Eksamensopgave Informationsvidenskab
Bacheloruddannelsen - Studieordningen 2005

INSTITUT FOR
INFORMATIONS- OG
MEDIEVIDENSKAB

Lærer: Finn Olesen Afleveringsdato: 16.01.2007

Censor *(udfyldes af sekretariatet)*:

Intro. til programmering	☐	Organisationsanalyse	☐
Programmering og systemudvikling	☐	Studium generale	■
Kommunikation 1	☐	Informationsvidenskabelig metode	☐
Kommunikation 2	☐	Valgfrit projekt	☐
Digital æstetik	☐	Bachelorprojekt	☐
Teknologihistorie	☐		

Opgavens titel:

Kunstig intelligens og at forstå kinesisk

Afleveret af:

Årskort: 20051132 Navn: Lasse Knud Damgaard E-mail: lasse.damgaard@gmail.com

Årskort: Navn: E-mail:

Årskort: Navn: E-mail:

Årskort: Navn: E-mail:

KUNSTIG INTELLIGENS
OG
AT FORSTÅ KINESISK

Lasse Knud Damgaard

Eksamen i videnskabsteori/studium generale, 2006
Institut for Informations- og Medievidenskab
Aarhus Universitet

Indholdsfortegnelse

Opgaven indeholder ca. 35.953 tegn, inklusive mellemrum.

Antal er eksklusive indholdsfortegnelse, litteraturliste, fodnoter og referencer.

Indledning

Kunstig intelligens dækker i dag alt fra modstandere i computerspil, over aktiehandel og informationssøgning, til deciderede forsøg på at efterligne den menneskelige bevidsthed. Det er det sidste område, der er emnet for denne opgave, og det er i den forbindelse værd at overveje, hvad det er der imiteres. Intelligens, bevidsthed, følelser og tænkning er blot nogle af de begreber, der uden nogen klar definition bruges til at betegne det særegne produkt af menneskets hjerne. Det er ikke formålet med denne opgave at redegøre for denne forvirring. I stedet vil 'kunstig intelligens' (herefter AI)[1] dække over *alle* disse begreber, og dermed betegne forsøget på fuldt ud at efterligne samtlige aspekter ved menneskets mentale indhold i computeren.

Måske netop på grund af uklarhed omkring begreber blev forskning i AI ved sin fødsel spået enorme og snarlige succeser. De er indtil videre udeblevet, og computere evner kun at udføre funktioner tilsvarende meget begrænsede dele af menneskets mentale kræfter. Eksternalisering af sådanne funktioner er imidlertid betydningsfuld og med til at redefinere computerens rolle. Derfor er AI interessant for det informationsvidenskabelige studie og arbejdet i spændingsfeltet mellem mennesker og computere.

En mulig total eksternalisering, altså det der i opgaven betegnes 'kunstig intelligens', vil naturligvis fundamentalt ændre computerens rolle og definition, men har også filosofiske og videnskabsteoretiske implikationer. Muligheden for AI udfordrer vores selvforståelse og rejser fundamentale spørgsmål om, hvordan viden om bevidsthedens mange facetter kan opnås. Lige så interessante er perspektiverne i, at det måske er umuligt at frembringe AI. Netop dette hævder den engelske filosof John Rogers Searle. I mere end 25 år har Searle argumenteret herfor, særligt via det tankeeksperiment han har døbt 'det kinesiske rum', og som i The Stanford Encyclopedia of Philosophy kaldes "et af de mest foruroligende og tankevækkende bidrag til denne litteratur" (Horst, 2004).

Det er således opgavens formål, at undersøge og diskutere John Searles argumenter imod muligheden for AI.

Opgavens disposition og metode

Forståelsen af Searles argumenter vil tage udgangspunkt i den historiske baggrund for AI forskning. Her vil fokus være på Alan Turings banebrydende arbejde i 1950'erne med grundlæggende koncepter og lovmæssigheder inden

[1] Af engelsk 'artificial intelligence'.

for computerteknologi, samt de forestillinger om computerens muligheder for intelligens som blev gjort i den forbindelse. Dette vil blive relateret til det primære paradigme inden for AI forskningen, der resulterede af Turings arbejde, og som motiverer Searles diskussion af emnet.

Undersøgelsen af Searles argumenter vil tage udgangspunkt i hans berømte tankeeksperiment om 'det kinesiske rum'. Argumentets grundlæggende præmisser vil blive kritisk behandlet individuelt, og relateret til de antagelser inden for AI forskning der argumenteres imod, hvorefter Searles konklusioner opsummeres.

En efterfølgende diskussion vil se på nogle af de generelle kritikker, der kan rejses mod Searles argumentation, samt det mest almindelige deciderede modargument.

Citater

For at gøre læsning af opgaven mere flydende er engelske citater oversat til dansk. Hvor der er fundet ambivalens i oversættelse angives, den original engelske formulering i fodnote. Der er kun benyttet original kursivering.

Kunstig intelligens

Ideen om tænkende eller bevidste maskiner går længere tilbage end det tyvende århundrede og computerens opfindelse. Den østrigske matematiker og filosof Gottfried Leibniz argumenterede fx i værket *Monadologie* fra 1714 imod en materialistisk (mekanisk) forklaring af sansning og bevidstheden, ved netop at forestille sig en maskine der kunne tænke, føle og sanse. Leibniz konkluderede, at uanset maskinens kompleksitet ville sådanne mentale fænomener aldrig kunne resultere af interaktion mellem maskinens mekaniske dele eller disses individuelle egenskaber. De måtte i stedet forklares metafysisk (Carlin & Kulstad, 2004).[2]

Naturvidenskabens succes med i de følgende århundreder at forklare fænomener, der tidligere blev tilskrevet sådan metafysik, kan antages at have været medvirkende til, at idéen om intelligent maskineri levede videre. Samtidig udvikledes i starten af 1800-tallet de første mekaniske computere, fx englænderen Charles Babbages 'difference engine' der kunne udregne forskellige matematiske funktioner. De omfangsrige muligheder blev dog først klarlagt, da de grundlæggende koncepter blev undersøgt teoretisk, fx af

[2] Som et kuriosum kan nævnes, at Leibniz definerede det moderne binære talsystem som det benyttes i dag, og således lagde en af computerteknologiens grundsten.

matematikeren Alan Turing, en af pionererne inden for datalogi[3], der i 1936 beskrev principperne i en abstrakt, universel computer[4] (Copeland, 2006).

