Bevissthet og Tenkende Maskiner

Kandidatnr. 30565

ANTALL Ord

15. desember 2021

EXPH0300 - Examen philosophicum for naturvitenskap og teknologi

Innhold

1 Innledning **1**

2 Bevissthetsfilosofi **2**

2.1 Hva er bevissthet? 2

2.2 Nagels dualistiske argument 2

2.3 Argumenter for en fysisk teori om bevissthet 2

3 Kan maskiner tenke? **5**

3.1 Hva vil det si å kunne tenke? 5

3.2 Imitasjonsleken 5

3.3 Hva skiller maskiner fra mennesker? 6

4 Konklusjon **7**

Referanser **8**

1 Innledning

Denne besvarelsen skal ta for seg de to deloppgavene som er beskrevet i den første oppgaveteksten. Den første delen av besvarelsen handler i all hovedsak om Thomas Nagles argumenter mot en ren fysisk teori om bevissthet, og hvorfor disse argumentene muligens ikke holder helt til. Den andre delen av besvarelsen skal ta for seg spørsmålet «*Kan maskiner tenke?*» ved å først prøve å forstå hva det vil si å tenke, og deretter å forstå hva som skiller et tenkende menneske og en tenkende maskin.

Besvarelsen kommer til svare på oppgaven ved å knytte problemstillingene opp mot relevant teori fra pensum og fra andre relevante tekster. Mer spesifikt blir Nagels argumenter og tankeeksperiment knyttet opp mot argumenter for identitetsteori og funksjonalisme, og blir drøftet ut i fra det. Spørsmålet om maskiner kan tenke blir brutt ned i to deler; hva det vil si og tenke, og om bevissthet er en vesentlig del av det å tenke. Det blir knyttet opp til René Descartes argumenter om tvil og tanke, Alan Turings imitasjonslek, og John Searles begrep «*weak AI*».

2 Bevissthetsfilosofi

2.1 Hva er bevissthet?

Bevissthet menes å være i følge filosofer som David Chalmers et av de, hvis ikke det, vanskeligste problemet vitenskapen har å bryne seg på (Chalmers, 1995). En måte å definere hva bevissthet er kan være å se på forholdet mellom bevisstheten og hjernen. Bevisstheten er avhengig av inntrykk den får fra hjernen enten direkte eller i form av sanser fra omverdenen. Bevisstheten er derfor opplevelsen man får av å sanse, tenke eller føle.

2.2 Nagels dualistiske argument

Det er blant annen mange av disse argumentene Thomas Nagel bruker i sin tekst, der han forsøker å se på problemet med forholdet mellom kropp og sinn. I teksten bruker Nagel et tenkt eksempel med en sjokolade, der man tar en bit av sjokoladen og kjenner smaken av denne sjokoladen. Nagel argumenterer med at følelsen av denne smaken er adskilt fra de fysiske prosessene i hjerne. Selv i det helt usannsynlige tilfellet der en person åpnet skallen til den som spiser sjokoladen og smaker på hjernemassen, kunne man umulig smake det samme som den som spiste sjokoladen. Opplevelsen den personen med sjokoladen hadde er unik fordi den ikke bare kan komme av tilstander fra hjernen, men heller noe utenfor kroppen din (Nagel, 2003, s. 32–34).

Nagels ideer om bevissthet minner en del om Descartes' argumenter for dualisme. *Dualisme* handler nettopp om det Nagel framlegger; skillet mellom det fysiske, det som eksisterer i vår verden, og det mentale, det som eksisterer i vår bevissthet. Eller som Descartes kaller det, det *objektive* og det *subjektive* (Dybvig & Dybvig, 2003, s. 156).

2.3 Argumenter for en fysisk teori om bevissthet

Nagel forklarer først at detaljene for fysisk teori om bevissthet er «*opp til vitenskapen å avdekke*» som i seg selv veldig riktig. For å kunne forklare en fysisk teori om bevissthet trengs det at det forskes på, og undersøkes mer om nevrologi og menneskehjernen. Dette er da ikke lengre er et rent filosofisk spørsmål men også et vitenskapelig.

