A A R H U S U N I V E R S I T E T

Eksamensopgave
Bacheloruddannelse i Informationsvidenskab (2006-studieordning)

Eksaminator: Finn Olesen **Afleveringsdato:** 22.01.2008

Censor (udfyldes af sekretariatet)**:**

Intro til programmering, intern best./ikke best.	☐	(sæt kryds)
Programmering og systemudvikling, intern 13-skala	☐	(sæt kryds)
Kommunikation-1, intern best./ikke best.	☐	(sæt kryds)
Kommunikation-2, ekstern 13-skala	☐	(sæt kryds)
Digital æstetik, intern 13-skala	☐	(sæt kryds)
Teknologihistorie , ekstern 13-skala	☐	(sæt kryds)
Organisationsanalyse, ekstern 13-skala	☐	(sæt kryds)
Studium generale, intern 13-skala	☒	(sæt kryds)
Informationsvidenskabelig metodet, intern best./ikke best.	☐	(sæt kryds)
Valgfrit projekt, intern 13-skala	☐	(sæt kryds)
Bachelorprojekt, ekstern 13-skala	☐	(sæt kryds)

Opgavens anslag: 72.246

Afleveret af:

Årskort: 20063185 **Navn:** Villars Gimm

Årskort: 20030605 **Navn:** Nicki T. Hansen

Årskort: **Navn:**

Årskort: **Navn:**

Må eksaminators eksemplar af eksamensopgaven gøres til genstand for udlån? JA ☒ NEJ ☐

Studienævnet for Informations- og Medievidenskab, Helsingforsgade 14, 8200 Århus N

Stærk Kunstig Intelligens

Studium Generale

Af
Villars Gimm (20063185)
Nicki T. Hansen (20030605)
22-01-2008

Indholdsfortegnelse

Antal tegn: 72.246

Ansvarsfordeling:

Villars er ansvarlig for følgende afsnit: *Stærk Kunstig Intelligens* (s. 4-7), Turings Ni Indsigelser (s. 7-12) og implementeringsmæssige begrænsninger ved en Stærk Kunstig Intelligens (s. 18-23).

Nicki er ansvarlig for følgende afsnit: *Bevidsthed* (s. 12-18) og Searles Chinese Argument (s. 23-27).

I fald af at *alt* skal ansvarsfordeles, hører indledning og konklusion også til Nickis ansvar; ellers er de og resten fælles ansvar.

Problemformulering

Kan man principielt fremstille en Stærk Kunstig Intelligens ud fra et bevidsthedsmæssigt perspektiv?

Indledning og metode

Denne opgaves formål er, at afdække, hvorvidt en *Stærk Kunstig Intelligens* (AI[1])kan lade sig gøre fra et bevidsthedsmæssigt perspektiv. Vi vil derfor forklare begrebet Stærk Kunstig Intelligens, ud fra flere synspunkter, og finde frem til en fornuftig afklaring af termen. Vi vil derefter, med udgangspunkt i Alan Turings ni indsigelser, undersøge flere mulige problemer med en Stærk AI. Vi vil endvidere behandle bevidstheden, primært fra Daniel C. Dennetts teorier, da vi mener, at bevidstheden er en del af et intelligent væsen. Hvad bevidstheden egentlig *er*, er dog vigtigt at forklare, før man kan forfølge dette synspunkt. Vi anskuer efterfølgende de implementerings-mæssige problemer ved en stærk AI med udgangspunkt i Hubert L. Dreyfus, og afslutter med en gennemgang af John R. Searles Chinese Room argument, som proklamerer at afvise en Stærk Kunstig Intelligens fuldstændigt. Vi har en formodning om at dette argument ikke holder vand, men det må komme an på en prøve. I sidste ende vil vi ud fra ovennævnte undersøgelser og diskussioner, konkludere hvorvidt Stærk Kunstig Intelligens er principielt mulig fra et bevidsthedsmæssigt perspektiv.

Stærk Kunstig Intelligens

Udtrykket *Kunstig Intelligens* (AI[2]) blev oprindeligt opfundet af John McCarthy i 1955 og havde fra hans side betydningen ”the science and engineering of making intelligent machines”(McCarthy 2007, ”basic questions”). Udtrykket er blevet diffuseret og tillagt ny mening de seneste års udvikling indenfor AI- og hjerneforskning, hvorfor vi i dette afsnit vil gøre det helt klart, hvad vi i denne opgave mener, når vi skriver ”Kunstig Intelligens”.

Turingtesten

Alan Turing fremstiller i 1950 en afart af en gammel leg kaldet *Imitation Game* (imitationsspil), som senere er blevet døbt *Turingtesten*. Turingtesten bliver fremstillet for at give et bedre alternativ til spørgsmålet ”Kan maskiner tænke?”, da han finder, at spørgsmålet ligger op til en kvantitativt svar, da der er mange forskellige meninger for hvad ”tænke” er. Vi vil ikke gennemføre en grundig gennemgang af Turingtesten her, men kun fremvise et kort oprids af den.

Imitation Game går ud på at isolere en mand (A) og en kvinde (B) fra en tredje person (C), det handler så for C igennem spørgsmål at finde ud af, hvem der er kvinden og hvem er manden. Spørgsmålene bliver typisk medieret igennem en fjerde person eller i nyere tid igennem et

[1] Forkortelsen *AI* er baseret på den engelske oversættelse, *Artificial Intelligence*.
[2] Fra den engelske oversættelse *Artificial Intelligence*.

tekstuelt baseret medie. Grunden til dette er at det kun er meningen, der skal afsløre noget; ikke stemmen eller andre indlysende ting.

Turing foreslår et imitationsspil, hvor person A eller B udskiftes med en maskine, og hvis maskinen kan snyde C, kan man kalde den *intelligent*. Det er dog vigtigt at notere sig, at Turing ikke siger ét spil er nok; på samme måde som det oprindelige imitationsspil, skal man tage flere spil og se på den statistiske sammenhæng – tages der ligeså meget fejl her mellem maskine og menneske, som der før blev taget mellem mand og kvinde? (Turing 1950, ”1 The Imitation Game”)

Styrker og svagheder ved Turingtesten

Styrkerne i Turingtesten er, at det er muligt at snakke om *alt*, som Turing selv skriver:

> ”The question and answer method seems to be suitable for introducing almost any one of the fields of human endeavour that we wish to include.” (Turing 1950, ”2 Critique of the New Problem”)

Der er herved et implicit krav, at maskinen ikke blot skal forstå ordene, men også emnet; det vil sige, maskinen skal ikke blot forstå syntaksen, men også semantikken. Dette er et punkt, som John Searle også berører, og som vi omtaler nedenstående. Endvidere skal en maskine, for at kunne bestå en veldesignet Turingtest, kunne bruge naturligt sprog, ræsonnere, og besidde og kunne tilegne sig viden. Man kunne endvidere udvide testen til at inkludere visuelt input og andre sanse input i form af objekter, der kunne sendes ind til en give person, som så skulle omtales enten i form af identificering eller egenskabsmæssige spørgsmål: ”Er objektet jeg sendte ind blød eller hård?”, ”Hvordan ser objektet ud?”. Dette betyder, at en velformuleret Turingtest kan omhandle alle – eller næsten alle – egenskaber som mennesket besidder.