Papir maskiner

Det teoretiske arbejde illustrerede computerens muligheder og potentielle kræfter, hvilket ledte Turing til i 1948 at beskrive, hvordan manipulation af symboler på basis af formelle regler konceptuelt kunne fremstå som intelligent handling. Turing forestillede sig en skakspillende 'papir maskine', der består af et menneske, som følger en række instruktioner for håndtering af modstanderens træk, nedskrevet i almindelige sætninger på papir. Med disse instrukser kan mennesket behandle input som fx 'Qg5' og fremstille passende modtræk af samme symbolske form som output (Cole, 2004). Dette gøres ved udelukkende at eksekvere "formelle manipulationer af abstrakte symboler på grundlag af eksplicitte formelle regler", hvilket er definitionen af computation (Wackerhausen, 1989, p. 121).[5] Manipulationen er her omdannelsen af modstanderens træk (abstrakte symboler) til egne modtræk ud fra papirinstruksernes eksplicitte regler, og den udføres ved, at mennesket udelukkende forholder sig syntaktisk (formelt) til symbolerne, uden (nødvendigvis) at vide det fjerneste om skak eller at symbolerne overhovedet repræsenterer træk i et spil skak. Pointen er selvfølgelig, at det ikke desto mindre kan fremstå som om, at papir maskinen, som er det samlede system af menneske og instrukser, udviser intelligent handling i skakspillet.

Turing testen

Tankeeksperimenter som dette foranledigede Turing til at undersøge, hvordan det generelt kunne afgøres, om en maskine - og mere specifikt en digital computer – er i stand til at tænke: "Kan maskiner tænke?" (Turing, 1950, p. 433). For at besvare dette foreslår han en test, således at spørgsmålet

---

[3] Jeg bruger det danske begreb 'datalogi', til at betegne det der på engelsk kaldes 'computer science', selv om begrebet har visse konnotationer som den engelske ækvivalent er foruden og vice versa.

[4] Uformelt kan en 'Turing maskine' beskrives som en universel symbolmanipulerende tilstandsmaskine. Den er uafhængig af konkret fysisk realisering og alle Turing maskiner kan udføre de samme operationer, med hastighed som den eneste forskel. En konkret realisering, som er den alle moderne computere er baseret på, er fx 'von Neumann arkitekturen', opkaldt efter den østrigsk/ungarsk/amerikanske matematiker John von Neumann. Den er opdelt i regneenhed, lager og kontrol og opererer sekventielt.

[5] I resten af opgaven betegner 'computation' dette koncept, da det danske 'beregning' ikke har de samme nyttige konnotationer.

omformuleres til, hvorvidt en computer kan gennemføre testen et statistisk overbevisende antal gange. Turing testen udgøres af en "imitations leg", der er bestået, hvis computeren kan foregive at være et menneske. Dette afgøres ved, at en interviewer, der kommunikerer skriftligt med henholdsvis et menneske og en computer, ikke med spørgsmål kan afgøre, hvem der er hvad. Det er udelukkende deltagernes evne til handling gennem sproglige ytringer, der afgør hvorvidt de er tænkende eller bevidste. Turing har altså erstattet det oprindelige spørgsmål, med spørgsmålet om hvorvidt en computer kan udføre arbitrære sproglige handlinger af en kompleksitet tilsvarende menneskets. Testens værdi er således baseret på antagelsen af, at "intet kunne muligt bestå Turing testen ved at vinde Imitations Legen uden at være i stand til at udføre uendeligt mange andre klart intelligente handlinger" (Dennet & Hofstadter, 1981, p. 93). Havde legen i stedet bestået i at skelne en menneskelig modstander fra en computer i to spil skak, ville den således ikke have været af megen værdi, da IBM's Deep Blue computer som bekendt slog den daværende verdensmester Garry Kasparov, uden at udføre nogen som helst andre handlinger der kunne tilskrives intelligens. Man kan også som Searle (1997, p. 58) påpege, at Deep Blues evne til at spille skak ikke har nogen funktionel lighed med menneskelig intelligens, og blot benytter "rå regnekraft" til at slå modstanderen ved at kunne regne på millioner af træk hvert sekund, hvilket ifølge Searle er "totalt anderledes end tanke processerne hos en virkelig spiller." Som det vil blive beskrevet senere, er Searles primære argument dog, at computeren, ligesom den menneskelige operatør i Turings 'papir maskine', kun forholder sig syntaktisk til skakspillets symboler, som altså ikke for computeren har noget semantisk indhold. Tilsvarende indvendinger kan gøres mod gyldigheden af Turing testen; succesfuld deltagelse i imitations legen er ikke nogen garanti for computerens bevidsthed og indre, kvalitative følelser. Turing (1950, p. 446) henviser til, at mennesker normalt vælger at tage hinandens bevidsthed for givet, hvilket ikke i sig selv udgør et argument for testen gyldighed. Dog mener Turing (ibid.), at den i praksis bruges til at "[...] opdage hvorvidt nogen virkelig forstår noget [...]".