Senere rent benekter Nagel at fysiske hendelser i hjernen kan utgjøre en smaksopplevelse.

"Det er ikke mulig at et stort antall fysiske hendelser i hjernen, uansett hvor kompliserte de måtte være, kan være de komponentene som utgjør smaksopplevelsene våre."
(Nagel, 2003, s. 36)

Begrunnelsen Nagel bruker når han kommer med denne påstanden er:

"vi [må] analysere noe mentalt - ikke noen ytre, observerbar fysisk substans (...) Et fysisk hele kan analyseres i mindre fysiske deler, men en mental prosess kan ikke det, og fysiske deler kan ikke utgjøre et mentalt hele." (Nagel, 2003, s. 36)

Begrunnelsen til Nagel er altså at en fysisk teori om bevissthet ikke kan eksistere fordi bevisstheten, det mentale, er adskilt fra det fysiske. Rettere sagt, den er dualistisk. Her forkaster Nagel tanken om at bevisst kan være fysisk i bunn og grunn på hans egen oppfatning om at den er dualistisk.

Tankeeksperimentet hans kunne også bli brukt til å forklare f.eks. en identitetsteori. *Identitetsteori* går ut på at en fysisk prosess er identisk med en mental tilstand (Hansen, 2021b). Slik f.eks. en datamaskin kan utføre den samme operasjonen, hvis den får repetert den samme kommandoen kontinuerlig.

Her kan vi også se at Nagels eksempel med sjokoladebiten i utgangspunktet ikke gir oss svar på så mye. Vi kan like enkelt tenke oss at når man spiser denne sjokoladebiten, er det fysiske tilstander i hjernen som produserer opplevelsen av smaken. Man kan da slikke så mye man vil på hjernen til den som spiser sjokoladebiten uten å smake sjokoladen, fordi smaken dannes av hjernen gjennom sanseinntrykk fra omverdenen. Spiser man en ny bit av sjokoladen vil den smake likt som den forrige fordi hjernen utfører de samme fysiske tilstandene. Denne identitetsteorien sier dermed at en mental tilstand er ikke noe mer enn en nevral tilstand, som kan sammenliknes med at en tilstand i en datamaskin ikke noe mer enn en binær tallstreng.

Funksjonalisme har også et eksempel i Nagels tankeeksperiment. *Funksjonalisme* handler formelt sett forklarer funksjonalisme mentale tilstander i hjernen ved å fokusere på rollen eller funksjonen de spiller (Hansen, 2021a). For eksempel kan smerte forklares med at den har rollen som en tilstand som sier at noe galt med kroppen, fordi man har blitt påført en skade.

En måte Nagels tankeeksperiment passer med funksjonalismen kan være som dette. Se for deg at du er i samme situasjon som Nagel la fram, og spiser denne sjokoladebiten. Når du spiser denne sjokoladebiten vil det gi deg en mental tilstand av lykke eller godhet,

gitt at du liker smaken, eller vemmelse eller avsky hvis du ikke liker smaken. Hvis en annen person også smakte på en lik sjokoladebit vil den ha samme funksjon og fylle den samme rollen for denne personen. Hos andre organismer vil den samme biten spille en annen rolle, for eksempel for en hund vil den ha en rolle som tilsier at hunder reagerer med vemmelse eller avsky fordi den har en funksjon lik som gift.

Man kan derfor på grunnlag av at identitetsteorien og funksjonalismen passer inn i Thomas Nagles tankeeksperiment si at hans påstand om at argumentene mot en fysisk teori om bevissthet ikke holder helt til, i all hovedsak fordi tankeeksperimentet er veldig vagt og upresist, og ikke greier å forklare en spesifikk isme.

3 Kan maskiner tenke?

3.1 Hva vil det si å kunne tenke?

Hva betyr det å kunne tenke? René Descartes mente at det å tenke var det som definerte at han faktisk eksisterte, «*Je pense, donc je suis*» (Descartes, 1641, s. 30). For han var tanken helt grunnleggende, han hadde overbevist seg selv til å betvile alt annet, han kunne ikke vite om det virkelig eksisterte. Det eneste han visste var at han kunne tvile, tvilen er en tanke, derfor måtte han tenke for å kunne eksistere.

