Kan maskiner tenke?

av Asle H. Kiran

Innhold

1 Kunstig intelligens 2

1.1 Bør vi frykte ultraintelligente maskiner? 2

1.2 Kunstig intelligens er både teknologi og filosofi 3

2 Tenkende maskiner 4

2.1 Turingtesten 4

2.2 Har noen dataprogrammer bestått Turingtesten? 5

2.3 Er det å simulere intelligens det som samme som å være intelligent? 6

3 Det kinesiske rom: Kan kunstig intelligente systemer *forstå*? 8

3.1 Hva er det Turingtesten egentlig tester? 8

2.5 Er Searles argument bare en intuisjon? 10

2.6 Utenfor det kinesiske rom 11

Kilder 14

1 Kunstig intelligens

I et intervju på festivalen *South by Southwest* i 2017 sa den humanoide roboten *Sophia* at hun vil ødelegge menneskeheten (Weller, 2017). Bør vi nå være redde for en fremtid befolket med roboter? Kanskje ikke pga. Sophia. Hun ønsket bare å være imøtekommende og hjelpsom. Dvs., hun *ønsket* ikke. Sophia kjører på en algoritme som bestemmer det. Så da intervjueren spurte «skal du ødelegge menneskeheten?» svarte hun bare så hyggelig hun kunne. Hadde hun *forstått* spørsmålet, så hadde hun nok (forhåpentligvis) svart annerledes.

Kan kunstig intelligens, dvs. et dataprogram som f.eks. en robot kjører på, *forstå*? Nå, eller en gang i fremtiden? Eller er det tilstrekkelig at det kunstige systemet oppfører seg *som om* det forstår for at vi kaller det intelligent? Mange frykter en fremtid med avansert kunstig intelligens. Har vi en grunn til å frykte det? For å ta stilling til et slikt spørsmål, må vi vite hva det egentlig betyr når vi sier at et kunstig system, en teknologi, er intelligent.

1.1 Bør vi frykte ultraintelligente maskiner?

I forrige kapittel så vi at det kan være vanskelig å ha full kontroll med hvordan teknologi påvirker samfunnet. Allikevel er det slik at all teknologi i utgangspunktet er avhengig av tilsyn og vedlikehold, og ikke minst, tilførsel av energi. Vi kan derfor velge å slå av teknologiene hvis vi skulle ønske det, selv om det vil komplisere livene våre.

I dag står vi imidlertid overfor en teknologi som potensielt kan endre dette bildet: kunstig intelligente systemer som kan lære. Dette er teknologi som mange frykter kan utgjøre en trussel mot menneskeheten. Dette fordi kunstig intelligens (KI) muligens vil ha kapasitet til å mangfoldiggjøre seg selv, og dermed utvikle – og forbedre – seg utenfor menneskets kontroll.

Allerede i 1966 beskrev informatikeren Irving Good (1916-2009) det han kalte for ultraintelligente maskiner. En slik maskin vil være mer intelligent enn mennesker, og kan selv designe nye, og enda mer intelligente maskiner. Dette vil være starten på en *intelligens-eksplosjon* som menneskeheten ikke vil kunne ta del i (Good, 1966, s. 33). I dag omtales punktet der maskinene blir mer intelligente enn mennesket som *singulariteten*, og er av enkelte datofestet til år 2045 (Kurzweil, 2005).

Singularitet er et kontroversielt begrep. Ikke alle er overbevist om at et slikt punkt vil oppstå – til det er det for store forskjeller på hva biologisk intelligens er og hva teknologisk intelligens er. Det er heller ingen enighet om hva en eventuell singularitet vil bety for menneskeheten. Good selv var forsiktig optimist, og *håpet* at de ultraintelligente maskinene ville la mennesket beholde kontrollen.

I dag er den gjennomgående tonen i diskusjonen mer pessimistisk. Fysikeren Stephen Hawking (1942-2018) er en av de som har advart mot hva fremtidens KI kan innebære. KI *kan* være det beste som har skjedd menneskeheten, men han frykter at det i stedet vil bety slutten på menneskeheten, og tar til orde for en forsiktig utvikling av KI (Cellan-Jones, 2014). Premisset for Hawkings frykt er at menneskets intelligens er knyttet til den trege biologiske evolusjonen, mens den teknologiske evolusjonen som driver kunstig intelligens fremover er langt raskere. Som Good, mener altså Hawking at KI-systemer, intelligensmessig sett, vil legge menneskeheten bak seg.

Andre er mindre engstelige for fremtidens KI. Filosofen Alva Noë mener at tanken om en intelligenseksplosjon bygger på en misforståelse av begrepet intelligens. KI er informasjonsteknologi (IT), og det all IT gjør er i bunn og grunn bare regelstyrt informasjonsprosessering, om enn kompleks og lynraskt. Det gjelder enten reglene er programmert av mennesker, eller generert ved maskinlæring (når et dataprogram skanner store datamengder og klarer å skille ut mønstre (Tidemann & Elster, 2019)). Menneskelig intelligens handler derimot ikke om prosessering, men om kreativitet og mening. Vi skaper vår egen tilværelse gjennom å forholde oss til og tilpasse våre omgivelser til de behov, ønsker og preferanser vi har. Det er derfor kvalitative forskjeller på biologiske systemer som det menneskelige, og teknologiske systemer – forskjeller som ikke vil bli utlignet av at informasjon blir prosessert enda kjappere og mer effektivt enn hva dagens KI-systemer kan gjøre.