Hvordan en computer bringes til at bestå testen og dermed antageligt have ægte forståelse, er der intet konkret bud på, men Turing (1950, p. 455) mener, at "problemet hovedsageligt vedrører programmering" og hans "[...] håb er, at der er så lidt mekanisme i barne-hjernen, at noget som den nemt kan programmeres" (Turing, 1950, p. 456). Det antages altså ikke blot, at computere har muligheden for bevidsthed og intelligent handling gennem formel manipulation af abstrakte symboler, men også at menneskets intelligens kan

forstås på denne måde: ”intelligent opførsel består antageligt i en afvigelse fra den fuldstændig disciplinerede opførsel brugt i computation, men en temmelig lille en […]” (Turing, 1950, p. 459).[6]

Fysiske symbol systemer

Idéerne, som Turing giver udtryk for i 1950’erne, ligger til grund for den retning inden for AI forskning, der betegnes symbolsk eller funktionalistisk, og som udgjorde det altdominerende paradigme frem til starten af 1980’erne (Dreyfus & Dreyfus, 1988, p. 34). Funktionalisme dækker den antagelse, at bevidsthed og andre mentale fænomener skyldes abstrakte kausale strukturer, der kan repræsenteres som regelstyret manipulation af symboler, altså computation (Chalmers, 1992, p. 1).[7] Det antages altså, at der på det rette abstraktionsniveau kan findes en fælles funktionel beskrivelse af den menneskelige hjerne og computere (Dreyfus & Dreyfus, 1988, p. 16). Dette formuleres af pionererne Allen Newell og Herbert Simon i hypotesen om fysiske symbol systemer, der siger, at ”et fysisk symbol system har de nødvendige og tilstrækkelige midler til generel intelligent handling”, hvilket skal forstås som værende af samme omfang som menneskelig intelligens, sådan at ”i enhver situation kan opførsel, der er passende for systemets mål og adaptiv til miljøets krav, opstå” (Newell & Simon, 1976, p. 116). Foruden denne forståelse af intelligens indebærer hypotesen også, at ”[…] ethvert system der fremviser generel intelligent handling, ved analyse vil kunne fastslås at være et fysisk symbol system” (ibid.). Konkret postuleres altså to ting: et prototypisk computer system af ”tilstrækkelig størrelse” kan udvise ”generel intelligens”, og menneskelig intelligens er resultatet af et fysisk symbol system (ibid.).

Newell og Simon understreger, at hypotesen om fysiske symbol systemer er af empirisk karakter og argumenterer herefter for, at der findes en mængde symbol manipulerende systemer – i form af computere - som udviser intelligent handling inden for en række afgrænsede domæner og fx kan ”forstå naturligt sprog” mv. (Newell & Simon, 1976, pp. 118-119). Det er naturligvis i den forbindelse vigtigt, at betragte Newell og Simons forståelse af hvordan det kan afgøres, om et givent system er intelligent. Som vi så før, anlægges for det første det synspunkt, at intelligens manifesterer sig i det

---

[6] Turing medgiver, at der er ”mysterier” forbundet med bevidstheden, fx i forhold til at lokalisere den (Turing, 1950, p. 447).

[7] Troen på at formel og regelstyret manipulation af abstrakte symboler ikke kun er én mulig, men *den* korrekte beskrivelse af mentale fænomener, kaldes ”computationalisme”, men betragtes i praksis oftest som synonym med funktionalisme (Chalmers, 1992, pp. 1-2).

menneskelige eller elektriske systems opførsel. Det betyder, at "intelligent handling […] er en form for opførsel, vi kan kende på dens effekt, uanset om den er udført af mennesker eller ej" (Newell & Simon, 1976, p. 116). Funktionalismen er altså helt i tråd med Turing opfattelse af, at input/output ækvivalens er ensbetydende med lighed i funktionelt indhold, og dermed den samme indre semantik.

Searle og semantikken

Et af funktionalisternes eksempler på en intelligent computer der udviste ægte forståelse, var et program udviklet af Roger Schanks. Det gives en mængde facts om ting eller aktiviteter og kan på basis af dette svarer på spørgsmål om resultaterne af et hændelsesforløb, som forelægges det i en historie. Computeren kan altså tilsyneladende frembringe ny viden, ved at korrelere forskellige facts. Dette udlægges af funktionalister således, at computeren "bogstaveligt talt kan siges at forstå historien og komme med svar på spørgsmål" (Searle, 1981, p. 284). Denne holdning, at computere kan have forståelse og "andre kognitive tilstande" identisk menneskets, er den ene halvdel af det, Searle (1981, p. 282) kalder "stærk AI", hvilket modstilles "svag AI", der blot hævder at kunne simulere menneskelige bevidsthedsfænomener, uden at computeren dermed selv er bevidst.[8] Den anden del af hvad Searle (2002, p. 669) forstår ved stærk AI er, at "bevidstheden er for hjernen, som programmet er for computeren"[9], hvilket da også er en omformulering af Newell og Simons hypotese om, at ethvert intelligent system, herunder menneskets, kan forklares som computation.

Det kinesiske rum

Funktionalisternes to postulater om menneskelig bevidsthed og stærk AI, samt de konkrete forsøg på "Turing maskine simulation af menneskelige mentale fænomener" som fx Schanks, vil Searle (1981, p. 283) tilbagevise. Til formålet opstiller han følgende tankeeksperiment baseret på Schanks program: Searle er låst inde i et rum og får ind af døren to sæt kinesiske skrifttegn, samt regler på engelsk for hvordan han skal manipulere dem. Han forstår intet kinesisk og kan kun skelne tegnene på deres form. På basis af

[8] Stærk AI er således det, der i indledningen blev defineret som opgavens emne. Terminologien er opfundet af Searle, men har fået bred anvendelse, og kan således findes mange steder i litteraturen om AI.

