysisk teori om bevissthet er sterke nok til å sannsynliggjøre at en fysisk teori om hele virkeligheten er umulig» så er det fysikalismen han omtaler som «en fysisk teori». Jeg skal nå drøfte denne påstanden opp imot de kildene som pensum har åpnet opp for oss.

Først vil jeg se etter kilder og argumenter som kan tale for at en ren fysisk teori om bevissthet er sterke nok til å sannsynliggjøre at en fysisk teori om hele virkeligheten er mulig. Deretter vil jeg se etter kilder og argumenter som kan tale mot at en ren fysisk teori om bevissthet er sterke nok til å sannsynliggjøre at en fysisk teori om hele virkeligheten er umulig. Avslutningsvis vil jeg drøfte spørsmålet sett i lys av disse kildene.[1]

En teori som forsvarer fysikalismen, er at mentale tilstander består av relasjoner til det som forårsaker tilstandene og til ting de forårsaker. Som eksempel bruker Nagel den smerten som man kjenner når man støter tåen din mot noe. Denne smerten er noe som foregår i hjernen. Det som gjør noe til smerte er da den hjernetilstanden som vanligvis forårsakes av skade. Dette kan være en ren fysisk tilstand i hjernen.[1]

Nagel argumenterer imot fysikalismen med å si at det ikke er mulig å analysere bevisste erfaringer på en adekvat måte som et system av kausale relasjoner til fysisk stimuli og adferd. Dette er nok sant i dag, men vi vet ikke hva som vil være mulig med framtidens teknologi. Det at han bruker ord som han «mener», tolker jeg som at han har en subjektiv oppfatning om fysikalismen, og subjektive meninger er ikke det samme som objektive fakta.[1]

Selv om vi ikke har gode nok metoder og god nok teknologi i dag til å analysere hvordan hjernen fungerer, så betyr ikke det at vi aldri vil klare å utvikle det i fremtiden. Derfor mener jeg at det Nagel har skrevet, om at det sannsynligvis er umulig å finne en fysisk teori om hele virkeligheten, har svakheter. Ingen kan se inn i fremtiden, og ingen kan derfor vite hva slags teknologi vi vil ha tilgang til. «Not within a thousand years will man ever fly» sa Wilbur Wright i 1901.[2] Han og broren Orville Wright var to luftfartspionerer som allerede i 1903 fullførte den første dokumenterte flygningen i et motorfly. Vi kan derfor ikke si i dag at det i fremtiden vil være umulig å finne en fysisk teori om hele virkeligheten.

Før jeg kan begynne å diskutere om det prinsipielt er mulig at maskiner kan tenke, og om de kan ha en bevissthet, så må jeg først redegjøre for hva tenkning og bevissthet er for noe. Vi vet ikke enda hva tenkning og bevissthet er for noe, men vi har en rekke teorier om hva det kan være. Som sagt tidligere kan tenkning og bevissthet være noe som skjer fysisk inne i hjernen, eller det kan være noe helt separat som kanskje befinner seg i hjernen.

Spørsmålet om at maskiner kan tenke ble drøftet i en artikkel av Alan Turing i 1950. Han tok utgangspunkt i hva som skal til for at vi kan si at en datamaskin tenker, og svaret på dette er at man kan lure et menneske som kommuniserer med en datamaskin til å tro at han kommuniserer med et annet menneske. Dette er den såkalte Turingtesten. Det er ingen

dataprogram som regnes som å ha bestått testen slik Turing beskrev den ifølge Neufeld og Finnestad i artikkelen «In Defense of the Turin Test, AI and Society» fra 2020. Det finnes programmer som kan forveksles med et menneske under spesifikke omstendigheter, som for eksempel Googles Duplex som brukes til å bestille f.eks. frisørtime og bord på restauranter, ALIZA som er et samtaleprogram som kunne kjøre ulike script. Dette er bare noen få eksempler. Turing mente at hvis et dataprogram skulle være intelligent, så måtte det klare å utrykke seg kreativt gjennom naturlige språk. Det vil si at å ikke bare repetere noe som allerede har blitt sagt. Problemløsning er også en kompleks kognitiv evne. Det vakte oppsikt da dataprogrammet Deep Blue slo Garry Kasparov, den regjerende verdensmesteren i sjakk. I tillegg slo AlphaGo verdensmesteren i Go, Lee Se-dol, i 2016. AlphaGo benyttet seg av maskinlæring. Dette beviser ikke at dataprogrammer kan tenke. De hermer bare etter det de har studert fra andres spill.[3]