Noë sammenligner KI med en klokke: Klokken *holder* tiden, men den *vet* ikke hva klokka er. Et KI-system vil aldri kunne være intelligent i seg selv, men ved å utvikle KI, så gjør vi *oss selv* mer intelligente (2014). KI er med andre ord en måte vi skaper nye betingelser for våre liv, slik mennesket har gjort gjennom all teknologi. Frykten for at vi skal bli etterlatt i en intelligenseksplosjon bygger slik sett på det som Noë hevder er en misvisende forestilling om at menneskelig intelligens kun utvikles gjennom biologisk evolusjon.

1.2 Kunstig intelligens er både teknologi og filosofi

Noë betrakter KI som et *filosofisk problem*, og tar stilling til spørsmål som: Hva er grensen mellom biologisk og kunstig intelligens? Hva skal til for at vi er villige til å si at det kunstige systemet er like, eller mer, intelligent enn det biologiske? Eller, hvorfor vil det kunstige systemet aldri kunne kalles intelligent på samme måte som vi omtaler biologisk intelligens.

Det aller meste av forskningen på og utviklingen av KI er imidlertid ikke opptatt av slike filosofiske spørsmål. KI er et praktisk og teknologisk felt, der man bygger systemer som kan utføre bestemte oppgaver, enten på egenhånd (autonome systemer) eller ved å assistere mennesker. Eksempler er:

selvkjørende biler (også vanlige biler har mye KI i seg), *MetaOptima* (som «ser etter» hudkreft ved å skanne en persons hud), *Siri* og *Alexa* (Apple og Amazons digitale assistenter), *Neurensic* (som overvåker og analyserer utviklinger i aksjemarkedet), og det meste som har fått prefixet *smart*: smarttelefoner, smartklokker, smarthus etc. Dagens KI-systemer er riktignok ganske spesialiserte, og langt unna det som Good kalte ultraintelligens – som i dag omtales som kunstig *generell* intelligens (KGI).

Mange spesialiserte KI-systemer er både maskinlærende og automatiserte, og kommer slik sett med noen etiske utfordringer. Ett av dem er utfordringen med *kontroll*: hvordan kan slike systemer, når de lærer og er automatiserte, fortsette å støtte oss på et sett som vi ønsker? I 2016 lanserte Microsoft *Tay*, en maskinlærende chatbot. I løpet av noen få timer snudde Tay seg fra å være en hyggelig og imøtekommende samtalepartner til å bli en rasistisk og kvinnefiendtlig en. Årsaken var at Twitter-brukere fôret chatboten med bemerkninger i denne retningen, som den da lærte av (Hunt, 2016). Hvordan kan vi kontrollere en vellykket KGI når vi ikke en gang har kontroll med hvordan en chatbot utvikler seg?

De etiske utfordringene knyttet til at flere og flere av teknologiene vi bruker til daglig er et KI-system eller består av KI-komponenter er selvfølgelig svært viktig. Imidlertid, en nærmere redegjørelse og diskusjon av disse må vente. Her skal vi se nærmere på det filosofiske spørsmålet om KI, dvs. om den menneskelige intelligensen omfatter egenskaper som et kunstig system i prinsippet ikke kan ha. Når vi skal bestemme hvordan forskning på og utvikling av kunstig intelligens skal foregå i fremtiden, så må vi legge til grunn en forståelse av grensene for KI og på hvilke sett kunstig intelligens er forskjellig fra naturlig intelligens. *Det* er hva det filosofiske spørsmålet om KI hjelper oss med.

2 Tenkende maskiner

2.1 Turingtesten

I en artikkel fra 1950 spurte den britiske matematiker og kryptografen Alan Turing (1912-1954): «kan maskiner tenke?» Turing mente riktignok at spørsmålet ikke ga mening i 1950, men at datateknologien ville ha endret seg så mye innen år 2000 at spørsmålet da ville være rimelig. Slik sett spurte han egentlig om: hva skal til for at vi vil si at en datamaskin tenker? Eller, hvordan skal en datamaskin oppføre seg, slik at vi vil si at den kan tenke? Svaret han kom opp med var at en slik maskin – som han kaller det – må kunne lure et menneske som kommuniserer med den til å tro at hen kommuniserer med et annet menneske.

Turing tenkte ikke her på en robot med menneskelignende fremtoning og snakkeevne. I stedet ser han for seg en datamaskin – vi vil kanskje heller kalle det et dataprogram, som kommuniserer skriftlig gjennom å svare på konkrete spørsmål. Dersom programmet svarer overbevisende nok på spørsmålene til at utspørreren tror hen får svar fra et annet menneske, så mener Turing at vi må si at datamaskinen tenker. En slik utspørring har i ettertid blitt hetende *Turingtesten*.