[9] En.: "the mind is to the brain […]". Begrebet "mind" er svært at oversætte til dansk, og kan betyde "sind", "tanker" eller "bevidsthed" m.v.

reglerne kan han imidlertid korrelere et tredje sæt kinesiske tegn med de to første og returnere kinesiske tegn ud af døren. Searle tillægger naturligvis stadig ikke tegnene betydning, men uden for døren udgør de svar på de spørgsmål, som blev sendt ind af døren sammen med facts og en historie, altså indholdet i hver af de tre sæt kinesiske skrifttegn. Konceptet er altså helt tilsvarende Schanks program og hvis reglerne (som Searle forstår) er formuleret præcist nok, vil de svar han producerer, men intet forstår af, ikke kunne skelnes fra de svar, en indfødt kineser vil give uden brug af reglerne (Searle, 1981, pp. 284-285).

Searle mener det kinesiske rum viser to ting. For det første forstår en computer intet, når den afvikler et program ved at foretage formel manipulation af abstrakte symboler i henhold til programmets regler, selv om input og output af information af identisk med et menneskes. Dette følger for Searle direkte af, at i tankeeksperimentet "[…] er computeren mig: og i tilfælde hvor computeren ikke er mig, har computeren ikke mere, end jeg havde i det tilfælde hvor jeg intet forstår" (Searle, 1981, p. 285). Det må bemærkes, at Searle ved denne parallel sætter lighed mellem "mig", som er menneskelig bevidsthed i en hjerne og krop, og en "computer", som er en samling mikrochips og elektriske kredsløb. Mediet, der eksekverer programmet uden at forstå, er altså af radikalt forskellig karakter, hvilket vil blive behandlet yderligere i forbindelse med kritikken af tankeeksperiment.

Med sammenligningen ønsker Searle at illustrere forholdet mellem formelle, syntaktiske regler og forståelse, ved at "uanset hvilke rent formelle principper du putter ind i computeren vil ikke være tilstrækkelige for forståelse, eftersom et menneske vil kunne følge de samme formelle principper uden at forstå noget" (Searle, 1981, p. 287). Det følger heraf, at eksperimentet for det andet viser, at eftersom programmet altså ikke kan give manden i rummet "tilstrækkelige forudsætninger for forståelse", kan sådan ikke forklares som abstrakt symbol manipulation (Searle, 1981, p. 286). Det er ikke dermed vist, at computationelle processer ingen betydning har for forståelse, men Searle mener ikke, at der er noget belæg for at antage en sammenhæng.

Præmisser

Det kinesiske rum illustrerer på en endog meget intuitiv måde umuligheden af stærk AI, for som Searle siger, virker det "indlysende […], at jeg ikke forstår et ord af de kinesiske historier", når han flytter rundt på abstrakte skrifttegn i henhold til formelle regler (Searle, 1981, p. 285). Det virker formålsløst at betvivle dette. Situationen er fx parallel til den menneskelige operatør i Turings skakspillende papir maskine, som heller ikke forstår noget

som helst om skak, selvom der ligesom i Searles eksperiment er lighed mellem output fra maskinen og en menneskelige modpart, som faktisk forstår skak eller kinesisk.[10] Konklusionerne på de to tankeeksperimenter er imidlertid diametralt modsatte. Turing formulerer sin test for, hvornår et sådant system kan siges at forstå og være intelligent, under den antagelse at input/output ækvivalens i en tilstrækkelige kompleks situation kun kan være resultat af bevidsthed. Testen er således ifølge Searle (1981, p. 304) typisk for hele den funktionalistiske tradition "ved at være uforskammet behavioristisk og operationalistisk."[11] Dette skyldes naturligvis, at Searle netop mener at have vist med det kinesiske rum, at uanset hvor avanceret og raffineret en Turing test en computer består, vil den aldrig være bevidst, intelligent eller forstående, netop fordi den er computationel. Testen bibringer ergo ingen viden om det indre indhold.

I senere udlægninger af tankeeksperimentet identificerer Searle (1997, p. 109) tre mere fundamentale præmisser for denne påstand. Disse må undersøges nærmere, inden kritikken af hans argumentation studeres.

Den menneskelige bevidsthed

Argumentation imod muligheden af stærk AI må indebære antagelser om karakteren af den menneskelige bevidsthed, som det forsøges at eftergøre i computeren. Som beskrevet er den drivende kraft i Searles tankeeksperiment (ganske rimelige) antagelser om menneskets evne til at forstå. Mere generelt er dette udtryk for, at bevidstheden har mentalt indhold af en semantisk karakter, hvilket opfattes således, at "overbevisninger, ønsker og intentioner" […] "er rettet mod eller om objekter eller situationer i verden"[12] og altså er intentionelle tilstande (Searle, 1981, p. 288). Searle tillægger ikke sådanne mentale tilstande nogen metafysisk status, men fastslår gentagne gange at "bevidsthed er et naturligt biologisk fænomen" (Searle, 1997, p. xiv). "Mentale processer er forårsaget af elementer i hjernens opførsel. Samtidig er de realiseret i strukturen som består af disse elementer" (Searle, 2002, p. 669). Bevidstheden skal dermed ifølge Searle (1997, p. 18) forstås som et emergent produkt af hjernens neurale struktur, idet den forklares som en kausal effekt af hjernens elementer uden dermed at være en egenskab ved individuelle

[10] Det skal bemærkes at Turings artikel er relativt ukendt og udgivet posthumt i 1969, cf. http://plato.stanford.edu/entries/turing/

[11] Behaviorisme er den retning inden for psykologi, der studerer opførsel naturvidenskabeligt uden hensyntagen til subjektets indre mentale tilstand. Operationalisme definerer ting, som den proces eller operation der tilvejebringer viden om tingen.