Det store spørsmålet blir da, er vi de eneste som kan tenke? Er vi alene om å være bevisst om at det eksisterer? Dette er store spørsmål som kan være svært vanskelige å besvare. Denne besvarelsen skal kun ta for seg en liten del av det første spørsmålet, kan maskiner tenke?

3.2 Imitasjonsleken

For å kunne si at en maskin kan tenke, må den kunne greie å overbevise mennesker til å tro at den kan tenke. Dette var konklusjonen til matematikeren Alan Turing da han tok for seg spørsmålet «*Kan maskiner tenke?*» i 1950 (Kiran, 2021). Turing mente at selve spørsmålet i seg selv var meningsløst, for Turing var det mer interessant å finne ut som en maskin kunne, når man observerte isolert utenfra, imitere et menneske eller greie å overbevise andre mennesker til å tro at den var et menneske.

Som resultat kom han fram til en test, en imitasjonslek, som gikk ut på at maskinen skulle prøve å fremstå som en menneske. Kort fortalt gikk den ut på at i et rom satt en mann og en dame og i et annet rom satt en undersøker. Undersøkeren kunne kommunisere med de skriftlig, og oppgaven var å finne ut hvem som var mannen og hvem som var kvinnen. Det Turing lurte på da var hva som ville skje hvis mannen eller kvinnen ble byttet ut med en maskin. Ville maskinen være i stand til å imitere mennesket? (Turing, 1950, s. 433–434).

Men vil en slik test som Turing legger frem kunne fungere? Turing anslo selv at innen 50 år ville datamaskinene ha nok minne til at et slik program kunne lages, og at den gjennomsnittlige undersøkeren etter fem minutter ikke klare å peke ut maskinen i mer en 70 % av tilfellene (Turing, 1950, s. 442). Dessverre for Turing hadde han nok bare rett i estimatet om datamaskinens minne, for etter 71 år fra han kom med antydningen har den enda ikke vært et tilfelle der testen har blitt bestått etter hans kriterier (Kiran, 2021, s. 5; Todorović, 2015).

3.3 Hva skiller maskiner fra mennesker?

Et viktig spørsmål å stille seg er om turing-testen i utgangspunktet er en god test. Den svarer nemlig ikke på det som var det opprinnelige spørsmålet, kan maskiner tenke? Som sagt mente Turing spørsmålet var meningsløst, men mye har endret seg på 70 år. I dagens samfunn er kunstig intelligens mye utbredt, og i senere år har selvlærende maskiner blitt veldig populært. En selvlærende maskin, eller maskinlæring, er i bunn og grunn en maskin eller et program som ikke er forhåndsprogrammert til å utføre en bestemt oppgave, men som gjennom å prosessere store mengder med data «lærer» seg hva den skal gjøre (Kiran, 2021, s. 3). Men å snakke om læring i denne sammenhengen er litt missvisende, ihvertfall hvis vi sammenlikner med slik vi mennesker lærer.

Når et menneske skal lære noe nytt, for eksempel en ny ferdighet, et nytt språk eller noe liknende, er man bevisste i handlingen man tar. Man tar selv initiativet til å lære og bevisst på eget ferdighetsnivå og eventuelle forkunnskaper, en maskin er ikke bevisst på samme måten. Når en selvlærende maskin skal lære noe nytt, for eksempel å gjenkjenne et menneske i et bilde, tar den ikke selv initiativet til å finne ut hva den må gjøre for å sette igang. Det er et menneske som må lage programmet som skal sette i gang maskinlærings prosessen, og som må finne fram testdataen som skal brukes til å trene opp maskinen. Maskinen er ikke selv bevisst på hva den skal gjøre slik mennesket er. Dette er et vesentlig skille mellom maskinen og mennesket, og en slik beskrivelse likner det den amerikanske filosofen John Searle kaller «*weak AI*» eller svak KI på norsk (Kiran, 2021, s. 9).