Der er dog også svagheder ved Turingtesten. Den er gennemgående *antropomorfisk*, idet den direkte sammenligner et menneske og en maskine. Maskinen skal således ikke blot være intelligent, men være *menneskeligt intelligent*. Menneskelig opførsel og intelligent opførsel er ikke nødvendigvis identisk, og der er flere scenarier man kan forestille sig, hvor en ellers intelligent maskine vil kunne fejle. Et menneske har mange begrænsninger, som f.eks. beregningshastighed, hvorfor en ellers intelligent maskine vil fejle Turingtesten, hvis den kan udregne et givent regnestykke *for hurtigt*. Endvidere kan mennesker vildlede og lyve, hvilket er svært at inkorporere i en maskine. Turing overvejer selv antropomorfismen, og siger:

> ”The game may perhaps be criticised on the ground that the odds are weighted too heavily against the machine. If the man were to try and pretend to be the machine he would clearly make a very poor showing. He would be given away at once by slowness and inaccuracy in arithmetic. May not machines carry out some-thing which ought to be described as thinking but which is very different from what a man does? This objection is a very strong one, but at least we can say that if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection.” (Turing 1950, ”2 Critique of the New Problem”)

Turingtesten er endvidere eksplicit behavioristisk, idet at den kun ser på deltagernes opførsel. Dette betyder, at en maskine, der kan gennemføre Turingtesten ikke nødvendigvis har en sind og

er intelligent, men blot er en simulation af menneskelig kommunikativ opførsel, som kan udføres igennem et regelstyret system. En af de store proponenter af kritikken af den behavioristiske del af Turingtesten er John Searle, som med sit *Chinese Room* argument forsøger at afvise Turingtesten, som visende ægte intelligens. Vi behandler Searles argument senere i nærværende tekst. Turing fremfører endvidere selv et modargument[3] til denne tvivl af bevidstheden i maskinen, som vi behandler, i afsnittet om bevidsthed.

En sidste umiddelbar svaghed – og styrke – ved Turingtesten er, at et menneske kan fejle det enkelte imitationsspil, da den baseres på en subjektiv bedømmelse af opførslen. Dette udbedres dog idet, at man skal se på en kvantitativ optælling af flere imitationsspil, og se om maskinen scorer markant anderledes end de menneskelige aktører.

Delingen mellem Stærk og Svag AI

John Searle (1980) er den første til at lave opdelingen mellem *Svag* og *Stærk AI* i forbindelse med hans *Chinese Room* argument, som vi behandler senere i nærværende tekst. Da Searle fremligger sit argument, er det i forbindelse med den løbende udvikling af AI i forbindelse med simulation af menneskets kognitive kapaciteter. I denne sammenhæng vil han opdele mellem en *Svag AI hypotese*, hvor computeren blot er et værktøj til at formulere og afprøve hypoteser mere udtømmende og præcist, og en *Stærk AI hypotese*, hvor man ser computeren som ikke blot et værktøj, men, hvis passende programmeret, et virkeligt *sind*[4] som har kognitive tilstande og *forstår*, hvad den "simulerer". Det vil sige man kan opsummere Searles opdeling således:

- Svag AI: Et fysisk symbolsystem, der kan *opføre sig* intelligent.
- Stærk AI: Et fysisk symbolsystem, der har et sind og mentale tilstande; og *er* intelligent.

I Searles terminologi ville både en Stærk og Svag AI kunne gennemføre en Turingtest, men kun den Stærke AI ville egentlig være intelligent. Dette er ikke den gængse opfattelse af Svag AI, som i dag også kaldes *anvendt AI* eller *snæver AI*, og den benyttes til meget specifikke formål, såsom at optimere data, kontrollere specifikke organisatoriske problemer eller anti-virus heuristik. Den Svage AI eksisterer således i dag, og har en "problemknuser"-rolle på specifikke problemer, som den typisk løser hurtigere og mere effektivt end et menneske kan.[5]

Den gængse opfattelse af Stærk AI

Den gængse opfattelse af Stærk AI kommer meget an på, hvilken teoretiker man spørger om emnet. Futuristen Raymond Kurzweil (2005, s. 260) opsummerer den gængse opfattelse, således; *en Kunstig Intelligens, som matcher eller er større end menneskelig intelligens*. Denne opfattelse er intuitivt nem at forstå, men medfører at man er nødt til at forklare, hvad menneskelig intelligens er, og hvilke krav der eksisterer før man kan regne noget for intelligent.

[3] Turings "polite convention", en art af "other minds" argumentet. Se nærmere i afsnittet om bevidsthed.
[4] Bemærk at vi bruger oversættelsen "sind" for det engelske "mind", som man måske kan mene indeholder mere end det danske ord.
[5] Hvis man betvivler dette, så kan 1 minuts afbenyttelse af www.google.com formentlig fjerne denne tvivl.

”Hvad indebærer menneskelig intelligens?” er dog ikke et let spørgsmål at besvare. Searle fremstiller det synspunkt, at et intelligent væsen har et sind, hvor Daniel C. Dennett fremstiller synspunktet, at et intelligent væsen blot skal have alle de underliggende materialistiske elementer for et sind, så har man det eneste vi selv har som grundlag for et sind. Et ”sind” er der dog ikke en koncis og klar ordforklaring for, da det typisk fungerer som et paraply-ord, der omfatter mange, mange under-egenskaber. Disse egenskaber inkluderer bevidsthed, intellekt, fornuft, sansning, tanker, læring, forståelse og flere andre. Sindet er også den kontinuere tankestrøm, *bevidsthedsstrømmen,* om man vil, af alle de egenskaber, der eksisterer i hjernen.

Flere egenskaber, udover sindet, kan endvidere opsættes for, at noget er intelligent, såsom at kunne kommunikere i naturligt sprog (se Turingtesten), selvbevidsthed – eksistensen af et *jeg* og derved en identitet – og en situering i verden, således at der kan ageres i verden og manipuleres med objekter (hvilket nærmer sig robotstudier). Andre vil argumentere for intentionalitet og den frie vilje, og derfra original tanke og ideer, er essentiel i et intelligent væsen.

Disse mange forskellige krav kan alle diskuteres, og bliver det hyppigt, hvorfor vi ikke vil behandle dem alle i nærværende tekst grundet sidemæssig begrænsning, men i stedet udvælge nogle som vi mener, er centrale og interessante for en diskussion om muligheden for Stærk Kunstig Intelligens. Visse teoretikere mener at *alle* de ovenstående krav, og i visse tilfælde flere til, skal overholdes, før man kan regne noget for et Stærkt intelligent væsen. Andre mener, at ikke alle er nødvendige, da et intelligent væsen ikke behøver at have alle muligheder som et menneske har. Vi kan anføre, at kravene ofte bliver sat så højt for en AI, at visse mennesker også vil falde igennem, hvorfor en række etiske spørgsmål rejser sig vedr. mentalt handicappede og lignende. Denne diskussions-sump vil vi dog ikke vade ud i, i denne tekst.

Turings Ni Indsigelser

Turing (1950) fremstiller ni mulige indsigelser mod Turingtesten i teksten omhandlende samme. Vi behandler i dette afsnit indsigelserne og undersøger de mulige problemer, der er med Turingtesten.

Det teologiske argument

> “Thinking is a function of man's immortal soul. God has given an immortal soul to every man and woman, but not to any other animal or to machines. Hence no animal or machine can think.” (Turing 1950, ”6 Contrary Views on the Main Question”)

Turing afviser det teologiske argument, på baggrund af at den tidligere er blevet benyttet til at modbevise nye teorier om verden. Ved Galileos tid brugte teologerne et bibelvers som forklarede at Solen stod højt på himlen en hel dag og at jorden ikke skulle bevæge sig, hvilket senerehen viste sig at være humbug. Turing svarer dog den teologiske indsigelse ved at svare i samme sprog: hvis Gud er almægtig, er der ingen grund til at tro at kun mennesket er blevet – eller vil blive –

velsignet med en sjæl (og derved bevidsthed). (Turing 1950, ”6 Contrary Views on the Main Question”)

”Hovedet i sandet” argumentet

> “The consequences of machines thinking would be too dreadful. Let us hope and believe that they cannot do so." (Turing 1950, ”6 Contrary Views on the Main Question”)

Denne frygt for en tænkende maskine værdiger Turing ikke et svar, da han mener at emnet er substansløst, og at frygten for tænkende maskiner, og derved ikke-unikheden ved mennesket, er en tåbelig.