[12] "Situationer" oversat fra en.: "states of affair in the world"

elementer eller summen af disses egenskaber. Netop hjernens "kausale egenskaber, dens evne til at producere intentionelle tilstande" fremhæves som en særegen egenskab (Searle, 1981, p. 295). Searle (1997, p. 59) forstår den som noget konkret neurologisk, idet "neuronerne handler *kausalt* for at forårsage bevidsthed og andre mentale fænomener ved specifikke biologiske mekanismer."

Computerens natur

Searle argumenterer som nævnt mod den funktionalistiske eller computationelle retning inden for kunstig intelligens. Denne var som også nævnt den altdominerende, da han udviklede det kinesiske rum, selv om Frank Rosenblatts 'Perceptron', et tidligt eksempel på neurale netværk, også blev udviklet i 1950'erne, og konceptet således eksisterede (Dreyfus & Dreyfus, 1988, pp. 18-19).[13] Som David Chalmers (1992, p. 39) påpeger, er computere baseret på neurale netværk tilsyneladende ikke så modtagelige for Searles kritik, hvilket vi vil se på senere, men det er dog klart udfra Searles forståelse af forholdet mellem computere og menneskelige bevidsthed, at han ikke erkender dette. Uanset fremskridt i computer teknologi, og fx parallel behandling af data hvorved Searle muligvis hentyder til nogle af neurale netværks egenskaber, fastslås det, at "hvis det virkelig er en computer, må dens operationer defineres syntaktisk" (Searle, 2002, pp. 672-673). Disse operationer er indeholdt i programmer, og "en digital computer [...] er defineret ved dens evne til at implementere programmer. Og disse programmer er rent formelt specificerbare – det vil sige, de har intet semantisk indhold" (Searle, 2002, p. 671). Dette er helt i tråd med opgavens tidligere definition af computation, men det er værd at uddybe denne, og specifikt den relationen mellem syntaks og semantik som udtrykkes i formel symbol manipulation, og som funktionalisme er baseret på.

Formalisering er et matematisk begreb, der groft sagt betegner det at udtrykke semantiske egenskaber gennem relationer mellem syntaktiske egenskaber. Det skal forstås sådan, at fx aksiomerne i Euklids geometri, som oprindeligt kun var baseret på Euklids intuitive forståelse, kan formaliseres til udelukkende at være de syntaktiske relationer mellem elementer i et matematisk udtryk. På den baggrund definerede Turing computere som den klasse

[13] Neurale netværk består af knuder, der reagerer på inputsignaler, og forbindelser mellem disse knuder med forskellig vægtning. De er således en stærk simplificeret model af den menneskelige hjerne. Netværkets elementer skrives i et almindelige programmeringssprog og afvikles på en almindelig computer, mens højerestående funktioner ikke programmeres, men trænes ved at feedback justerer vægte mellem knuder.

af maskiner, der kan beregne en hvilken som helst formaliseret funktion. Det vil sige, at computere ifølge Turing, og som Searle påpeger, kun er syntaktiske, selv om visse semantiske symboler kan udtrykkes rent syntaktisk. Funktionalisme er derfor baseret på antagelsen af, at mentale fænomener er de syntaktiske relationer mellem grundlæggende atomare symboler i bevidstheden, som udgør en art 'tankernes sprog' tilsvarende matematikkens. Disse symboler har, ligesom de matematiske, både syntaktiske og semantiske egenskaber, og kausalitet skyldes symbolernes syntaktiske egenskaber (Horst, 2004). Der kan være uenighed om, på hvilket niveau de atomare symboler findes, og hvilke der blot er sammensatte, men sikkert er det for funktionalister, at semantik kommer fra symboler, der altid repræsenterer noget i verden ved at betegne det. I en konkret computationel repræsentation af bevidstheden er hvert atomart symbol således repræsenteret af en unik bit-streng, og de syntaktiske manipulationer af disse skal resultere i semantik på et højere plan (Chalmers, 1992, p. 31).

Som vi så i forbindelse med det forrige præmis, afviser Searle en sådan computationel forklaring af bevidstheden, idet han mener, at symboler formet af bit-strenge ikke i sig selv kan have et semantisk indhold. De kan kun tilskrives det af mennesker, fx af en programmør. Det følger, at "nullerne og ettallerne er rent abstrakte" og ifølge Searle (1997, p. 59) kun har kausal kraft fordi, "det implementerende medium, hardwaren" kan "producere det næste stadie i programmet." Det er svært at forstå, præcis hvad Searle mener hermed, og hvorfor han fokuserer på de individuelle binære værdier, som jo helt klar ikke er de semantiske enheder i en funktionalistisk model. Ligeledes konstaterer Searle (1981, p. 302) ganske korrekt, at "mentale tilstande og begivenheder er et produkt af hjernens operationer, men programmet er ikke på den måde et produkt af computeren", hvilket forekommer at være en sammenblanding af program (syntaks), og det programmet skal frembringe gennem den syntaktisk manipulation af semantiske enheder, altså mentale tilstande.

Essensen af det andet præmis må imidlertid være, at det vigtigste karakteristika ved funktionalistiske forsøg på kunstig intelligens er, at programmet udfører sine syntaktiske manipulation direkte på de atomare symboler, som tilskrives semantisk indhold (Chalmers, 1992, pp. 33-34). Syntaksen ligger altså på samme niveau som de semantiske (ifølge funktionalismen) enheder, hvilket svarer til forholdet i et naturligt sprog og altså til situationen i det kinesiske rum, hvor de semantiske kinesiske skrifttegn manipuleres i henhold til de syntaktiske regler.

Forholdet mellem syntaks og semantik

Searles pointe med to første præmisser er altså, at menneskets bevidsthed er semantisk, mens computeren er defineret ved at udføre rent syntaktiske operationer. Det tredje præmis, som relaterer de to andre, angår således forholdet mellem syntaks og semantik og består ganske enkelt i, at Searle (2002, p. 674) anser det for en "konceptuel sandhed", at semantik ikke kan opstå af ren syntaks. I tankeeksperimentet betyder dette konkret, at "programmets syntaks er ikke tilstrækkelig for forståelsen af sprogets semantik" (Searle, 1997, pp. 128-130). Dette er naturligvis problematisk for funktionalismen.