Mer konkret beskrives testen slik: Deltagerne er to personer og en datamaskin. De er alle skjult fra hverandre. Den ene personen stiller en rekke spørsmål, og svarene kommer enten fra den andre personen eller fra datamaskinen. Dersom spørsmålsstilleren etter fem minutters utspørring ikke klarer å skille svarene fra datamaskinen fra svarene fra den andre personen i mer enn 70% av tilfellene, så har datamaskinen bestått testen (Turing, 1950, s. 442).

Spørsmålene må være spesifikke og velformulerte, men kan hentes fra all type menneskelig aktivitet. Maskin og menneske må stille på så like premisser som mulig. Etter spørsmål om matematiske oppgaver eller om et gunstig sjakktrekk må datamaskinen derfor programmeres til å svare med en tilstrekkelig pause, ellers vil tempoet i svargivingen røpe den. På spørsmål der mennesket har en fordel fremfor datamaskinen, f.eks. om å skrive et kort dikt, må mennesket i sin tur tilpasse svarene for ikke å røpe for mye.

2.2 Har noen dataprogrammer bestått Turingtesten?

Ingen dataprogrammer besto testen innen år 2000, men en chatbot kalt *Eugene Goostman* klarte det – muligens – i 2014, i en konkurranse arrangert av Universitetet i Reading. Dette var imidlertid ikke en fullverdig Turingtest fordi chatboten utga seg for å være en 13-åring fra Ukraina. Slik kunne litt forvirrende og misvisende svar (på engelsk) lett bortforklares. Det er fortsatt ingen dataprogrammer som regnes som å ha bestått testen test slik Turing beskrev den (Neufeld & Finnestad, 2020).

Men det finnes programmer og chatbots som kan forveksles med et menneske under mer spesifikke omstendigheter. Googles *Duplex* er et programtillegg til Google Assistant som kan brukes for å bestille time hos frisøren og for å bestille bord på restauranter. Programmet høres ut som en virkelig stemme (slik mange digitale assistenter gjør i dag), og «krydrer» samtalen med tenkepauser og nølende «umm» og «hmm» for å høres mer realistisk ut (Vincent, 2018). Skulle tematikken dreie seg bort fra bestilling av time/bord, så vil Duplex røpe seg ganske fort, så noen Turingtest er det lite trolig at programmet vil bestå.

Et tidlig dataprogram som brukte naturlig språk, men kun skriftlig, var ELIZA, som ble utviklet på midten av 1960-tallet av informatikeren Joseph Weizenbaum (1923-2008), som var ansatt på Massachusetts Institute of Technology (MIT). ELIZA var et samtale-program som kunne kjøre ulike

scripts. Det mest kjente scriptet var DOCTOR, som simulerte en psykolog. Litt som Sophia, responderte ELIZA/DOCTOR ut fra nøkkelord i setninger som en bruker skrev inn. Svarene var som oftest spørsmål, som skulle få brukeren til å utdype det som ble fortalt, men det kunne også være noen standardfraser for å få samtalen videre. Hvis man skrev inn «alle menn er like», så kunne programmet svare: «hvordan da?». Eller hvis man skrev «jeg er så deprimert», så kunne svaret bli «det er leit å høre at du er deprimert». Eller: «jeg kommer ikke overens med min mor», hvorpå ELIZA/DOCTOR kunne si: «fortell meg mer om familien din». (Weizenbaum, 1976, ss. 3-4)

DOCTOR var en parodi på samtaleterapi der det aldri blir gitt tydelige svar eller konkrete innspill til pasientene. Til Weizenbaums store overraskelse, så satte folk god pris på å snakke (skrive) med programmet, selv når de var klar over at det bare var et dataprogram. Weizenbaum mente at dette var et klart tegn på at folk snakket med programmet som om de hadde en samtale med en virkelig person. Det kan slik sett kanskje virke som ELIZA har bestått Turingtesten. Det er allikevel like trolig at effekten kom *på grunn av* at folk visste at de snakket med et dataprogram i stedet for en person.

Test selv! Versjoner av ELIZA/DOCTOR kan lett finnes på internett.

Programmet *Project Debater* som er utviklet av IBM er en langt mer avansert, og maskinlærende versjon av ELIZA (som var programmert). Dette programmet kan debattere, noe det bla. gjorde i 2019 om førskoleutdanning (i USA) bør være subsidiert med offentlige midler. Programmet finner frem til, og det fremfører egne argumenter (det må da fôres med nyheter, leksikaoppslag og artikler på et tema), og kan også motargumentere. Til tross for sofistikert teknologi og opptreden, så er ikke programmet et forsøk på en generell intelligens. Det er en samling spesialiserte KI som virker sammen, en såkalt *kompositt KI*. Det muliggjør en mer kompleks atferd enn hva spesialiserte KI-systemer, som ansiktsgjenkjenning, bok- og filmanbefalinger etc. kan gjøre. F.eks. bruker programmet IBMs «*Watson*» til å omdanne meddebattantenes tale til skrift, som så blir grunnlaget for å finne motargumenter, og det er separate moduler for konstruksjon av argumenter og for konstruksjon av motargumenter, mm. (Slonim, et al., 2021).