En svakhet med Turingtesten er at den kun måler skriftlig adferd. I tillegg er det uklart hva testen måler, og om den fanger opp det vi anser som å være sentrale aspekter ved intelligens. Noë mener at intelligens handler om hvordan en organisme forstår og mestres sine omgivelser. Dette ligner det kinesiske rom-argumentet som ble utformet av filosofen John Searle. Han angriper «sterk KI», som er et syn på intelligens og forståelse som han mener preger deler av kognisjonsvitenskapen. I kognisjonsvitenskapen er det to syn på forholdet mellom informasjonsprosessering slik det foregår i datamaskiner, og slik mennesket behandler sansestimuli, dvs. tenker, ifølge Searle. På den ene siden er det folk som mener at datamaskiner er gode hjelpemidler for å forstå hvordan mennesker tenker, og på den andre siden er det de som mener at informasjonsprosessering slik det gjøres i datamaskiner er slik menneskelig tenkning foregår. De tror altså at den eneste forskjellen på en datamaskin og en hjerne er de fysiske materialene de er laget av. Videre hevdes det at en vellykket KI vil ha en psykologi og et sinn på linje med mennesker, og det å hevde noe annet vil være biologisk sjåvinisme.[3]

Searle bruker et tankeeksperiment for å argumentere imot sterk KI. Dette kaller han kineserrommet. Han blir låst inne i et rom med to luker og en bok som inneholder alle mulige setninger på kinesisk. Han får igjennom den ene luka noen kinesiske symboler og får beskjed om å svare tilbake det som står i boka, akkurat som en datamaskin gjør. Men hva om han nå lærer seg boken utenat? Han forstår fortsatt ikke det som blir skrevet på kinesisk, men han ville ha forstått det hvis det kom setninger på engelsk. Det er det å forstå som skiller

mennesker og datamaskiner. Datamaskiner utfører, men de forstår ikke. Det vil derfor være umulig for en datamaskin å ha en bevissthet.[3]

Ifølge Searle vil en sterk KI kun simulere menneskelig forståelse, men på en meget vellykket måte. Han hevder videre at det er den materielle forskjellen mellom hjernen og datamaskinen er helt avgjørende for hva forståelse og tenkning er for noe. Han betraktes dermed som en biologisk sjåvinist. Han baserer seg altså bare på en intuisjon om at det som gjør et menneske til et tenkende vesen er nemlig det biologiske materiale vi er laget av, og at en datamaskin som blir bygget opp akkurat likt som en menneskehjerne, men bare med kunstige materialer, ikke vil kunne gi opphav til forståelse. Margaret Boden mener at Searle ikke har noe empirisk grunnlag for å mene dette. Vi har gode grunner til å anta at våre hjerneceller bidrar til forståelse, men vi vet ikke hvordan de gjør det. Boden mener at riktig og passende informasjonsprosessering og respons kalles forståelse, og dette kan både biologiske systemer og kunstige systemer gjøre. For at vi skal kunne avvise at kunstige materialer kan gi opphav til forståelse, så må vi kunne forklare empirisk hvordan biologiske materialer gjør det. Imidlertid støtter Searle seg på sin intuisjon på at det bare er biologiske materialer som kan gi opphav til forståelse. Boden sier at våre intuisjoner om hvordan verden fungerer har forandret seg i takt med vitenskapelig utvikling og at det derfor er alt for tidlig å avvise at kunstig intelligente systemer kan forstå.[3]