Det matematiske argument

Turing laver dette argument, baseret på matematiske uløseligheder, såsom Halting problemet, der aldrig kan løses af en Turingmaskine. Turing fremstiller formildende omstændigheder i form af rammer for spørgsmålene og deres output i dette imitationsspil. Han antager at man søger binære (ja/nej) spørgsmål, og derfor ikke spørger til subjektive ting, som f.eks. holdninger til Beethovens 5. symfoni. Endvidere medgiver han, at der er visse spørgsmål som ikke kan besvares korrekt af en maskine eller besvares nogensinde, ligegyldigt hvor meget tid der tildeles.

Selvom der er på bevist begrænsninger ved beregnelighed, er der ikke fundet beviser for at disse begrænsninger ikke eksisterer den menneskelige hjerne, og selvom computeren laver fejl i dens besvarelse, er mennesker ej heller ufejlbarlige, hvorfor den stadig ikke nødvendigvis vil fejle en Turingtest. Turing advarer endvidere mod at vi ikke skal føle os for overlegne, da vi ofte tager fejl i vores antagelser og sanser. Opsummerende giver Turing dette svar:

> “In short, then, there might be men cleverer than any given machine, but then again there might be other machines cleverer again, and so on.”(Turing 1950, “6 Contrary Views on the Main Question”)

Argumentet fra bevidstheden

> "Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain-that is, not only write it but know that it had written it. No mechanism could feel (and not merely artificially signal, an easy contrivance) pleasure at its successes, grief when its valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, be angry or depressed when it cannot get what it wants." (Turing 1950, ”6 Contrary Views on the Main Question”)

Turing citerer Professor Jefferson for ovenstående, hvilket ikke blot peger på bevidsthed, men også om subjektive følelser. Jefferson ønsker her, at maskinen skal have en semantisk forståelse for, hvad den fremstiller; ikke blot tilfældigt lave præcis det som den gør. Turing afviser dette argument med det såkaldte *other minds* svar, som falder tilbage på, at vi ikke kan vide om andre personer end os selv har en bevidsthed, da bevidstheden ikke er direkte målelig, og han tilføjer:

> ”Instead of arguing continually over this point it is usual to have the polite convention that everyone thinks.” (Turing 1950, ”6 Contrary Views on the Main Question”)

Han har således en *høflig konvention*, som antager, at hvis en maskine (eller et væsen) opfører sig så intelligent som et menneske, så er det ligeså intelligent som et menneske. Denne høflige konvention finder de fleste filosoffer dog ikke tilfredsstillende, hvorfor vi har dedikeret et helt afsnit til bevidstheden i denne tekst.

Argumenter fra forskellige mangler

> “I grant you that you can make machines do all the things you have mentioned but you will never be able to make one to do X" (Turing 1950, ”6 Contrary Views on the Main Question”)

Turing indsætter følgende som eksempler på X:

> “Be kind, resourceful, beautiful, friendly, have initiative, have a sense of humour, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make someone fall in love with it, learn from experience, use words properly, be the subject of its own thought, have as much diversity of behaviour as a man, do something really new.” (Turing 1950, ”6 Contrary Views on the Main Question”)

Disse mangler er af forskellige karakter, nogle mere frivole end andre, som for eksempel smagen af jordbær. Umiddelbart virker denne mangel ikke så nødvendig for, at en maskine virker intelligent, men denne kan være støttende til de mere seriøse mangler, såsom venskabelighed. Hvor det måske er muligt at fremstille en maskine, der kan simulere glæden ved at spise jordbær, ville dette blot være idiotisk, ifølge Turing. (Turing 1950, ”6 Contrary Views on the Main Question”)

Tanken om at en maskine ikke kan lave fejl er af en anden størrelse. Turing antager, at denne er ment med henblik på et imitationsspil. Gennem et sådant spil, vil der være en forhørsleder og et subjekt under forhør. En forhørsleder vil gennemskue maskinen, da den svarer med for høj præcision og hastighed i forhold til et menneske, som i matematiske problemer typisk vil svare i cirkatal og langsommere.

Dreyfus understøtter dette argument, med regelbaserede eksempler, såsom at skulle skifte gear i en bil ved 20 km/t, hvor en maskine vil skifte præcis på 20 km/t, hvor et menneske vil slække denne regel og skifte *omkring* 20 km/t. (Dreyfus og Dreyfus 1986, s. 63-64)

Turing har en anden holdning, da han mener at det er muligt at programmeres sig ud af disse problemer, hvorfor det er muligt at få en maskine til at give et svar upræcist. Dette kan gøres aritmetisk, så maskinen beregner sig til et sandsynligt svar, eller en laver en sandsynlig afskæring af svaret. For eksempel kunne den ved et spørgsmål om π-værdi svare 3,14 i stedet for 3,14159… og så videre i talrækken. Dette vil *vildlede* forhørslederen, således at han tvivler på at det er en maskine der kommunikeres med. Turing mener ikke, at denne påstand behøver dybere forklaring, da denne er baseret på en mangel af forståelse for computere og deres muligheder. Det er korrekt at computeren ikke kan tage fejl i en funktion, men der kan indlægges videnskabelig induktion i en

funktion og ved induktionen kan computeren *simulere* regnefejl. Maskinen kan således fordumme sig selv, så den minder mere om et menneske. (Turing 1950, ”6 Contrary Views on the Main Question”)

De der siger at computere ikke kan være subjektive ved deres egne resultater, er uenige med programmører. En programmør vil, ifølge Turing, mene, at en computer, der laver følgende udregning x^2-40x = 0 være subjektiv omkring sit svar, da den har beregnet sig frem til det rigtige svar ved at følge computerens egne beregningsmetoder. Derfor er den subjektiv omkring sit svar. Det vil endog være muligt for computeren at skabe sine egne funktioner, og i et tilfælde, hvor en computer har skabt sin egen funktion, vil den være subjektiv omkring sit svar. (Turing 1950, ”6 Contrary Views on the Main Question”)

Den sidste af den lange række af mangler en kunstig intelligens kan have, er at en maskine ikke kan have differentieret adfærd, hvilket vil sige, at den ikke opfører sig overraskende for mennesker. Den er deterministisk. Turing mener, det er et spørgsmål om at have kapacitet nok, hvorved en kunstig intelligens kan tillære sig differentieret adfærd ved at have tilstrækkeligt forskellige muligheder, således at der stadig overrasker med ”nye” handlinger. (Turing 1950, ”6 Contrary Views on the Main Question”)

Lady Lovelaces argument

> “The Analytical Engine has no pretensions to *originate* anything. It can do *whatever we know* how *to order it* to perform” (Turing 1950, ”6 Contrary Views on the Main Question”)

Lady Ada Lovelace var en af pionererne indenfor datalogi og var praktisk talt den første programmør. Hendes argument omhandler *originalitet*, som hun ikke mener at en maskine kan have. Douglas Hartree tilføjer at:

> “This does not imply that it may not be possible to construct electronic equipment which will 'think for itself', or in which, in biological terms, one could set up a conditioned reflex, which would serve as a basis for 'learning'. Whether this is possible in principle or not is a stimulating and exciting question, suggested by some of these recent developments. But it did not seem that the machines constructed or projected at the time had this property.” (Turing “6 Contrary Views on the Main Question”)

I citatet siger Hartree at det er muligt at opsætte en refleks, således at en maskine kan have en indlæringsevne. Hvis denne kan have en indlæringsevne, er det så ikke også muligt at skabe en kunstig intelligens? En digital computer med en uendelig stor hukommelse og uendelig hastighed ville muliggøre en kunstig intelligens, der kan imitere en maskine med adskilte tilstande.