Som antydet tidligere henter Searle tilsyneladende distinktionen mellem syntaks og semantik fra den oprindelige "lingvistiske jargon" (Searle, 1981, p. 300). Her er begreberne klart denotativt adskilt, og regler for sammensætningen af ord (syntaks) kan ikke i sig selv give ordene mening (semantik). Tankeeksperimentets umiddelbare drivkraft må derfor siges at bestå deri, at det netop er sproglige tegn mennesket manipulerer, og forholdet mellem syntaks og semantik derfor ses gennem en konventionel sproglig optik. Men det er ikke nødvendigvis givet, at man kan sammenligne et menneskes syntaktiske operationer med en computer, som ikke forstår, at den eksekverer programmets syntaktiske operationer. Programmet kan i stedet forstås som regler, der strukturerer operationen af en bunke transistorer, der derved bliver en maskine som skifter fysisk tilstand kausalt i henhold til program og input (Cole 2004). Programmet er intet uden computeren og vice versa, modsat situationen i det kinesiske rum hvor mennesket allerede har forståelse (fx for andre sprog), ligesom skrifttegnene faktisk repræsenterer noget (på kinesisk). Searle vil ganske givet påpege, at uanset hvor dynamisk computeren kan opføre sig, er dette altid baseret på syntaktisk manipulation af symboler. En sådan kan udføre kraftfulde simulationer af egenskaber ved menneskelig bevidsthed, men "ingen simulation kan i sig selv konstituere en duplikation" (Searle, 2002, p. 673).

Med hensyn til semantik påpeger Chalmers (1992, pp. 29-30) at må man skelne mellem muligheder for semantisk indhold i bevidstheden, hvilket naturligvis er afgørende for kunstig intelligens' potentiale. Den ene type er afhængig af forbindelser eksternt bevidstheden, for at have mening, altså om fx ting eller begivenheder. Den anden type semantisk indhold er kun afhængig af hjernens interne mekanismer, og er således mulig uden nogen forbindelse til omverdenen, om end fraværet af forbindelser vil påvirke kvaliteten af det mentale indhold. Det er denne sidste type semantik Searle hævder principielt identificerer den menneskelige bevidsthed, naturligvis uden

dermed at fornægte at det meste semantiske indhold er forbundet til den ydre verden.[14] Man kan derfor ikke afvise argumentet ved at hævde, at computeren kunne opnå forståelse hvis den blev udstyret med passende sanseapparatur, hvilket er essensen af det modargument Searle (1981, p. 293) kalder "robot svaret".[15]

Searles konklusioner

De tre præmisser er i det foregående undersøgt gennem beskrivelser og definitioner hentet bredt fra Searles skriverier. Tilsyneladende anser Searle i praksis postulaterne for at være egentlige aksiomer, altså sandheder af en selvindlysende eller tautologisk karakter der således ikke behøver videre forklaring. De to første er da også relativt ukontroversielle. Det kan hævdes, at Searles beskrivelsen af menneskets bevidsthed er overfladisk, men formålet er da også her blot at påpege dens semantiske og kvalitative karakter. Dette behøver Searle ikke argumentere for overfor målgruppen, som er tilhængere af stærk AI og jo netop ønsker at reproducere disse karakteristika. Der kan ligeledes heller ikke være megen tvivl om, at computere ifølge Turings definition, som er den, der ligger til grund for funktionalistisk tilgang til stærk AI, er styret af formaliserede og rent syntaktiske programmer. På trods af nogle mulige, mindre misforståelser er Searles pointe her helt klart, at funktionalismen forsøger at skabe semantiske bevidsthedstilstande, som dem Searle postulerer i det første aksiom, gennem programmers manipulation af atomare semantiske enheder. Disse enheder er naturligvis på et meget lavere semantisk niveau end det tilstræbte mentale produkt, men Searle (1981, p. 300) mener i øvrigt at "symbolerne ikke symboliserer noget."

Altså er det centrale præmis, og det som det kinesiske rum egentlig skal demonstrere, at semantik ikke kan opstå af syntaks, hvilket gør Searle (2002, p. 674) i stand til at konkludere, at "ingen computer er i sig selv tilstrækkelig til at give et system en bevidsthed."[16] Erkender man, at menneskets bevidsthed faktisk eksisterer, følger af konklusionen også at denne bevidsthed ikke kan forklares som formel manipulation af abstrakte symboler gennem

---

[14] Chalmers henviser til: Searle, J. R. (1980b). Intrinsic intentionality. *Behavioral and Brain Sciences*, *3*, 450–457. Her skriver Searle bl.a. at "If I were a brain in a vat I could have exactly the same mental states I have now […]."

[15] Searle benytter ikke skellet mellem eksternt og internt afhængig bevidsthed i argumentet mod AI, hvor det egentlige argument ligger i forholdet mellem syntaks og semantik.

[16] "Bevidsthed" oversat fra en.: "mind", se fodnote nr. 9 mht. dette ord.

afviklingen af et program. Searle mener dermed definitivt at have afvist det funktionalistiske projekt og stærk AI generelt.

Slutningen fra disse tre præmisser, og den måde hvorpå det kinesiske rum demonstrerer sammenhængen mellem syntaks og semantik, kan imidlertid problematiseres, hvilket vil blive forsøgt i den følgende diskussion.