Et annet spørsmål en kan stille seg er da om det å være bevisst i det hele tatt er vesentlig når man lærer. Det høres kanskje absurd ut å ikke være bevisst om egen læring, men mange har muligens erfart å lære noe uten at man gikk inn for det. Ta for eksempel noe så grunnleggende som å gå. De aller fleste mennesker uten noe form for funksjonell nedsettelse kan gå. Vi lærer det som regel ganske tidlig i livet, og før det er krabbing den eneste formen for bevegelse vi kan. Når et barn lærer seg å gå, er det som regel ikke det selv som har tenkt «*idag skal jeg lære meg å stå på to ben å gå*», det er ofte en eller begge av foreldrene som lærer barnet hva det skal gjøre. De løfter barnet på to ben og holder det i hendene mens det febrilsk forsøker å ta sine første skritt. Når foreldrene ikke aktivt lærer barnet, sitter det og observerer hvordan foreldrene går og beveger seg. Dette er litt på samme måte en maskin kan lære seg noe. Den blir implisitt fortalt hva den skal gjøre uten at den egentlig er bevisst over hva det er, og blir «holdt i hånden» slik som barnet. Den blir fortalt hva som er forventet resultat og hva som er forventet å gjøre, slik som når barnet observerer hvordan foreldrene går. Man kan derfor heller si at man muligens ikke må være bevisst for å lære, men at man heller er avhengige av noen som er bevisste som kan fortelle deg hvordan du skal begynne å lære.

4 Konklusjon

Så for å konkludere, hva har besvarelsen kommet fram til? Kort sagt har den kommet fram til at bevisstheten er et stort og vanskelig tema å definere, og at det er veldig relevant når man skal prøve å finne svar på spørsmål som «*kan maskiner tenke?*». Den har også kommet fram til at Thomas Nagels sjokolade-tankeeksperiment ikke er det beste når man skal prøve å avvise argumentene for en fysisk teori om bevissthet.

Videre har den kommet fram til at problemstillingen om å finne ut om maskiner kan tenke kan angripes fra to standpunkter. Man kan enten forsøke å finne ut om maskinene kan imitere mennesker slik som Alan Turing forsøkte, eller man kan prøve å sammenlikne måten mennesker lærer med måten maskiner lærer. Den har også kommet fram til at maskiner ikke er bevisste på samme måte som oss mennesker, men at det ikke nødvendigvis er et hinder for at maskiner kan lære slik som mennesker lærer. Likevel sitter vi her enda og diskuterer rot-spørsmålet «*kan maskiner tenke?*», som for Turing i 1950 ville være utenkelig.

"Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted" (Turing, 1950, s. 442)

Referanser

Chalmers, D. (1995). Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies*. Hentet 5. desember 2021, fra http://consc.net/papers/facing.pdf

Descartes, R. (1641). *Discourse on the Method* (J. Veitch, Overs.). Internet Archive. Hentet 8. desember 2021, fra https://archive.org/details/rmcg0001/

Dybvig, D. D. & Dybvig, M. (2003). *Det tenkende mennesket: Filosofi- og vitenskapshistorie med vitenskapsteori*. tapir akademisk forlag.

Hansen, M. K. (2021a). funksjonalisme. *Store norske leksikon*. Hentet 2. desember 2021, fra https://snl.no/funksjonalisme

Hansen, M. K. (2021b). identitetsteori. *Store norske leksikon*. Hentet 2. desember 2021, fra https://snl.no/identitetsteori

Kiran, A. H. (2021). *Kan maskiner tenke?* NTNU.

Nagel, T. (2003). Problemet med forholdet mellom kropp og sinn. I *Hva er meningen? En kort innføring i filosofi* (s. 31–39). Oslo: Libro.

Todorović, A. (2015). *Has The Turing Test Been Passed?* Hentet 14. desember 2021, fra http://isturingtestpassed.github.io/

Turing, A. M. (1950). I.—Computing Machinery and Intelligence. *Mind*, *59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433