En variant af Lady Lovelaces’ argument er, at computeren aldrig kan overraske, hvor Turing mener at computeren, som sådan, ofte overrasker. Turing blev selv overrasket af computeren, ved hans beregninger af computerens beregningstid. Denne beregning ville være et hurtigt overslag, hvilket inkluderer lidt risiko, som derfor ofte er forkert. En overraskelse, for Turing, er hvor skævt dette

overslag kunne være i forhold til den egentlige beregningstid. Dette er blot et af de ting, hvor computeren kan overraske, i henhold til Turing, men det er det stadig vigtigt at tage i mente, at overraskelser stammer fra en kreativ kilde, og det er ens om overraskelsen stammer fra en bog, menneske eller en kunstig intelligens. Turing anviser, at det er en fejltagelse, som filosoffer og matematikere ofte laver, at når et faktum præsenteres for et sind, bliver alle konsekvenser af denne åbenbare for sindet. Dette er en fejl, da man tillærer sig konsekvenserne, hvorfor en konklusion draget ud fra erfaringer, der ikke er komplette også vil være overraskende. (Turing 1950, ”6 Contrary Views on the Main Question”)

Argumentet fra kontinuitet i nervesystemet

> “The nervous system is certainly not a discrete-state machine. A small error in the information about the size of a nervous impulse impinging on a neuron, may make a large difference to the size of the outgoing impulse. It may be argued that, this being so, one cannot expect to be able to mimic the behaviour of the nervous system with a discrete-state system. ” (Turing 1950, “6 Contrary Views on the Main Question”)

Dette argument er konstrueret ud fra nogle gamle standarder til kunstig intelligens, og er derfor ikke fyldestgørende længere. I halvtredserne, da Turing fremlagde ideen om en Turingtest, var computeren af gode grunde ikke så udviklet som den er i dag. Af denne årsag mente Turing, at det var tilstrækkeligt for en kunstig intelligens at svare skriftligt. Det skriftlige svar invaliderer spørgsmålet om en kontinuerthed i nervesystem (og bevidstheden), da det ikke kan bestemmes om det ene eller andet er på spil ”bag gardinet”. Endvidere er der i dag tvivl om hvorvidt bevidstheden egentlig er en kontinuer strøm. Dette kigger vi nærmere på i afsnittet om bevidsthed. Turing affærdiger spørgsmålet, da hans test ikke kan afgøre det ene eller andet.

Argumentet fra opførslens uformalitet

> “It is not possible to produce a set of rules purporting to describe what a man should do in every conceivable set of circumstances.” (Turing 1950, ”6 Contrary Views on the Main Question”)

Denne kritik er baseret på regler for opførsel. Kritikerne siger, at det ikke er muligt at lave regler for alt mulig opførsel et menneske er i stand til at udføre. De nævner som et eksempel, at reglen for en fodgængerovergang, hvor det er almen kendt at det ved grønt er lovligt at krydse, og ved rødt er forbudt. Hvis et signal er gået i stykker og den således viser begge farver, hvilken en af de to regler skal så håndhæves?

Turing besvarer dette ved, at omstrukturere spørgsmålet til dette:

> “If each man had a definite set of rules of conduct by which he regulated his life he would be no better than a machine. But there are no such rules, so men cannot be machines.” (Turing 1950, “6 Contrary Views on the Main Question”)

Argumentet er formentlig aldrig blevet fremsagt på denne måde, siger Turing, men i sidste ende er det argumentet der bliver benyttet. Turing mener endvidere at der er misforståelse i mellem

adfærdsregler, som han definerer som legale, sociale regler, og opførslens love, som er de rent kausale love, der gør at man reagerer når man sanser noget. Det er svært at forestille sig at opførselslovene kan defineres konkret, men det mener Turing godt de kan.

> "If we substitute "laws of behaviour which regulate his life" for "laws of conduct by which he regulates his life" in the argument quoted the undistributed middle is no longer insuperable." (Turing 1950, "6 Contrary Views on the Main Questsion")

Dette skaber en problematik, da kritikerne mener, at det kun er maskiner som er underlagt regler, hvorimod mennesker ifølge Turing, også er regelbetyngede, da interaktion med andre mennesker ellers vil være besværliggjort eller nærmest umulig. (Turing 1950, "6 Contrary Views on the Main Questsion")

Dreyfus er blandt kritikerne af dette regelstyrede opførsel, og han bruger et argument omkring symbolgenkendelse. Dreyfus mener, at det er muligt at sige *generelt*, hvordan genkendelsen tager sted, hvilket kan være ved hjælp af observationer af et givent symbol, identificere det og erkende det. De generelle genkendelsestrin kan forklares udefra, men hvordan de egentlig foregår i hjernen, mener Dreyfus, at vi ikke kan identificere, i hvert fald ikke på nuværende tidspunkt. (Dreyfus 1992, s. 175)

Argumentet fra ESP

> "I assume that the reader is familiar with the idea of extra-sensory perception, and the meaning of the four items of it, *viz.* telepathy, clairvoyance, precognition and psycho-kinesis. These disturbing phenomena seem to deny all our usual scientific ideas. How we should like to discredit them! Unfortunately the statistical evidence, at least for telepathy, is overwhelming." (Turing 1950, "6 Contrary Views on the Main Question")

Ekstrasensoriske Perceptioner, såsom telepati, var på Turings tid ikke videnskabeligt undersøgt, selvom Turing selv anerkender de statistiske "beviser". Vi har i dag eftervist, med fornuftige videnskabelige beviser, at denne form for paranormal aktivitet er kategorisk falskt og de der udøver det er bedragere. På grund af disse overvældende beviser imod ekstrasensoriske perceptioner, vil vi ikke behandle emnet yderligere i nærværende tekst.

Bevidsthed

> "I'm writing a book on magic," I explain, and I'm asked, "Real magic?" By real magic people mean miracles, thaumaturgical acts, and supernatural powers. "No," I answer: "Conjuring tricks, not real magic." Real magic, in other words, refers to the magic that is *not real*, while the magic that is real, that can actually be done, is *not real magic*. (Dennett 2007, 07:50-08:18)

Daniel C. Dennett citerer her en kendt magiker og noterer sig, at på samme måde som magi, har folk det med bevidsthed. En typisk opfattelse af bevidstheden er, at den netop omgår forklaring, fordi når der fremstilles en forklaring, er svaret "det er *ikke ægte bevidsthed*". Der er en tendens

til, at folk forfalder til *mysterianisme*[6] med hensyn til bevidstheden, hvilket er et argument baseret på uvidenhed eller, som Turing skriver i indsigelserne ovenfor, dog med en lidt anderledes kontekst: ”heads in the sand” argument. Hvor Turing afviser diskussionen om bevidstheden, qua hans høflige konvention, er filosoffer meget interesserede i netop denne diskussion. Vi vil derfor fremvise Dennetts teorier om, hvad bevidstheden kan være og vise Searles argumenter mod dette synspunkt.

