Utfordringer og løsninger innen kunstig intelligens

AKTUELT: **Første helgen i november ble konferansen «*Philosophy & Theory of Artificial Intelligence*» avholdt ved universitetet i Leeds. Blant temaene var bevissthet, maskinlæring og sikkerhet, knyttet til kunstig intelligens.**

Av Solveig N. Selseth

Utviklingen av kunstig intelligens (KI) er i en spennende situasjon for øyeblikket. Vi vet ennå ikke om vi kan klare å lage teknologi med intelligens på menneskelig nivå, men utsiktene er gode. I dag finnes det for eksempel flere androider som er oppsiktsvekkende gode til å imitere mennesker. Av de mest kjente androidene er *Sophia*, som i år ble tildelt statsborgerskap i Saudi Arabia som den første roboten med statsborgerskap i verden.[1] Sophia er hovedsakelig en samtalerobot. Hun reagerer på tiltale med både bevegelser og mimikk, og kan formulere egne svar. Hun er programmert til å forbedre sine reaksjoner og svar basert på alle sine interaksjoner med mennesker, og blir dermed flinkere til å konversere jo flere samtaler hun har.

Sophia er åpenbart designet for å ligne mennesker, så det er lett å anse henne som en av de mest avanserte teknologiene innen KI, men det er flere typer teknologi som er vel så lovende, om ikke mer lovende. Spesielt arbeidet med førerløse biler er kommet veldig langt. Mange bilprodusenter jobber med utvikling av automatisk kjøring og flere selskaper har allerede biler som kan kjøre uten fører.[2] Flere av disse har allerede vært ute på veiene i ulike forsøk, og flere er planlagt.[3] Den perfekte selvkjørende bilen vil kunne ta alle vurderinger som potensielt må tas på en kjøretur. I alt fra hyppige vurderinger om det er klart for forbikjøring, til manøvrering i trange gater med ulike hindringer, til sjeldne situasjoner med navigering i krasjer eller steinras-situasjoner. En del av innføringen av selvkjørende biler blir å rydde opp i og endre veiene slik at man minsker eller eliminerer eksterne farer. Det kan bety forbedring av asfalt, bedre skille mellom fotgjengerområder og kjørefelt, utvetydig merking, osv. Det er allikevel viktig at bilene kan ta gode avgjørelser i alle mulige situasjoner, særlig i perioden før resten av trafikkbildet er optimalisert og de fortsatt deler veiene med førerstyrte biler.

Med slik teknologi melder det seg flere interessante filosofiske problemstillinger. I tilfellet med selvkjørende biler presser etiske spørsmål seg frem: Hvis vi skal programmere inn vurderinger bilen skal ta, hva da med ekstreme tilfeller der bilen må velge mellom to onder? Svinge til venstre og krasje i en motorsyklist, eller til høyre inn i en topp moderne familiebil? Familien i bilen vil ha en større sjans til å overleve krasjen, men det er potensielt snakk om flere liv enn den ene motorsyklisten. Hva slags etikk skal vi gi de selvkjørende bilene? Og hvem skal ta skylden i tilfelle en krasj? Eier av bilen? Bilprodusenten? En av programmererne? Ingen?

I tilfellet med androider som Sophia oppstår det interessante spørsmål om definisjoner av begreper

som «intelligens» og «forståelse». Hva er et «menneskelig intelligensnivå» og hva skal til for at et maskinprogram skal nå det? Hva er relasjonen mellom intelligens og forståelse? Sophia kan formulere egne setninger, men kan hun forstå dem? Hva skal til for at hun skal forstå det hun snakker om? Er det mulig at Sophia en dag kan bli bevisst?

I åpningstalen til konferansen «Philosophy & Theory of Artificial Intelligence» som gikk av stabelen i Leeds 4. og 5. november sa filosof Vincent Müller at feltet kunstig intelligens fortsatt er ganske lite innen filosofifaget, men at det ser ut til å ha fått en oppsving i det siste. Selv om de fleste deltakerne var filosofer, var noen av dem også godt kjent med programmering, informatikk eller hjerneforskning. Selv var jeg til stede for å få en oversikt over hva som skjer innen feltet i dag, og jeg er veldig fornøyd med hva jeg har fått med meg fra konferansen.

Foredragene ble delt opp i fire forelesningsrekker med temaene begreper, utfordringer, etikk og metoder. Dag 1 bestod av forelesningsrekkene «begreper?» og «utfordringer», og dag 2 av «etikk?» og «metode?», hvor sistnevnte tok for seg metoder innen blant annet problemer og løsninger innen programmeringen av KI.