Diskussion

Searle fremkom med det kinesiske rum i 1981 og har siden destilleret meningen i det på forskellige måder, fx ved at eksplicitere de tre aksiomer som udtrykker det grundlæggende og uændrede indhold i tankeeksperimentet. Argumentet er i den oprindelige udgave suppleret med en lang række modargumenter samlet af Searle og alene frem til midten af 1990'erne blev endvidere vel over 100 artikler publiceret, som diskuterede, problematiserede eller forsøgte at tilbagevise Searles argumentation (Cole, 2004). I det følgende vil nogle af de generelle problematikker omkring argumentet blive berørt sammen med det mest udbredte modargument, det Searle kalder 'system svaret'.

Generelle problematikker

Argumentet er af mange blevet beskyldt for høj grad at spille på intuitioner, om hvilke substanser der kan producere bevidsthed eller intentionalitet, ligesom betydningen af 'forståelse' ikke ekspliciteres af Searle (ibid.).

Med hensyn til det første, så kan Turings universelle computer, som beskrevet tidligere, implementeres i en række forskellige maskiner af fx elektrisk eller mekanisk karakter eller noget helt andet. Dette benytter Searle (1981, p. 301) til at fremkomme med en række eksempler på "alle slags skøre realiseringer", som består af "den forkerte slags ting til at have intentionalitet"[17], fx "sten, toiletpapir, vind og vandrør". Fælles for disse er at computere af tilstrækkelig kompleksitet til at afvikle et avanceret AI program, i sådanne implementeringer ville være gigantiske, Storm P. lignende, mekaniske konstruktioner, der i øvrigt med al sandsynlighed ville være umulige at bygge. Praktiske forhold er imidlertid uden betydning for Searle, som blot vil vise, at når sådanne substanser i princippet kan realisere et AI program på samme måde som en 'von Neumann' computer, så kan man heller ikke tilskrive dennes mikrochips mulighed for at skabe intentionalitet. Problemet er, at dette vises ved at spille på læserens intuitioner om, at fx "en bunke

[17] En.: "the wrong kind of stuff […]"

små sten" aldrig vil kunne realisere intelligent handling (ibid.). Men hvorfor er det naturligt for Searle, at 1,5 kilogram grå masse i menneskets kranie kan producere forståelse, intentionalitet, bevidsthed? Det skal her siges, at Searle fuldt ud anerkender problemet i at forstå, "hvordan det er muligt, for fysiske, objektive, kvantitativt målbare neuroner som tænder, at forårsage kvalitative, private, subjektive oplevelser."[18] Dette spørgsmål er faktisk det primære problem i forholdet mellem krop og bevidsthed (Searle, 1997, p. 28). Ligeledes erkender Searle (2002, p. 675), at der kan findes andre biologiske substanser, fx hos marsmænd, som hvis de producerer bevidsthed må have "kausale kræfter lig den menneskelige hjernes." Dette følger naturligvis direkte af Searles argumentation, men det er uklart, hvorfor microchips ikke vil kunne have sådanne kausale kræfter, og hvad disse kræfter egentligt består i.

I det samlede argument er forholdet mellem syntaks og semantik det essentielle, men Searle underbygger alligevel sin pointe med disse intuitioner om, at kun særligt udvalgte substanser kan producere bevidsthed. På dette punkt leder det kinesiske rum således tankerne hen på Leibniz, der blev omtalt indledende. I sit tankeeksperiment undersøger Leibniz maskinens indre og finder "kun dele, som virker på hinanden, og aldrig noget hvorved at forklare en perception. Derfor er det i en simpel substans og ikke i en sammensætning eller i en maskine, at perception må søges efter" (Dennett, 2005, p. 3). Hos Leibniz er den simple substans af metafysisk karakter, mens den hos Searle er specifikt biologisk, men i begge tilfælde er det intuition der afgør, at netop denne substans har "de rigtige sager til intentionalitet"[19] (Searle, 1981, p. 301).

Sådanne antagelser må dog være underordnet det primære forhold i det kinesiske rum, nemlig placeringen af den menneskelige operatør som den der eksekverer programmet. Det er for Searle "indlysende" at mennesket intet kinesisk forstår ved at følge de formelle regler, hvilket da også er en rimelig konklusion. Formålet er netop at vise, hvordan det formelle program ikke kan tilføre forståelse til et menneske, som i forvejen kan forstå andre ting (Searle, 1981, p. 301). Men denne sammenblanding af operatøren og det som rent faktisk består Turing testen, nemlig det kinesisk rum som helhed, skaber ifølge Douglas Hofstadter (Dennett & Hofstadter, 1981, pp. 373-375) "et helt utroligt urealistisk koncept om relationen mellem intelligens og symbol manipulation." Dette sker ved, at Searle placerer forståelsen hos mennesket, som allerede forstår og som det derfor er logisk at spørge: "forstår du mere

[18] En.: "[…] quantitatively describable neurons firing […]"

[19] En.: "the right kind of stuff for intentionality"

kinesisk nu?" Men dette er forfejlet ifølge Hofstadter (ibid.), idet næsten al forståelsen må ligge i papirinstruktionerne, der af Searle ikke tillægges nogen videre betydning eller omfang, men som i praksis ville bestå i millioner eller billioner af sider. Det står klart, at *hvis* det kinesiske rum faktisk forstod, hvilket ifølge Searle er umuligt, så ville denne forståelse ikke være en projektion af den menneskelige operatørs forståelse, men det er ikke desto mindre sådan det fremstår i det kinesiske rum (Dennett, 1991, p. 438).

System svaret

Blandt forskere inden for AI er det mest almindelige argument, mod de konklusioner Searle drager fra det kinesiske rum således, at det er hele systemet, udgjort af både regler, menneske og hjælpemidler (papir og blyant), der forstår (Dennett, 1991, p. 439). Det svarer til, at mennesket i det kinesiske rum er computerens CPU, og den vil de færreste nok tilskrive forståelse i et potentielt bevidst AI system. Man ville i stedet sige, at det var det samlede system der forstod, ligesom Searle heller ikke hævder, at det er selve hjernen eller den individuelle neuron der forstår, men at mentale tilstande såsom forståelse er et emergent fænomen af hjernens samlede neuronstruktur. På baggrund af Searles biologiske forståelse af bevidstheden må man antage, at han også ville medgive, at denne struktur er styret af en eller anden form for regler, naturligvis af en anden beskaffenhed end computer programmer.