Afvisning af Dualisme

Dennett afviser i korte termer dualismen, hvor sind/hjerne er to adskilte, men alligevel forbundne, i *Consciousness Explained* fra 1991. Han skriver:

> ”This fundamentally antiscientific stance of dualism is, to my mind, its most disqualifying feature. (…) It is not that I think I can give a knock-down proof that dualism, in all its forms, is false or incoherent, but given the way dualism wallows in mystery, accepting dualism is giving up.” (Dennett 1991, s. 37)

Der er således en opgiven i at acceptere dualismen, da der indlejret er en omgåelse af videnskabelig bevisførelse, hvilket betyder at, i det dualistiske synspunkt, vil sindet *ikke kunne* bevises. Dette peger tilbage på Descartes tro på, at der eksisterer en immateriel sjæl *udover* den fysiske krop, som alligevel har en indvirkning på denne. Dette afviser Dennett.

Searle afviser også dualisme, dog med andre implikationer, som kan ses nedenstående:

> ”Consciousness has a first-person ontology; neuronal processes have a third-person ontology. For that reason you cannot ontologically reduce the former to the latter. Consciousness is thus an aspect of the brain, the aspect that consists of ontologically subjective experiences. But there are not two different metaphysical realms in your skull, one ’physical’ and one ’mental.’ Rather, there are just processes going on in your brain and some of them are conscious experiences.” (Searle 2004, s. 89)

Searle opstiller en anden form for todeling, hvor bevidstheden er eksplicit førstepersons ontologi og de fysiske processer i hjernen er tredjepersons ontologi; de fysiske processer kan således observeres *udefra*, hvor bevidstheden kun kan ses indefra. Dette er en forunderlig blanding af dualisme og monisme. Searle påstår selv, at han er hverken materialist eller dualist, med følgende argumenter:

> ”I say, “Consciousness is just a brain process” (…) but what I mean is that consciousness as an irreducibly qualitative, subjective, first-personal, airy-fairy, and touchy-feely phenomenon is a process going on in the brain.” (Searle 2004, s. 88)

Her afviser han den materialistiske tilgang til bevidstheden, som siger at bevidstheden som irreducérbart kvalitativt, subjektivt (…) fænomen ikke eksisterer, ved at påstå at de netop eksisterer som en del af hjernen.

> “I say, “Consciousness is irreducible to third-person neurobiological processes” (…) what I mean is that consciousness is causally reducible but not ontologically reducible. It is part of the ordinary physical world and is not something over and above it.” (Searle 2004, s. 88)

[6] En tro på at noget er så mystisk, at det aldrig kan forklares.

Her afviser han den klassiske Cartesianske dualisme, som siger at bevidsthedens irreducérbarhed allerede antyder, at bevidstheden er noget over og andet end den neurologiske base. Searles bevidsthedsteori eksisterer således i et grænseland imellem materialismen og dualismen, hvor bevidstheden eksisterer i den fysiske hjerne, men med en lettere dualistisk tilbøjelighed, da den kun kan tilgås fra førsteperson; en fænomenologisk tilgang til den.

Det Cartesianske Teater

Når vi først har afvist den cartesianske dualisme, er næste skridt at se på den monistiske afstikker af den, der kaldes den cartesianske materialisme, som forstiller sig et centraliseret sted i hjernen, hvor "det hele" samles. Dette kalder Dennett *det Cartesianske Teater*, og han bruger derved en metafor for, hvor bevidstheden i denne teori er placeret. Dette centrale center er en af de mest vedholdende ideer ved bevidsthedsteori, som bliver ved med at vise sig i nye klæder. En af grundene til denne ide bliver ved med at komme tilbage, siger Dennett, er den naive opdeling mellem "jeg" og "verden udenfor". Når vi oplever at vi ikke kan bevæge en arm, fordi den er faldet i søvn eller er blevet lam, bliver "jeget" mindre og verden udenfor trænger sig ind i os. På samme måde vil vi ikke forvente at kunne se, hvis vores optiske nerve blev overskåret. Den visuelle oplevelse foregår således et sted imellem mine øjne og min stemme, når/hvis jeg fortæller hvad jeg ser. Et sted imellem disse to ekstremiteter skulle man således kunne finde det Cartesianske Teater. Den store fejltagelse i teater-teorien er, at der bliver antydet *et mål*, som sanseindtryk skal nå, før fænomenet opleves, hvor man, ifølge Dennett, begynder at opleve et fænomen allerede når et billede indskrives på øjet. (Dennett 1991, s. 108)

Dette argument falder i sidste ende tilbage på et homunculus-argument, hvor sjælen før i tiden havde rollen som en individuel observant af vores oplevelser, og derved var en lille mand – en homunculus – adskilt fra os selv, men som alligevel havde indflydelse på vores handlinger. En mand, der sidder ved kontrolcenteret i hjernen og ser hvad vi ser, og styrer hvordan vi bevæger os. Dette argument fejler idet, at det skaber en uendelig regression, da den lille mand, også må have en lille mand[7] osv. Hvis regressionen afbrydes, og man således ender på *den sidste homunculus*, vil man være tvunget til at forklare, hvordan denne forstår og tænker, og man har derfor blot forskudt problemet.

Adskillige Udkast-modellen

Dennett (1991) introducerer *Adskillige Udkast-modellen*, som et alternativ til det Cartesianske Teater. Dennett skriver følgende:

> "According to the Multiple Drafts model, all varieties of perception – indeed, all varieties of thought or mental activity – are accomplished in the brain by parallel, multitrack processes of interpretation and elaboration of sensory input." (Dennett 1991, s. 111)

[7] Det skal klargøres, at homunculus er naturligvis er en metafor og den "lille mand" ikke er en fysisk lille mand, men en form for samlingssted for information.

Der er således *adskillige udkast* til forståelse hele tiden. Dette betyder, at der udføres en konstant revision og fortolkning af den akkumulerede information. Vi oplever således ikke, hvad der sker i vores øjne eller vores ører – eller for den sags skyld i alle de andre sanser – vi oplever et produkt af mange forskellige tolkningsprocesser. Det specielle ved denne model er, at en given diskriminering af en sansning kun sker én gang. For hvert stimuli, startes der en handlingskæde, som gradvist afkaster en skelnen med større og større detaljegrad. Visuelle stimuli bevirker, at dele af hjernen går i tilstande, der skelner mellem forskellige særpræg:

> ”First mere onset of stimulus, then location, then shape, later color (through a different pathway), later still (apparent) motion, and eventually object recognition.” (Dennett 1991, s. 134)

Der er således flere specialiserede, lokaliserede dele af hjernen som fikserer indholdet i en observation. Der behøves derfor ikke en *re*præsentation af informationen et andet sted i hjernen, i et Cartesiansk Teater.

Det er muligt at placere denne indholdsfiksering i både rum og tid, men dette hjælper os ikke specielt, da indholdsfikseringen ikke betyder, at informationen bliver bevidst. Spørgsmålet, hvornår information bliver en bevidst oplevelse, er et forvirrende spørgsmål, da meget af den givne information aldrig bliver bevidst og blot ”overses”. Dennett skriver:

> ”These distributed content-discriminators yield, over the course of time, something rather like a narrative stream or sequence, which can be thought of as subject to continual editing by many processes distributed around in the brain, and continuing indefinitely into the future.” (Dennett 1991, s. 113)

De mange individuelle processer afkaster den narrative tankestrøm, som vi oplever som bevidstheden. Dette er en model som intuitivt ikke giver mening, da vi individuelt oplever vores egen tankestrøm og har personlig erfaring med denne. Dette er fristelsen ved det Cartesianske Teater – det er intuitivt nemt at forstå – dette betyder dog ikke at det er sandt.