2.3 Er det å simulere intelligens det som samme som å være intelligent?

Det som har vært utfordringen for KI-systemer siden 1950-tallet har vært hvor gode de er på oppgaver som regnes som det ypperste den menneskelige intelligensen har å by. Descartes mente i sin tid at det er mye en maskin kan gjøre, men det som vi kaller høyere kognitive evner, kunne man ikke få dem til å gjøre. Meningsfull bruk av språk er iallfall utelukket, ifølge Descartes (1985 [1637], s. 140). Dette er bakgrunnen for Turings tanke om at dersom et dataprogram klarer å uttrykke seg

kreativt gjennom naturlig språk – altså, ikke bare repeterer noe som har blitt sagt, så vil vi innrømme den iallfall noe intelligens (Copeland, 1993, s. 38).

Rasjonelle, målrettede handlinger som krever langsiktig planlegging (f.eks. problemløsning) er også en krevende og kompleks kognitiv evne. Det er noe vi gjør hver eneste dag uten å reflektere over kompleksiteten. I sin mest rendyrkede form vises denne evnen i spill som sjakk og Go. Av den grunn har utviklingen av KI-systemer som kan vinne over de beste menneskelige spillerne vært regnet som en målestokk på hvor langt KI-teknologien har kommet.

Det var derfor en sensasjon da *Deep Blue* slo den regjerende verdensmesteren Garry Kasparov i 1997. Kasparov mente riktignok at spillet tydet på menneskelig innblanding, men dette ble tilbakevist av IBM, som sto bak programmet. Det var også oppsiktsvekkende da *AlphaGo* slo verdensmesteren i Go, Lee Se-dol, i 2016. Reglene i Go er enklere enn sjakk, men antallet mulige trekk er langt større (10^{360} mot «bare» 10^{123} i sjakk), noe som selvsagt er svært krevende å programmere et dataprogram til å vurdere før hvert trekk. Derfor benytter *AlphaGO* seg av maskinlæring. Programmet ble trent opp ved å «studere» 30 millioner variasjoner fra 160 000 faktiske spill (Koch, 2016). Slik Deep Blue var en målestokk i forhold til hva programmerbar KI kan gjøre, så er *AlphaGO* en milepæl i forhold til hva maskinlæring kan gjøre på KI-feltet.

Men hva så, om et dataprogram skulle bestå Turingtesten, bestiller frisørtime, eller klarer å slå en verdensmester i Go, vil vi da si at programmet «tenker»? At slik atferd indikerer tenkning, eller at systemet som utfører den har en form for intelligens, er omstridt. Men det er også uklart om det å stryke på testen kan si oss noe definitivt om *mangel* på intelligens (Copeland, 1993, s. 44). Tenk på hvordan sjimpanser og gorillaer samhandler, planlegger handlinger og bruker verktøy. Eller hvordan delfiner hjelper mennesker. Er dette dyr uten intelligens? Ingen av dem ville iallfall kunne bestå en Turingtest. Betyr det at de ikke tenker? De tenker iallfall ikke ved bruk av menneskelig språk, men å si at de av den grunn ikke er intelligente bygger på en snever forståelse av intelligens.

Men heller ikke det å bestå testen, om et system skulle klare det, kan egentlig si oss noe sikkert om dataprogrammet tenker, eller er intelligent. Om noe *simulerer* en egenskap eller evne, betyr det nødvendigvis at det har eller er den egenskapen eller evnen? Et smykke med gullmaling, er jo ikke et gullsmykke. En skuespiller som spiller død, er jo ikke død.

Å si at kunstig simulering av intelligens ikke er intelligens kan være et uttrykk for biologisk forutinntatthet. Hvis et system av ikke-biologiske materialer som silisium, aluminium etc. skulle lykkes fullt ut, i alle henseende, med å opptre intelligent, hvorfor skulle det ikke kunne kalles intelligent? Vi bedømmer andre menneskers intelligens ut fra det de sier og oppfører seg – vi kan jo

ikke se inn i hjernen deres og faktisk se intelligensen. Hva er det da som gjør at vi ikke vil innrømme dataprogrammet intelligens, mens vi f.eks. er villig til å innrømme dyr intelligens, til tross for de ikke kan ytre seg i et naturlig menneskespråk?

En mulig faktor her er at Turingtesten kun måler skriftlig atferd. Det er en ganske snever form for atferd, om enn avansert nok til at ingen dataprogram ennå har bestått testen (når tematikken er mer generell enn bestilling av frisørtimer). Allikevel kan det i noen tilfeller være tilstrekkelig; tidligere nevnte Stephen Hawking ytret seg de siste årene av sitt liv kun gjennom skrift. Denne skriften ble oversatt til tale av et dataprogram, og de fleste er kjent med hans robotaktige stemme. Hawking fikk denne på midten av 80-tallet og nektet siden å bytte den ut, selv om tekst-til-tale teknologi har blitt sterkt forbedret siden da.