Det første foredraget ble holdt av Mark Sprevak som snur spørsmålet om kunstig intelligens på hodet og diskuterer den biologiske hjernen som maskin. I foredraget diskuterte han en nyttig måte å forstå hjernens evne til å vurdere sannsynlighet. Susan Schneider derimot er opptatt av *ulikhetene* mellom hjernen og maskiner. Hennes foredrag tok for seg substansuavhengighet og spørsmålet om maskinbevissthet er mulig. I denne teksten vil jeg gå nærmere inn på disse foredragene, og til slutt vil jeg nevne flere av utfordringene og noen av løsningsforslagene som ble fremmet på konferansen.

LES OGSÅ: Det teknologiske menneske • Humanioras utfordringer i det digitale landskapet

Hjernen som maskin

Mark Sprevak diskuterte muligheten for hjernen som sannsynlighetsinferensmaskin (*probabilistic inference machine*). Tanken er at hjernen kan ha en sannsynlighetsrepresentasjon av informasjon, istedenfor at vi representerer enkeltstående tilstander i verden med 0 eller 100% sikkerhet, som Sprevak beskriver som den tradisjonelle teorien. Forslaget er at hjernen har mulighet til å representere flere tilstander i verden med sannsynlighet 1 distribuert over dem. La oss si vi har en representasjon som sier at enten *x* eller *y* er tilfellet. Hvis *x* tilskrives sannsynlighet 0,2, så tilskrives *y* 0,8, slik at deres samlede sannsynlighet er 1.

Sprevak hevder at denne sannsynligheten fastsettes baseres på ulik *tiltro* (*credence*) til alternativene. Med tiltro innføres også et usikkerhetsmoment: Noen ting blir representert til oss med en viss grad av usikkerhet. Når vi har lav tiltro til noe – det er usikkert – tildeler vi det en lavere grad av sannsynlighet, mens de alternativene som er sikrere har en høyere sannsynlighet. Anser vi hjernen som en sannsynlighetsinferens-maskin, er det nærliggende å tenke at det er *neural circuits*

eller enkeltstående neuroner som holder disse sannsynlighetene, ifølge Sprevak.

Så, hvordan bestemmes et subjekts tiltro? Mulige alternativer kan være veddemål eller preferansestrukturer, men Sprevak avviser begge disse. Neuroner foretar seg ikke veddemål, for eksempel. Sprevak foreslår heller David Lewis' teori om at tiltro representeres av en sannsynlighetsfunksjon som tilhører et *utility* funksjon/sannsynlighetsfunksjonpar som kan rasjonalisere atferd på best mulig måte. Det vil si, tiltro spiller en overordnet rolle i den handlenes beslutningstaking. Hennes atferd bestemmes av hennes tiltro.

På denne måten unngår man å hevde at neuroner foretar seg for eksempel veddemål. En av hakene er at tiltro nødvendigvis er en egenskap som tilhører intensjonelle, fullt ut rasjonelle aktører, ifølge Lewis' teori. Altså, hvis det er neuroner som tilskrives tiltro, får man et prekært problem. Tar man et realistisk syn på teorien, blir man nemlig nødt til å forplikte seg til at neuronene selv er fullt ut rasjonelle aktører. Dette må være feil, og Sprevak foreslår at vi istedenfor tar et instrumentalistisk? syn på teorien. Standpunktet blir dermed at det er *nyttig* for oss å anse neuroner som subjekter for tiltro, og at vi derfor burde holde teorien.

Sprevaks foredrag baseres på et pågående arbeid, og han bemerker i ettertid at flere av kommentarene han fikk i løpet av konferansen i Leeds vil være behjelpelige i den videre utarbeidelsen. En av kommentarene til foredraget pekte blant annet på at teorien forutsetter at sannsynlighetsvurderingen sitter i neuronene, noe som ikke er ukontroversielt.