Hvornår er en given observation så tilgængelig i bevidstheden, kan man spørge. Dette er ikke altid tydeligt for den enkelte, der oplever det, og det kan ikke fast defineres uafhængigt af når vi spørger ind til det. Når vi prøver at opfange, hvad der sanses, gør vi bevidstheden opmærksom på netop det der sanses. Eksempelvis, er der et klassisk eksempel på, hvad man intuitivt kan kalde ”ubevidst bevidsthed” følgende:

> ”Are you constantly conscious of the clock ticking? If it suddenly stops, you notice this, and you can say right away what it is that has stopped; the ticks ’you weren’t conscious of’ up to the moment they stopped and ’would never have been conscious of’ if they hadn’t stopped are now clearly in your consciousness” (Dennett 1991, s. 137)

Dennett kalder dog ikke dette ubevidst tænkning, da man ikke nødvendigvis ikke er bevidst om uret. Han kalder det i stedet ”rolling consciousness with swift memory loss”. Dette baserer han på, at man aldrig ville have tillagt observationen til hukommelsen, hvis ikke man blive signaleret ved at urets tikken stopper. En observation kan således komme op i bevidstheden, ved at gøre opmærksom på sig selv, men den givne person er ikke opmærksom på det – bevidst om det – før

det *er* sket. Dette modarbejder den centraliserede "kontrolrums"-tankegang i den cartesianske materialisme.

Dennett konkluderer at bevidstheden således er en godartet illusion, som skabes af de mange udkast ved at et udkast påkalder sig opmærksomhed.

Heterofænomenologi

> "Serious phenomenology is in even greater need for a clear, neutral method of description, because, it seems, no two people use the words the same way, and everybody's an expert." (Dennett 1991, s. 66)

To mulige problemer med den klassiske fænomenologi, som Dennett kalder *autofænomenologi*, er, at den subjektive introspektion bliver taget for gode varer, og at den bliver generaliseret til alle andre mennesker. Vores personlige observation of vores egen bevidsthed er ikke så ufejlbarlig, som de fleste mener. Siden Descartes *Cogito ergo sum*, er vores egen tilgang til vores bevidsthed blevet regnet for ufejlbarlig (altid rigtig), eller i hvert fald uforbederlig (aldrig muligt at korrigere), da en person udefra ikke har vores egen eksklusive adgang.

> "The problem with autophenomenology is not that it is (always, or typically) victim to illusion and distortion but that it is (always) vulnerable to illusion and distortion." (Dennett 2006, s. 19)

I sit forsvar mod heterofænomenologiens kritikere, anfører Dennett endnu en gang svaghederne ved den klassiske autofænomenologi. Disse svagheder er forholdsvis nemme at fremvise, da et givent subjekt kan *snydes*, således at hjernen producerer et falskt billede i bevidstheden. Vi henviser til McGurk effekten[8] og neon-color spreading effekten[9], og der kan findes flere af disse typer effekter. Med den klassiske fænomenologi kan man argumentere for, at det vi oplever netop er det der er på grund af den immanente førstepersons tilgang, men med den videnskabelige tredjepersons tilgang, kan vi se, at sanserne bliver snydt, og at vi derfor ikke udelukkende kan stole på dem. Dette invaliderer fænomenologien alene som videnskabelig proces, da den ene ikke kan reduceres til den anden.

Heterofænomenologi kombinerer ordinære førstepersons fænomenologiske observationer med tredjepersons videnskabelige observationer, for således at kunne determinere om en given person tager fejl om sit eget sind. Der er en immanent neutralitet i heterofænomenologi, da man hverken skal acceptere eller udfordre subjektets påstande, men i stedet vedholde en konstruktiv og sympatisk neutralitet, for derved at kunne samle en definitiv forklaring af verden i overens-stemmelse med subjektet.

Pointen med heterofænomenologi er, at den definerer alt der kan vides om bevidstheden, da den sammenhængt med neurologiske data, der kan samles fra et subjekts hjerne udgør det totale

[8] http://www.youtube.com/watch?v=aFPtc8BVdJk. Lydsporet siger "ba-ba" og billedsporet 'siger' "ga-ga", kombineret hører man "da-da". Prøv at lukke øjnene og høre, og derefter se på videoen samtidig.
[9] http://www.blelb.ch/english/blelbspots/spot05/expspot05_en.htm, her skabes illusionen af en farve, hvor der egentlig er hvidt.

mængde information der skal bruges for en teori om bevidstheden. For at kunne studere bevidstheden er det derfor oplagt at benytte denne metode og afskrive sig den klassiske fænomenologi.

Bevidsthedens Vanskelige Problem..?

En af de indlejrede og vedholdende ideer med hensyn til bevidstheden er, at der eksisterer et specielt vanskeligt problem, der gør at vi ikke kan forklare den i alle detaljer. David Chalmers (1995) fremstiller som den første det såkaldt *vanskelige problem* ved bevidsthed, ved at adskille det fra alle de (relativt) lette problemer. Han definerer de to som følgende:

> ”The easy problems of consciousness are those that *seem directly susceptible to the standard methods of cognitive science*, whereby a phenomenon is explained in terms of computational or neural mechanisms. The hard problems are those that *seem to resist those methods*.” (Chalmers 1995, afsnit 2. Vores fremhævninger.)

Det vanskelige problem, ifølge Chalmers, er *oplevelse*. Udover den store informationsbearbejdning der foregår, når vi sanser, er der et subjektivt aspekt, hvad Chalmers kalder oplevelse[10]. Vi oplever følelsen af farver, af lyde og musik, men hvorfor gør vi dette? I sidste ende er oplevelse bevidsthedens ”magi”, som gør at den undgår eller omgår videnskabelig forklaring. Chalmers udvider endvidere dette problem med, at hvert af de ”lette” problemer, såsom hukommelse, perception og indlæring, er associeret til bevidstheden, og at vi kan gå til disse problemer med kognitiv videnskab og derved forklare disse funktioner ved at forklare deres ydelse, men dette gælder ikke for oplevelse. Som han forklarer:

> “What makes the hard problem hard and almost unique is that it goes *beyond* problems about the performance of functions. To see this, note that even when we have explained the performance of all the cognitive and behavioral functions in the vicinity of experience - perceptual discrimination, categorization, internal access, verbal report - there may still remain a further unanswered question: *Why is the performance of these functions accompanied by experience?*” (Chalmers 1995, afsnit 3)

Det vanskelige problem går *ud over* problemerne med funktionsydelsen, hvilket betyder at vi ikke kan studere bevidstheden ved blot at se på hjernens funktioner.

Chalmers bringer selv en parallel med vitalistisk skepticisme, som går ud på at vitalister tvivlede om de fysiske processer kunne forklare livet, og derfor fremlagde hypotesen af en ”vital ånd” var en alternativ forklaring. Da det viste sig at de fysiske processer forklarede alle disse funktioner, forsvandt den vitalistiske tvivlen. Han skriver:

> “What drove vitalist skepticism was doubt about whether physical mechanisms could perform the many remarkable functions associated with life” (Chalmers 1995, afsnit 5)

Han mener dog ikke at sin egen bevidsthedsmæssige skepticisme fungerer på samme måde:

[10] Oplevelse bliver også i visse kredse kaldet for ”Qualia” og er den subjektive oplevelse af ”ting”. F.eks. farven rød.