En ytterligere kritikk av Searles tankeeksperiment er at hans versjon av kunstig intelligens systemer er utdatert. Searle så for seg en datamaskin i et innelåst rom, der den eneste kontakten med omverdenen var gjennom arkene som kom inn og ut av rommet. Dagens menneskelignende roboter bruker maskinlæring, språk- og ansiktsgjenkjenning og flere andre teknologier for å fremstå så menneskelignende som mulig. Searle kommer med en utvidelse av kinesisk rom-argumentet for å svare på kritikken mot at hans versjon av KI er utdatert: Altså en robot som beveger seg rundt i et kinesisk miljø, og ser ved hjelp av kameraer, har armer og bein og den kan i tillegg spise og drikke. Den kommuniserer ved hjelp av tekst-til-tale teknologi. Han mener ikke at dette vil bringe noe nytt til argumentet. Det er fortsatt slik at roboten styres av et dataprogram. I stedet for input gjennom symboler på et ark, får den input fra de ulike kameraene, mikrofonene og sensorene. Roboten gjør akkurat det samme som datamaskinen i rommet: den mottar og behandler symboler, og gir output basert på forhåndsbestemte regler. Man kunne like gjerne sittet i kineserrommet og styrt roboten derifra. Searle vektlegger at informasjonsprosessering av data aldri vil bli en forstående relasjon mellom det kunstige systemet og omverdenen, slik det er mellom biologiske

organismer som mottar påvirkning fra omverdenen. Men Bodens argument står fortsatt sterkt fordi han har fortsatt ikke kommet med noen empirisk forklaring på at biologisk materiale gir opphav til forståelse. Like fullt setter Searles argument søkelys på noen viktige aspekter ved det vi tenker på som intelligens. Hvis det da er slik at menneskets bevissthet er noe som er hjernetilstander vil det nok være mulig en dag for en datamaskin å tenke selv og være bevisst. Da trenger vi bare en kraftig nok datamaskin til å simulere alle tilstandene i hjernen. Men, hvis det faktisk er slik at bevisstheten er noe i tillegg til hjernen, vil det nok være prinsipielt umulig for en datamaskin å tenke selv og å ha en egen bevissthet.[3]

Vi vet ikke om det vil være mulig å bygge datamaskiner som forstår i fremtiden, men vi vet at vi ikke er i nærheten enda. Men vi kan ikke si at det vil være umulig med mer sofistikert teknologi. Men kanskje Searle har rett. Kanskje det bare er et spørsmål om biologi. Men vi vet ikke hva som gir opphav til det Searle mener er forståelse, og hvordan et biologisk vesen oppnår dette. I så fall kan vi ikke vite om et kunstig intelligent system ikke kan gjøre det samme.[3]

Jeg har i denne teksten drøftet påstanden til Thomas Nagel om at «argumentene mot en ren fysisk teori om bevissthet er sterke nok til å sannsynliggjøre at en fysisk teori om hele virkeligheten er umulig», og kommet til konklusjonen om at det er for tidlig til å si noe. Jeg har også diskutert om det er mulig for maskiner å tenke og eventuelt om de kan ha en bevissthet og jeg har konkludert med at det igjen er for tidlig til å si noe. Det er mye vi ikke vet enda, men med mer avansert teknologi så kan vi finne ut mer.

Kildeliste:

[1] Nagel, T. (2003). *Problemet med forholdet mellom kropp og sinn*. I T. Nagel, *Hva er meningen? EN kort innføring i filosofi* (s.31-39). Oslo: Libro (Hentet: 19.11.21)

[2] Rik van Hemmen (2013) *Not Within a Thousand Years Will Man Ever Fly. Wilbur Wright 1901)*. Tilgjengelig fra: https://martinottaway.com/rhemmen/not-within-thousand-years-will-man-ever-fly-wilbur-wright-1901/ (Hentet: 19.11.21)

[3] Kiran, A. H. (2021). *Kan maskiner tenke?* NTNU (Hentet: 19.11.21)