Alt dette er eksemplifiserer hvordan kunstig intelligente maskiner kan gi oss innsikt i hvordan menneskelig kognisjon fungerer. Når vi utforsker kunstig intelligens, både filosofisk og teknisk, baserer vi ofte våre begreper på menneskelig intelligens. Men man kan også snu det rundt og se på den biologiske hjernen som en maskin, og finne likheter mellom hvordan hjernen fungerer og hvordan maskiner fungerer. Dette er en metode som strekker seg tilbake til Alan Turing, som ønsket å finne ut hvilke matematiske beregninger et menneske kunne utføre. For dette utviklet han det som vi kaller Turingmaskinen – en hypotetisk maskin som med bare fem ulike typer handling skal kunne utføre alle de matematiske beregningene som mennesker kan.[4] Antagelsen om at *computation* er den relevante egenskapen som hjernen har for å produsere intelligens har vært en av de mest fruktbare innen kognitiv vitenskap, ifølge Sprevak.[5]

Det er også noen som mener at den biologiske hjernen og en maskin aldri vil kunne fungere helt likt, og ha de samme kognitive egenskapene. Susan Schneider for eksempel, nevner muligheten for at den biologiske hjernen og maskiner har ulikheter som leder til en ganske viktig forskjell: Hjernen kan være bevisst, det kan muligens aldri en maskin bli.

LES OGSÅ: Å samtale med maskiner • Bevisstheten – et mysterium?

Maskinbevissthet

Schneider innledet konferansens andre dag med et veldig engasjerende foredrag med

maskinbevissthet som tema. Schneider påpeker at hvis bevissthet ikke er substansuavhengig (*subrstate independent*), det vil si uavhengig av materiale, og maskinbevissthet dermed ikke er mulig, så vil forbedring av hjernen via maskinimplantater i form av for eksempel computerchips føre til tap av bevissthet. Tanken bak «hjerneforbedring» (*brain enhancement*) er at man skal kunne gi hjernen ekstra prosesseringsevne ved å implanterte datachiper i forskjellige deler av hjernen. Målet er at chipen skal gi ekstra prosesseringsevne til å gjøre spesifikke oppgaver. Men Scheiders argument er at hvis disse chipene aldri kan bli bevisste – eller ta del i bevisstheten – så vil hjernen bestå av en potensielt stor mengde fullstendig ubevisst prosesseringskapasitet. Eller verre; hvis en av chipene skulle føre til at personen mistet synet, for eksempel, så har man et tilfelle der prosesseringskapasiteten har økt, men bevisstheten har minsket i omfang. Legger man til store mengder chiper som en etter en tar bort litt av bevisstheten, kan man ende opp med en hjerne med potensielt massiv kapasitet, men uten noen bevissthet.

Schneider nevner i denne sammenheng et av Elon Musks neste prosjekter – Neuralink – som skal jobbe for nettopp slik hjerneforbedring.[6] Musks formål er å forbedre menneskets egen hjernekapasitet slik at vi blant annet har bedre sjanse til å holde en virkelig kunstig intelligens i sjakk, når vi eventuelt skulle klare å produsere en. Musk er en av flere som har uttalt frykt for at en kunstig intelligens skal kunne utvikle seg ut av kontroll og føre til store skader på, eller ren utryddelse av, menneskeheten. Schneider problematiserer allikevel Musks kontrollforslag. Skulle det vise seg at hjerneforbedring fører til tap av bevissthet, så blir vi selv en del av problemet. En overmenneskelig intelligens uten bevissthet er potensielt enda skumlere enn en bevisst en.

Schneider tilstår selv ha en viss bekymring over skrekkscenarioer hvor maskiner tar over og utrydder menneskeheten, særlig etter å ha lest Nick Boströms (ganske så) pessimistiske bok *Superintelligence* (2014). Boströms mål er å fremme viktigheten av å utforme sikkerhetstiltak i god tid før vi utvikler kunstig intelligens på et menneskelig eller høyere nivå. I boken skisserer han et slags «verstefall-scenario» der den kunstige intelligensen har blitt gitt som mål å maksimere produksjonen av binders, og dermed fortsetter å lage binders ut av alt tilgjengelig materiale – jordkloden, solsystemet og etterhvert også resten av universet.[7] Hvordan vi skal sikre at slike negative utfall ikke vil skje kalles *kontrollproblemet*.

Schneider påpeker også at eksperimenter med hjerneforbedring potensielt kan gi empirisk resultat som svekker teorien om substansuavhengighet. Hvis en pasient mister synet når en chip blir satt inn, kan man ta chipen ut og forbedre den med mål om at pasienten skal kunne beholde synet samtidig som chipen er implantert. Mister pasienten synet igjen når den forbedrede chipen er satt inn, må man forsøke nye forbedringer. Hvis man har forbedret og forbedret chipen, operasjonen og dens plassering, og pasienten fortsatt ikke har syn når chipen er inne, er det en indikator på at bevissthet ikke er uavhengig av materiale. Men slike forsøk er selvsagt ikke frie for etiske problemstillinger.