3 Det kinesiske rom: Kan kunstig intelligente systemer *forstå*?

3.1 Hva er det Turingtesten egentlig tester?

En annen grunn til å være motvillig til å kalle et dataprogram intelligent – selv om det skulle bestå Turingtesten – er at det er uklart hva testen måler. Fanger den egentlig opp det vi anser for å være de mest sentrale aspektene ved intelligens? Som vi så lenger opp, Noë mener at intelligens handler om hvordan en organisme forstår og mestrer sine omgivelser. Slik sett er det andre krav til hva som gjør en blekksprut intelligent og hva som gjør et menneske intelligent. I begge tilfellene handler det om hvordan en organisme forholder seg til omgivelsene, og aktivt og kreativt skaper den verden den lever i. Et slikt forhold har ikke f.eks. en klokke til tiden. En klokke stykker opp tiden og viser oss, gjennom tall eller visere, hva klokka er. Men den *vet* ikke, som vi så, hva klokka er. For Noë er en datamaskin «bare» en avansert klokke.

Noës argument er en variant av det *kinesiske rom-argumentet*, som har blitt utformet av filosofen John Searle. Searle argumenterer ikke primært mot Turingtesten; hans skyteskive er det han kaller «*sterk KI*», et syn på intelligens og forståelse han mener preger deler av kognisjonsvitenskapen (et fagfelt hvor psykologi, filosofi, lingvistikk, nevrovitenskap og informatikk kommer sammen i studiet av kognisjon og, mer generelt, det menneskelige sinnet) (Searle, 1980).

Ifølge Searle, er det i kognisjonsvitenskapen to syn på forholdet mellom informasjonsprosessering slik den foregår i datamaskiner og slik mennesket behandler sansestimuli, dvs. tenker. På den ene siden, de som mener at datamaskinen og dataprogrammer er gode *hjelpemidler* for å forstå hvordan mennesker tenker. I det ligger det at forskning på KI, og mer generelt, på data, gir oss verdifull innsikt i menneskelig tenkning: En abstrakt beskrivelse av hvordan en datamaskin prosesserer informasjon

ligner på en abstrakt beskrivelse av hvordan den menneskelige hjernen behandler sansestimuli og annen informasjon fra kropp og omgivelser. Det materielle underlaget er forskjellig (silisium, metall osv. i datamaskinen; nerveceller, aksoner, myelin osv. i hjernen), men strukturen ligner. Dette synet kaller Searle for «*svak KI*», og er et syn han ikke har noen problemer med.

Sterk KI, på den andre side, innebærer at informasjonsprosessering slik den gjøres av datasystemer *er* slik menneskelig tenkning foregår. Den materielle forskjellen på et kunstig system – en datamaskin – og et biologisk system – en hjerne – er underordnet. En hjerne er en biologisk datamaskin. Menneskelig intelligens er bare å prosessere et sansestimuli og gi en passende respons, f.eks. å dukke når du ser en fotball komme i full fart mot hodet ditt, eller si «hei» når noen hilser på deg. Konsekvensen av dette synet er at hvis vi mennesker tenker, forstår og har andre psykologiske tilstander, så kan også en datamaskin ha det, bare den blir like kompleks som den menneskelige hjernen. Mao., en vellykket KI vil ha en psykologi og et sinn på linje med mennesker, dersom sterk KI sitt syn på menneskelige intelligens er riktig. Å si noe annet, vil være *biologisk sjåvinisme*.

Searle bruker et tankeeksperiment som ramme for å argumentere mot sterk KI. Searle ser for seg at han er låst inne i et rom som har to luker. Gjennom den ene luken kommer det inn ark med kinesiske tegn. Searle kan ikke kinesisk og vet ikke om det er kinesisk, japansk eller bare noen meningsløse symboler. Searle har fått instruksjoner om å slå opp i en bok, gjenfinne symbolene han får, og deretter skrive noen andre symboler, som er spesifisert i boka, på et nytt ark. Deretter skal han sende dette arket ut den andre luka. Ettersom Searle ikke kan kinesisk, så vet han ikke at det første arket inneholder noen spørsmål og at det han har skrevet ned på det andre arket er svar på disse spørsmålene. Boken han har brukt inneholder alle mulige kinesiske setninger og spørsmål, med tilhørende riktige responser. Han kan dermed svare på alle spørsmål på kinesisk uten å kunne et ord kinesisk.

Searle ber oss så forestille oss at han også får spørsmål på engelsk, hans morsmål, inn den ene luka og at han svarer på disse på et ark som sendes ut den andre luka. Her trenger han ikke å slå opp i en bok, men kan svare ut fra det han allerede vet. For å gjøre disse to responsene enda mer lik, så kan vi se for oss at Searle lærer seg den kinesiske boken utenat, slik at han ikke trenger å slå opp i den for hvert ark som kommer inn, men kan respondere ut fra hukommelsen.