> "With experience, on the other hand, physical explanation of the functions is not in question. The key is instead the *conceptual* point that the explanation of functions does not suffice for the explanation of experience." (Chalmers 1995, afsnit 5)

Dennett mener at Chalmers simpelthen tager fejl i denne antagelse, og siger:

> "Whether people realize it or not, it is precisely the "remarkable functions associated with" consciousness that drive them to wonder about how consciousness could possible reside in a brain. In fact, if you carefully dissociate all these remarkable functions from consciousness – in your own, first-person case – there is nothing left for you to wonder about." (Dennett 1995)

Dennett afviser derved ideen, at "oplevelse" går ud over de kognitive funktioner, og noterer sig at hvis man afskærer sig de enkelte funktioner, vil man ende med intet. Bevidstheden – eller oplevelsen – er altså intet i sig selv, men derimod et produkt af alle de kognitive funktioner.

I sidste ende mener Dennett at man skal lade være med at stirre sig blind på "det vanskelige problem", da man så ikke kan se sammenhængen ved alle de lette problemer.

Bevidstheden kan ses som et magisk trick, hvor hele tricket er overstået før det er startet. Fordi vi allerede har defineret den som bevidsthed***en***, afskriver vi muligheden for at den slet ikke eksisterer og blot er et produkt af alle de individuelle processer i hjernen. Man kan således sige at bevidsthedens *magi* kun omgår forklaring, så længe vi accepterer den på dens egne præmisser og tager den for gode varer. Når vi først accepterer alle de ikke-mystiske måder, hvorpå hjernen kan skabe godartede illusioner, kan man begynde at lave en fornuftig teori om hvordan hjernen skaber bevidsthed.

Dennetts teori om Adskillige Udkast er på mange måder understøttende for en Stærk AI, da den fokuserer på opførslen, på samme måde som Turing testen. De underliggende funktioner i sindet, skaber bevidstheden, ved at gøre opmærksom på udkast og konstant revidere og revurdere det information, der behandles.

Implementeringsmæssige begrænsninger ved en Stærk Kunstig Intelligens

Vi vil i dette afsnit vise de implementeringsmæssige begrænsninger og problemer, der eksisterer ved en Stærk AI, som Hubert L. Dreyfus fremfører. Han er kritiker af teorien om en stærk AI, der er bevidst, og mener at vi ikke kan føre teorien ud i livet med digitale computere, da vi er nødt til at bruge komponenter, der er tilstrækkeligt tilsvarende menneskets egne dele. Dette vil muligvis kunne lade sig gøre, men det er ikke grundlaget for hans kritik, som er rettet mod digitale informationsbehandlingsmaskiner. (Dreyfus 1992, s. 250-251)

Klassifikationer af Kunstig Intelligens

Før vi kommer til Dreyfus' kritik, vil vi først forklare de to overordnede paradigmer indenfor AI-forskningen.

Klassisk Kunstig Intelligens

Klassisk AI, som også kaldes symbolsk AI, fordi den regner kognition for symbolmanipulation. Dette paradigme bliver støttet af forskerne Allen Newell og Herbert Simon, der i 1976 proklamerer, at symbolgenkendende computere på daværende tidspunkt er i stand til at udvise intelligens i form af generelle operationer. (Chalmers 1992, s. 7)

Turing, som vi har dækket tidligere i nærværende tekst, er også bag klassisk AI. Han skaber grundlaget for computeren, som den eksisterer i dag, med sin teoretiske Turingmaskine. Den klassiske AI er underlagt det samme regler for beregnelighed, som Turings oprindelige maskine, hvorfor den er serielt baseret.

Klassisk AI er syntaktisk baseret, da de kognitive processer er syntaktisk frembragte sekvenser af symboler. En klassisk AI er bundet af dens serielle beregningsmetode og faste løsningsmodeller. Fordi en klassisk AI er nødt til at være *forprogrammeret*, bliver den således komplet deterministisk, således at to maskiner, med samme programmering ender med det samme svar til et givent spørgsmål. (Turing 1936, ”On Computable Numbers, With An Application To The Entscheidungsproblem”)

I processen for at få viden formidlet til en klassisk AI, skal alt input atomiseres, fordi en klassisk AI ikke er lærende, men al viden skal forprogrammeres, hvilket betyder, at en klassisk AI har al dens viden fra dens ”fødsel”. Dette medfører et problem, da en klassisk AI ikke kan håndtere situationer, som ikke er forudset af programmørerne, og derfor altid er statisk. Klassisk AI skyder sig selv således i foden, da den, for at kunne eksistere i verden på lige vilkår med mennesker, er nødt til at formalisere *al viden*, hvilket er en umulighed.

Moderne Kunstig Intelligens

En moderne AI, eller forbindelses AI, er til modsætning af klassisk AI et lærende system, der ikke fra starten har nogen specifik viden, men skal ligesom en menneskelig intelligens tillære sig det gennem oplevelser og erfaringer deraf.

Den moderne AI baserer sig på et netværksparadigme, da den bruger parallel beregning, modsat den klassiske AI, som er seriel. Dette giver mulighed for flere simultane databehandlinger, hvor klassisk AI er låst fast til en beregning af gangen. Den parallelle databehandling sker ved hjælp af et *neuralt netværk*. Et neuralt netværk er en simulering af menneskehjernen, som er et netværk af individuelle neuroner, som hver har forbindelse med alle andre gennem flere led. Et neuron for sig selv, har ikke nogen større effekt, mens det netværk et neuron er medlem af, skaber forskellige tilstande. En enkel neuron bærer intet semantisk viden i dette paradigme, men i samarbejde kan den semantiske forståelse opstå. (Chalmers 1992, s. 7)

Modellen bag det neurale netværk er, udover at den er kompleks, meget effektiv idet at den paralleliserer opgaver. Endvidere tilføjer Brunak og Lautrup at:

Hvis der er for mange processorer, kan de ikke bringes til at virke sammen i et meget hierarkisk system. Den hierarkiske struktur gør det besværligt at konstruere et effektivt kommunikationsmønster, der kan tilpasses mange forskellige beregningsopgaver. (Brunak og Lautrup 1990, s. 82)

Det vil sige at det neurale netværk ikke er hierarkisk, men er forbundet parallelt i ét plan, hvilket øger beregningshastigheden.

Biologiske bemærkninger af Dreyfus

Dreyfus angriber den kunstige intelligens fra flere vinkler. Vi behandler i nærværende tekst to tilgange, de biologiske, fysiske problemer ved at genskabning af menneskets hjerne, og de psykologiske og bevidsthedsmæssige aspekter i denne genskabning.

I det biologiske aspekt, mener Dreyfus at det er vigtigt at pointere, at hjernen ikke udelukkende består af neuroner, og derfor er det ikke fyldestgørende udelukkende at lave nøjagtige kopier af hvert enkelt neuron i hjernen for derved at lave en komplet udskiftning fra biologiske til mekaniske dele i hjernen. Selv hvis man kunne lave denne nueron-udskiftning, ville der stadig eksistere kemiske blandinger, som påvirker neuronerne, og i visse tilfælde giver følelser og andre tilstande. (Dreyfus 1992, s. 160)

I bogen *Mind over Machine* skriver Dreyfus og hans lillebror Stuart E. Dreyfus, at der er tre niveauer, der skal forstås før det er muligt at forklare, hvordan noget virker. Disse tre niveauer er: *partikel*, *komponent* og *funktionsniveauet*. Disse tre niveauer er vigtige at klargøre for at give en forståelse for hvordan en maskinel eller noget biologisk væsen fungerer. Hvert niveau bliver opdelt for at påvise hvilke dele der skaber maskinen. Inddelingen i niveauer er nødvendig, fordi de beskriver hver deres syn på organismen eller mekanismen, som ikke kunne være dækket af et enkelt niveau for det hele.

På partikelniveauet beskrives de helt basale fysiske sammenhænge, og der kan determineres hvilket stof, der er med at gøre. På partikelniveauet kan man ikke forklare funktionelle egenskaber af en maskine, da dette kun kan udledes på et "højere" niveau.