Utfordringer

Hvilke kognitive egenskaper må med for at noe skal anses å tilsvare intelligens på menneskelig nivå? Hvilke mennesker skal tas med i utregningen? Hvis nivået settes til et slags gjennomsnittsmenneske, betyr det at mennesker som ligger under gjennomsnittet ikke selv har intelligens på menneskelig nivå, og vice versa? Disse spørsmålene stilte José Hernández-Orallo i sitt foredrag, som også drøftet hvilke utfordringer som må løses når det kommer til testing av intelligens.

Andre utfordringer ble fremmet av blant andre Chin Chuan Fei som påpeker problemene vi vil få med å fastslå om en maskin er bevisst eller ikke. Han trekker frem Commander Data fra tv-serien Star Trek – The Next Generation som eksempel: Data er en funksjonell kopi av et menneske. Dermed oppfører han seg akkurat som et menneske i forskjellige situasjoner. Med våre atferdsbaserte bevissthetstester vil han nødvendigvis fastslås som bevisst, selv om han ikke skulle være det.

Mens Chan Fei påpeker vanskelighetene med å fastslå maskinbevissthet, problematiserer Schlomo Danziger også hvordan man fastslår maskinintelligens. Danziger diskuterer nemlig muligheten for at samfunnet ikke vil akseptere at begrepet «intelligent» brukes om maskiner. Danzigers foredrag tok for seg Alan Turings tekst «Computing Machinery and Intelligence» (1950)[8]. Spørsmålet er om maskiner kan nå et menneskelig intelligensnivå, og hvordan vi som samfunn forholder oss til dette spørsmålet. Danziger påpeker først at Turings kjente imitasjonsspill (*imitation game*) som utarbeides i artikkelen ikke først og fremst har som formål å være en test for intelligens. Spillet tester etter en type *atferd* som man vil forvente av et intelligent vesen. Måleenheten er selvfølgelig menneskelig atferd. Selve spillet dreier seg om å vurdere hvilken av spillerne som er menneske basert på hvem av dem som har mest menneskelig oppførsel.

Danziger fremmer Diane Proudfoots tolkning av «Computing Machinery and Intelligence», hvor Turing anses å hevde at vi mennesker er sjåvinistiske mot maskiner og deres evne til å være intelligente. Tolkningen hevder at vi mennesker, eller samfunnet, uansett aldri vil godta maskiner som intelligente, uavhengig av deres testresultater i imitasjonsspillet. Intelligens tilskrives et subjekt bare når samfunnet er villige til å tilskrive det. Dermed er spørsmålet om en maskin kan være intelligent meningsløst, forklarer Danziger. Danziger påpeker at en maskin som hypotetisk sett skulle «bestå» imitasjonsspillet, hvor bestå vil si at den utviser atferd man vil forvente av et intelligent vesen, allikevel ikke vil anses som intelligent av samfunnet, nettopp fordi det er en maskin, og ikke et menneske.

Selv om Danzigers foredrag gav et tankevekkende blikk på begrepene intelligens og maskinintelligens, er jeg skeptisk til at denne sjåvinismen mot maskinintelligens har en dyp rot i samfunnet. Jeg hører ofte begrepet intelligens brukes om maskiner med ulikeevner og kapasitet. Det virker i alle fall ikke som at mange har problemer med å kalle Sophia og dataprogrammet AlphaGo for intelligente. Men det man legger i begrepet intelligens i disse tilfellene, er åpenbart ikke det samme som den menneskelige intelligensen Turing snakker om. Det er heller en begrenset type intelligens som ikke kan brukes til å løse oppgaver i nye, ukjente situasjoner.

Menneskers intelligens kan generaliseres: Vi kan bruke vår intelligens til å navigere oss gjennom ukjente problemer i helt andre typer situasjoner enn vi har opplevd tidligere. Maskiner er enn så lenge bundet til avgrensede områder. Kan vi en dag klare å gi maskiner den samme generelle intelligensen som mennesker? Er begrenset intelligens og generell intelligens to forskjellige typer intelligens? Kan man nå generell intelligens ved å bygge på begrenset intelligens?

Sander Beckers advarer mot en annen type urett vi potensielt kan begå mot maskiner, hvis vi skulle klare å produsere en overmenneskelig kunstig intelligens. Beckers argument er at med superintelligens kommer også super-smerte?, og at det er galt av oss å produsere en intelligens med potensiale til en smerte som vil være ubegripelig enorm. Det er selve intensjonen om å produsere en superintelligens som stiller oss klanderverdige, hvis noe skulle skje og den kunstige intelligensen skulle ende i en supersmertefull tilstand. Men her dukker også spørsmålet om intelligens nødvendigvis fører med seg qualia[9] og følelser som smerte opp. Det er også kontroversielt om superintelligens vil føre med seg super-smerte, og ikke bare smerte på menneskelig nivå. Er super-smerte i det hele tatt mulig?