Er det noen forskjell på det Searle gjør på engelsk og det han gjør på kinesisk? Searle sier at han i det engelske tilfellet *forstår* spørsmålene og svarer ut ifra det, mens han i det siste tilfellet, også etter at han har memorert alle symbolene, fortsatt ikke forstår kinesisk. Når Searle leser og skriver kinesiske

tegn oppfører han seg akkurat som en datamaskin: han manipulerer input, i form av symboler, ut fra bestemte instruksjoner og regler, og gir passende output, i form av andre symboler.

Hvis et dataprogram skulle bestå Turingtesten, så vil det være beskrivelsen av programmets «atferd». Programmet ville tatt imot spørsmålene, prosessert disse ut fra hvordan det har blitt programmert eller maskinlært, og gitt svar som utspørreren ville tatt for å være svar fra et menneske. Men kan vi si at programmet dermed har *forstått* spørsmålene? I følge Searle, så nei, det vil gå på tvers av det vi tenker at forståelse er. Det er nemlig en forskjell på kognisjon og forståelse: En datamaskin *utfører*, men forstår ikke. Og derfor kan vi heller ikke si at en datamaskin – uansett hvor godt den vil simulere et menneske skriftlig – har et sinn.

2.5 Er Searles argument bare en intuisjon?

Ifølge sterk KI simulerer ikke bare datamaskinen menneskelig forståelse: en suksessfull simulering *er* det den simulerer. Utover at hjernen og datamaskinen er laget av ulike materialer, kan vi ikke skille den forståelsen som en datamaskin oppviser fra den menneskelige dersom datamaskinen simulerer den menneskelige på en suksessfull måte. Forståelse kan nemlig bare måles ut fra det å gi en passende respons til et gitt input/sansestimuli, og gjør en datamaskin dette, ja, da forstår den. Derfor kan vi også se på menneskelig kognisjon og forståelse som å kjøre et dataprogram med hjernen i rollen som datamaskin. I motsetning til et dataprogram som er programmert av mennesker (eller, som mange av dagens KI, maskinlært), så er vi mennesker programmerte gjennom den naturlig evolusjon (og vanlig læring).

Searle hevder derimot at den materielle forskjellen er helt avgjørende for hva forståelse, og dermed tenkning, er. Det er ikke fordi han er et dataprogram at han tenker og forstår, sier Searle, men fordi han er en biologisk organisme, og er som andre biologiske organismer derfor i stand til å forstå, handle, lære osv. (Searle, 1980, s. 422). Searle er med andre ord en biologisk sjåvinist (og stolt av det). Searle sammenligner forholdet mellom hjerneprosesser og forståelse med forholdet mellom klorofyll og fotosyntese. En datasimulering av prosessene i fotosyntesen vil ikke produsere karbohydrater, det vil produsere en datasimulering av karbohydrater.

Men nettopp det er også en grunn til å spørre hva det kinesiske rom-argumentet egentlig sier oss. Searle baserer seg her nemlig på en intuisjon, en *magefølelse*, om at biologiske materialer, hvis de oppfører seg på et bestemt vis, gir opphav til forståelse, mens kunstige materialer, når de strukturelt sett oppfører seg likedan ikke vil gi opphav til forståelse.

Hvilket empirisk grunnlag har Searle for å sette opp denne distinksjonen? Ingen, ifølge filosofen og kognisjonsforskeren Margaret Boden. På mange vis er også vår kropp og hjerne et

informasjonsprosesserende system. Forståelse innebærer å være aktivt engasjert i de omgivelsene vi er i. Enten vi leker, handler, jobber, spiller fotball, klatrer i Mount Everest, eller kommuniserer med andre personer, tar vi imot inntrykk og informasjon, innser hva det innebærer for aktiviteten vi holder på med, og gir en passende respons. Men selv om vi har gode grunner til å anta at våre hjerneceller bidrar til forståelsen, så har vi ingen aning om *hvordan* de gjør det. For Boden er det derfor like mystisk at riktig og passende informasjonsprosessering og respons i et biologisk system skal kalles forståelse, som at det kalles det når et kunstig system gjør det samme (Boden, 2004, s. 256).

Boden peker på at den vitenskapelige forståelsen av fotosyntese er svært forskjellig fra den vitenskapelige forståelsen av sammenhengen mellom hjerne og forståelse. Vi kjenner til og kan peke på produktene av fotosyntese, vi vet hvordan de blir produsert og hvilke betingelser som skal til for at fotosyntesen skal fungere. Det er ikke tilfellet med forholdet mellom hjerne og forståelse: vi vet svært lite om hvordan forståelse, slik Searle bruker begrepet, oppstår. Egentlig vet vi også lite om hva det er, og skal vi peke på det, ja, da må vi peke på handlinger, ord o.l., altså *produktene* av forståelse, ikke forståelsen i seg selv.