Det kan siges, at system svaret ikke direkte adresserer dette centrale problem, altså forholdet mellem syntaks og semantik, men primært viser, at Searle forsøger at illustrere denne sammenhæng på en måde, der er så langt fra den måde et faktisk AI system fungerer på, at tankeeksperimentet måske slet intet viser. Searle benytter måske, at forholdet mellem syntaks og semantik i et programmeringssprog konceptuelt er det samme som i naturlige sprog, og overfører dette til det samlede computersystems operation. Som beskrevet tidligere er det ikke klart om en sådan overførsel er meningsfuld. Ligeledes er den lingvistiske syntaks/semantik distinktion muligvis indskrænkende eller misvisende mht. begrebet 'forståelse' som er centralt i Searles argument.

Chalmers (1992, p. 40) forsøger i stedet at tolke syntaks bredere som 'regelfølge', men det betyder, at Searles idéer om bevidsthedens emergens fra hjernens komplekse struktur bliver sårbare for de samme typer argumenter, som han selv benytter mod AI. Menneskets hjernen er i sidste ende styret af fysikkens regler, men af denne regelstyrede atomare substans opstår alligevel noget vi kalder bevidsthed og tilskriver semantik. Chalmers'

pointe er dog, at det måske netop er afgørende hvilket niveau den syntaktiske manipulation foregår på, relativt til det semantiske, og at regelfølgen i hjernen antageligt er på et meget lavere niveau. Han mener derfor, at hvad angår AI baseret på neurale netværk, hvor semantik er indeholdt i strukturer af mindre enheder som manipuleres syntaktisk, er Searles argumentation ikke så sikker (Chalmers, 1992, pp. 38-39). Dette løser naturligvis ikke spørgsmålet om sammenhængen mellem semantik og syntaks, og Chalmers konkluderer at "problemet omkring hvordan […] et semantisk, mentalt niveau kan opstå fra et mekanisk substrat er et af de mest langhårede i bevidsthed/krop spørgsmålet. Det ville ikke være overraskende om fremkomsten af semantik fra syntaks i computer systemer skulle være lige så svær at forstå" (Chalmers, 1992, p. 47).

Konklusion

Det kan således konkluderes, at John Searle fremhæver nogle centrale problemstillinger omkring frembringelsen af semantik fra syntaks og forholdet herimellem. Særligt påpeges mulige problemer i funktionalistiske modeller, men det kan ikke afgøres, hvorvidt Searle dermed har bevist noget konkret om mulighederne for kunstig intelligens. Ligeledes rejser tankeeksperimentet spørgsmål omkring, hvad forståelse indebærer og hvordan dette relaterer sig til syntaks. I brugen af disse og andre begreber appellerer 'det kinesiske rum' på visse punkter til intuitive slutninger og rejser således også implicit spørgsmål om tankeeksperimenters værdi.

Litteraturliste

Carlin, L. & Kulstad, M. (2004). Leibniz's Philosophy of Mind. I E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2004 Edition).* Hentet 13. januar, 2007, fra http://plato.stanford.edu/archives/fall2002/entries/leibniz-mind/

Chalmers, D. J. (1992). Subsymbolic Computation and the Chinese Room. I J. Dinsmore (Ed.), *The Symbolic and Connectionist Paradigms: Closing the Gap* (pp. 25-48). Hillsdale, N.J.: Lawrence Erlbaum.

Cole, D. (2004). The Chinese Room Argument. I E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2004 Edition).* Hentet 7. januar, 2007, fra http://plato.stanford.edu/archives/fall2004/entries/chinese-room/

Copeland, B. (2006). The Modern History of Computing. I E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Summer 2006 Edition).* Hentet 13. januar, 2007, fra http://plato.stanford.edu/archives/sum2006/entries/computing-history/

Dennet, D. C. & Hofstadter, D. R. (1981). *The Mind's I – Fantasies and Reflections on Self and Soul*. New York: Basic Books.

Dennett, D. C. (1991). *Consciousness explained.* Penguin Books.

Dennett, D. C. (2005). *Sweet Dreams.* Cambridge, M.A.: MIT Press.

Dreyfus, H. L. & Dreyfus S. E. (1988). Making a Mind Versus Modelling the Brain: Artificial Intelligence Back at a Branchpoint. I S. R. Graubard (Ed.), *The Artificial Intelligence Debate: False Starts, Real Foundations* (pp. 15-43). Cambridge, M.A.: MIT Press.

Horst, S. (2004). The Computational Theory of Mind. I E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2004 Edition).* Hentet 8. januar, 2007, fra http://plato.stanford.edu/entries/computational-mind/

Newell, A. & Simon, H. (1976). Computer science as empirical inquiry: symbols and search. *Communications of the ACM, 19(3),* 113-126.

Searle, J. R. (1981). Minds, Brains, and Programs. I J. Haugeland (Ed.), *Mind Design* (pp. 282-306). Cambridge, M.A.: MIT Press.

Searle, J. R. (1997). *The Mystery of Consciousness.* London: Granta Publications.

Searle, J. R. (2002). Can Computers Think? I D. J. Chalmers (Ed.), *Philosophy of mind – classical and comtemporary readings* (pp. 669-675). New York: Oxford University Press.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59(236),* 433-460.

Wackerhausen, S. (1989). Mennesket i computerens billede. *Philosophia, 18(1-2),* 111-146.