Michael Prinzing er mer optimistisk. Hans foredrag fremmer en mulig løsning på *kontrollproblemet.* Ifølge Prinzing, ligger løsningen på problemet i å gi den kunstige intelligensen tilgang til alle tekster vi har skrevet om kjærlighet og deretter gi den som mål å elske alle (i like høy grad). Når den kunstige intelligensen har lært seg hva kjærlighet er, vil den ta de valgene som er til det beste for alle mennesker – hvem vil vel skade dem man elsker? Noen utfordringer med dette er å bestemme hvem som skal tas med i «alle» – alle mennesker? Alle dyr? Noen mennesker står bak voldsom smerte for store folkegrupper; hvordan skal den kunstige intelligensen forholde seg til dette? Det er også problematisk hvordan man skal sikre at den lærer akkurat det kjærlighetsbegrepet som vi ønsker at den skal lære. Menneskelig kjærlighet fører ofte til en mengde irrasjonelle og uønskede situasjoner, som det ikke ville være gunstig å overføre til en overmenneskelig intelligens.

Veien videre

Konferansen viser at kunstig intelligens er et tema som trenger utarbeidelse på tvers av feltene filosofi dekker. Etikk er et åpenbart område: Flere filosofer mener det er helt avgjørende at vi løser kontrollproblemet i god tid før vi produserer kunstig intelligens på menneskelig nivå eller høyere. Men det virker for meg som at man ikke har kommet veldig langt i utarbeidingen av en god løsning. Jeg kan i alle fall ikke si at det ble presentert noen påfallende god løsning i løpet konferansen. Innen etikken er det også viktig å diskutere lovverket rundt autonome biler og våpen, blant annet. I disse tilfellene har vi allerede teknologien, men vi bør oppdatere lovverket før man benytter seg av den på et stort nivå.

Det er også usikkerhet rundt definisjonene av begrepene vi bruker. Hvordan skal vi teste intelligens hvis vi ikke er klare på hva det er, og hvordan det henger sammen med forståelse, bevissthet, læring, osv.? Som vi ser fra Schneiders foredrag er også spørsmålet om substratavhengighet

innen bevissthetsfilosofi kanskje mer interessant enn tidligere. Innen arbeidet med kunstig intelligens kan man muligens ta steget fra hypotetiske diskusjoner til det langt mer praktisk håndfaste.

LES OGSÅ: «Maskinisme» og destratifisering• Matematikk er en kampkunst

Litteratur

Boström, N. 2014, *Superintelligence. Paths, Dangers, Strategies*. Oxford University Press.

Copeland, J. Bowen, M. Sprevak & R. Wilson (Eds.) *The Turing Guide: Life, Work, Legacy*. Oxford: Oxford University Press.

Noter

[1] http://www.bbc.com/news/blogs-trending-41761856 [29.11.2017]
https://www.nrk.no/nyheter/robot-ble-saudiarabisk-statsborger-1.13771620 [29.11.2017]

[2] https://waymo.com/ [29.11.2017]
http://www.keolisnorthamerica.com/ [29.11.2017]

[3] Forsøk med selvkjørende buss i Las Vegas: https://www.theverge.com/2017/11/6/16614388/las-vegas-self-driving-shuttle-navya-keolis-aaa [18.12.2017]
Forsøk med selvkjørende busser i Danmark: http://nyheder.tv2.dk/samfund/2017-06-30-busser-uden-chauffoer-testet-i-aalborg-naeste-aar-bliver-det-hverdag [18.12.2017]
Forsøk med selvkjørede drosje i Singapore: https://www.bloomberg.com/news/articles/2016-08-25/world-s-first-self-driving-taxis-debut-in-singapore [18.12.2017]

[4] Copeland 2017, side 281

[5] Copeland 2017, side 278.

[6] https://www.neuralink.com/ [18.12.2017]
https://www.digitaltrends.com/cool-tech/neuralink-elon-musk/ [18.12.2017]

[7] Bostrom 2014, side 150.

[8] Turing, A. 1950, «Computing Machinery and Intelligence» i*Mind, 59*, 433-460.

[9] Qualia er et begrep som skal innebefatte alt av bevisste opplevelser, som for eksempel den indre opplevelsen av å se en rød bil, eller å høre en trompet.