Et minstekrav til en avvisning av at kunstige materialer som silisium og metall kan gi opphav til forståelse er derfor at den er ledsaget av en empirisk forklaring av hvordan biologiske materialer kan gjøre det. Og det har ikke Searle (eller noen andre), han støtter seg på en intuitiv forskjell på biologiske og ikke-biologiske materialer. Boden viser til at våre intuisjoner om hva som er riktig, har forandret seg i takt med den vitenskapelige utviklingen. At klorofyll spiller en rolle i fotosyntesen er selvfølgelig for oss nå, men var ikke det for et par hundre år siden. Følgelig er det prematurt å avvise at kunstig intelligente systemer kan forstå (Boden, 2004, s. 257).

2.6 Utenfor det kinesiske rom

Et annet ankepunkt mot Searles tankeeksperiment er at hans versjon av KI-systemer er utdatert. Status på KI i dag er noe helt annet enn det var i 1980. Searles metafor for en datamaskin var altså at han satt innelåst i et rom. Den eneste kontakten han hadde med omverden var gjennom arkene som kom inn den ene luka og arkene han sendte ut. Og det eneste han forholdt seg til var den ferdigskrevne, dvs. programmerte, oppslagsboka. Er dette et godt bilde på f.eks. dagens og morgendagens humanoide roboter, slike som tidligere nevnte *Sophia* og indiske *Rashmi* (som har egen twitterkonto, og muligens blir den første humanoide roboten med oppdrag i verdensrommet (Deogharia, 2019)). Disse menneskelignende robotene bruker maskinlæring, språk- og ansiktsgjenkjenning og en rekke andre teknologier for å fremstå som så menneskelignende som

mulig. Slike roboter er, og vil etter hvert være *i verden* i en helt annen grad enn Searle-i-rommet. Gjør det noe forskjell for Searles argument?

Searle foregriper dette i artikkelen sin, og ser for seg at en robot som kan bevege seg rundt (i et kinesisk miljø), «se» ved hjelp av kameraer, og har armer og bein som gjør at den kan gå, hamre inn spiker, og til og med spise og drikke. Vi kan også se for oss at en slik robot er utstyrt med tekst-til-tale teknologi som gjør at den snakker kinesisk. Roboten kan ikke skjelnes fra et biologisk, kinesisktalende menneske.

Searle ser ikke at denne utvidelsen av tankeeksperimentet bringer inn noe vesentlig nytt. Det er fortsatt slik at roboten styres av et dataprogram. Forskjellen er at roboten nå ikke bare får input gjennom symboler på et ark, men fra kameraer, mikrofoner og andre sensorer i robotkroppen. Denne informasjon må allikevel «oversettes» til symboler for at den styrende datamaskinen skal kunne prosessere den. En slik robot, slik Searle ser det, vil fortsatt være en datamaskin, selv om det er en som beveger seg rundt istedenfor å være fastlåst i et rom.

Det Searle så ber oss om å gjøre er å tenke han ut av det kinesiske rommet og inn i «hjernen» på roboten. Istedenfor å motta ark fra utsiden gjennom en luke, vil han sitte her og motta symboler fra kameraet, fra armene og beina, og magen, og istedenfor å sende ark ut en luke, så fôrer han beina, armene etc. med symboler som så bli omgjort til handlinger. Robot-Searle gjør akkurat det samme som Searle-i-rommet: mottar og behandler symboler, slår opp i boka, gir output basert på regler. Robot-Searle er ikke mer i dette kinesiske miljøet enn det Searle-i-rommet var. Den kontakten han har med omverden er fortsatt gjennom symboler, og han kunne slik sett like gjerne fortsatt sittet i det kinesiske rommet og styrt robotkroppen gjennom trådløs overføring. Robot-Searle kan derfor styre roboten suksessfull, uten å forstå noe som helst av robotens omgivelser (Searle, 1980, s. 420).

Dersom en robot utseendemessig, muntlig og i atferd er til å forveksle med et menneske (slik som f.eks. robotene i HBO-serien *Westworld*), så vil det uansett være slik at datamaskinen/roboten kun utfører, den forstår ikke. Searles poeng er at informasjonsprosessering av data aldri vil bli en *forstående relasjon* mellom det kunstige systemet og omverden, slik relasjonen er mellom en biologisk organisme som mottar sansestimuli fra omverden.

Searles tilbakevisning av forstående roboter løser allikevel ikke problemet med hans biologiske sjåvinisme – så lenge vi ikke har en empirisk tilfredsstillende forklaring på forståelse hos biologiske organismer, så kan vi ikke, som Boden argumenterer, avvise at fremtidens KI kan sies å forstå. Selv om maskinlærende, snakkende og humanoide roboter er svært avansert teknologi, ut fra dagens standard, så er slike systemer allikevel langt unna den kompleksiteten som finnes i den menneskelige

hjernen, med sine milliarder av nerveceller, synapser og nevrale forbindelser. Det kan være grunn nok til å være forsiktig med å avvise hvordan et kunstig system som nærmer seg en slik kompleksitet forholder seg til omverden. Kanskje er det kompleksiteten, og ikke biologien, som er betingelsen for forståelse? Hvorvidt det er mulig å lage et så komplekst kunstig system er et annet spørsmål.

Like fullt, det kinesiske rom-argumentet setter søkelys på et viktig aspekt ved det vi tenker på som intelligens, eller tenkning for å bruke Turings begrep, nemlig det å forstå. Og beslektet: bevissthet og opplevelsen av å være et bevisst vesen, som for oss mennesker også er uavlatelig knyttet til tenkning og intelligens. Vi vet ikke om vi i fremtiden vil kunne bygge datamaskiner, dataprogrammer eller roboter som forstår. Vi vet at vi ikke har gjort det så langt – og at vi ikke er i nærheten av å gjøre det, men kan vi av den grunn være sikker på at det ikke vil kunne skje med mer sofistikert teknologi, mer komplekse former for maskinlæring osv.?

På den andre side, vi vet heller ikke om kompleksitet er nok. Muligens har Searle rett, at det er et spørsmål om biologi. Men om han har grunnlag for å avvise på et prinsipielt grunnlag at «maskiner kan tenke» nå, det er mer tvilsomt. Det koker ned til at vi ikke vet hva som gir opphav til det som Searle kaller forståelse. Vi vet ikke hvordan et biologisk system gjør det, og da kan vi heller ikke vite at et kunstig system *ikke* gjør det.

Kilder

Boden, M. A. (2004). Escaping from the Chinese Room. I J. Heil (Red.), *Philosophy of Mind. A Guide and Anthology* (ss. 253-266). Oxford: Oxford University Press.

Cellan-Jones, R. (2014). *Stephen Hawking warns artificial intelligence could end mankind*. Hentet 11.06.2021 fra BBC.com: https://www.bbc.com/news/technology-30290540

Copeland, J. (1993). *Artificial Intelligence. A Philosophical Introduction.* Oxford: Blackwell Publishing.

Deogharia, J. (2019). *Isro keen to send humanoid Rashmi on a space mission.* Hentet 11.06.2021 fra The Times of India: https://timesofindia.indiatimes.com/city/ranchi/isro-keen-to-send-humanoid-rashmi-on-a-space-mission/articleshow/67605639.cms

Descartes, R. (1985 [1637]). Discourse on the Method. I R. Descartes, *The Philosophical Writings of Descartes, vol. 1* (ss. 111-151). Cambridge: Cambridge University Press.

Good, I. J. (1966). Speculations Concerning the First Ultraintelligent Machine. *Advances in Computers, 6*, ss. 31-88. doi:10.1016/S0065-2458(08)60418-0

Hunt, E. (2016). Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. *The Guardian*. Hentet 11.06.2021 fra https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter?CMP=twt_a-technology_b-gdntech

Koch, C. (2016). How the Computer Beat the Go Master. *Scientific American*. Hentet 11.06.2021 fra https://www.scientificamerican.com/article/how-the-computer-beat-the-go-master/

Kurzweil, R. (2005). *The Singularity Is Near: When Humans Transcend Biology.* New York: Penguin Viking.

Neufeld, E., & Finnestad, S. (2020). In defense of the Turing test. *AI & Society, 35*, ss. 819-827. doi:10.1007%2Fs00146-020-00946-8

Noë, A. (2014). *Artificial Intelligence, Really, Is Pseudo-Intelligence.* Hentet 11.06.2021 fra NPR.org: https://www.npr.org/sections/13.7/2014/11/21/365753466/artificial-intelligence-really-is-pseudo-intelligence

Searle, J. R. (1980). Minds, brains, and programs. *The Behavioral and brain sciences, 3*(3), ss. 417-424. doi:10.1017/S0140525X00005756

Slonim, N., Bilu, Y., Alzate, C., Bar-Haim, R., Bogin, B., & Bonin, F. (2021). An autonomous debating system. *Nature*(591), ss. 379-384. doi:10.1038/s41586-021-03215-w

Tidemann, A., & Elster, A. (2019). *Maskinlæring*. Hentet 11.06.2021 fra snl.no: https://snl.no/maskinl%C3%A6ring

Turing, A. M. (1950). Computing Machinery and Intelligence. *MInd. A quarterly review of Psychology and Philosophy, 59*, ss. 433-460.

Vincent, J. (2018). *Google's AI sounds like a human on the phone - should we be worried?* Hentet 11.06.2021 fra The Verge: https://www.theverge.com/2018/5/9/17334658/google-ai-phone-call-assistant-duplex-ethical-social-implications

Weizenbaum, J. (1976). *Computer Power and Human Reason.*

Weller, C. (2017). *Meet the first-ever robot citizen — a humanoid named Sophia that once said it would 'destroy humans'.* Hentet 11.06.2021 fra Businessinsider.com: https://www.businessinsider.com/meet-the-first-robot-citizen-sophia-animatronic-humanoid-2017-10?r=US&IR=T