På komponentniveauet begynder man at beskrive de enkelte dele i en maskine, og i en computer vil dette være de enkelte transistorer og kredsløb, og stadig på samme niveau eksisterer processorer og enkelte dele i computeren.

Det funktionelle niveau, som Dreyfus regner for det vigtigste, i denne sammenhæng, kan man forklare hvordan en enhed fungerer. En funktionel tilgang til en bil er relativt nem, da man kan se på alle delene for sig, som at skabe ydelsen ved den holistiske sammensætning af dem. Når man beskriver en computer, bliver det straks mere kompliceret, da der eksisterer, hvad Dreyfus kalder, en *virtuel maskine*, på computeren. I denne virtuelle maskine er der dele af software der udfører hver deres del, og i det samlede hele skaber den samlede softwares funktion. En computer, der

genkender symboler kan ikke deles op med fysiske dele, da nogle kun eksisterer rent virtuelt på computeren, og disse udfører blot logiske operationer internt i maskinen. Det er med denne baggrund, at teoretikere af klassisk AI, anser hjernen for at operere ligesom en computer, da det er et samarbejde af de forskellige komponenter, der giver et samlet resultat. (Dreyfus og Dreyfus 1972. S. 61-62)

Når man anskuer hjernen med de tre niveauer, som Dreyfus brødrene fremstiller, bliver proponenter af den klassiske AI sat til skamme, som Dreyfus skriver:

> "This Model is still uncritically accepted by practically everyone not directly involved with work in neurophysiology, and underlies the naîve assumption that man is a walking example of a succesful digital computer program." (Dreyfus 1992, s. 159)

Dreyfus siger her, at det er direkte forkert og naivt at anse hjernen som en digital databehandler. Det er altså kun individer, der ikke er involverede i neurofysiologi, der kan accepterer hjernen som en digital maskine. Dog skal der nævnes at hjernen også indeholder binære signaler i form af impulser til nervesystemet, men her nævner Dreyfus en anden teoretiker John von Neumann[11], der siger at:

> "The neuron transmits an impulse. … The nerve impulse seems in the main to be an all-or-none affair, comparable to a binary digit. This a digital element is evidently present, but it is equally evident that this is not the entire story.
> ... It is well known that there are various composite functional sequences in the organism which have to go through a variety of steps from the original stimulus to the ultimate effect- some of the steps being neural, that is, digital, an others humoral, that is, analog." (Dreyfus 1992, s. 160)

Dreyfus påpeger således igennem von Neumann, at selvom det er rent binære signaler, er der andre ting der har indvirkning på det egentlig signal, såsom intervaller mellem affyring. Impulsfrekvensen er analog, da den ikke er taktfast. Når hjernen således benytter et analogt system, følger det naturligt at det ikke er det samme som en digital maskine. For Dreyfus opstår der en tvivl om, hvorvidt hjernen behandler information på samme måde som en klassisk kunstig intelligens, da information kan være fordelt flere steder i hjernen. Dette ligner op med hvad Dennett skriver om bevidstheden; at hjernen netop fordeler information til forskellige delle af hjernen, og præcis *ikke* centrerer informationen et enkelt sted i hjernen i et cartesiansk teater. (Dreyfus 1992, s. 160)

Fra vores synspunkt virker det som Dreyfus angreb på den kunstige intelligens fra det biologiske aspekt, udelukkende rammer den klassiske AI. De forskelle og mangler som Dreyfus anfører, passer netop overens med forskellen mellem klassisk og moderne AI. Man kunne blive tilbøjelig til at tro, at med en fyldestgørende indsigt i netværksparadigmet, ville Dreyfus slippe sin kritik af den kunstige intelligens eller i hvert fald frafalde nogle af disse kritikpunkter.

[11] John von Neumann var den første til at praktisk bygge en fungerende Turingmaskine, og regnes sammen med Turing for faderen til den moderne computer.

Psykologiske bemærkninger af Dreyfus

Dreyfus kritiserer kunstig intelligens fra et psykologisk perspektiv, hvor han slår ned på at vi mennesker har et hjernen som en analogi på vores nyeste opfindelser, og computeren blot er nyeste skud på denne stamme:

"In the period between the invention of the telephone relay and its apotheosis in the digital comuter, the brain, always understood in terms of the lastest technological inventions, was understood as a large telephone switchboard or, more recently, as an electronic computer." (Dreyfus 1992. S. 159)

Mennesker er således uden egentlig forståelse af hjernen, men får det introduceret som den nyeste teknologi på markedet. Der laves således en simplificering af hjernen, som dog stadig bliver mindre simpel end forgængeren. Der simplificeres derfor til den serielle computer, da vi endnu ikke har noget der virker bedre som analogi. Gennem kritikken af denne simplificering tager Dreyfus afstand til sammenligningen mellem en digital computer og den menneskelige hjerne. Dreyfus selv er kritiker af kunstig intelligens, og ikke mener at det muligt at skabe kunstig intelligens med den teknologi vi har i dag, og som skrevet i det foregående afsnit er det ikke kun databehandlingsdelen der ikke er fyldestgørende, men der mangler også nogle kemiske blandinger, før det er muligt at skabe en kunstig intelligens. Han mener derved, der ikke er et beregneligt svar på alle mulige opførsler, hvilket han ekspliciterer i følgende citat:

"There is, indeed not the slightest justification for the claim that 'for each type of behavior in the repertoire of that organism, a putative answer to the question, How does one produce behavior of that type? takes form of a set of specific instructions for producing the behavior by performing a set of machine operations.'" (Dreyfus 1992. S. 167)

Dreyfus siger altså at opførsel ikke er beregnelig og da dette er tilfældet, er kunstig intelligens dømt til at mislykkes, så længe teknologien hænger fast i denne serielle databehandling. Argumentet om forprogrammering dukker her tilbage til overfladen, da Dreyfus ikke mener at alle mulige opførsler kan programmeres i forvejen. Han mener således at sammenligningen mellem den digitale computer og vores analoge hjerne er absurd, da de to fungerer på hver deres måde.

Hjernens unikke behandling af data er også tydelig i måden hvorpå der håndteres og lagres data. Hvor en digital maskine vil lave en streng med bits og gemme disse et unikt sted i dens hukommelse, vil en analog hukommelse tillægge mening til data, ifølge Dreyfus. Et eksempel på dette er ved ordet "mor", hvor en digital databehandler blot vil gemme strengen binært som en lang talrække af nuller og ettaller, vil en analog databehandler tillægge meninger til ordet, såsom referencer til ens egen mor, andres mødre, eller en forståelse af ordet mor osv. Ved at lagre ordet med mening og forståelse er der således en semantisk tilgang til data.

Semantik er en af nutidens mere vedholdende problemer for den kunstige intelligens, da den kunstige intelligens ikke kan forholde sig til verden omkring den. Hvis en kunstig intelligens ikke

har en semantisk forståelse af verden, vil den ikke kunne gennemføre en Turingtest, da den ikke har den fornødne viden om verden. Vi udforsker dette aspekt i næste afsnit vedrørende Searles Chinese Room argument.

En af Dreyfus' implementeringsmæssige argumenter mod kunstig intelligens går ud på, at at computere ikke kan lære nye ting. Den klassiske AI fungerer ganske vist på denne måde, vil vi medgive, men igen passer Dreyfus' argument ret præcist på den moderne AI, som netop kan og *skal* lære, for at fungere.

Vi finder, at de fleste af Dreyfus' modargumenter bliver overraskende godt svaret af den moderne AI, som opfylder mange af de krav som han stiller op for en AI. Der er stadig biologiske problemer, såsom at den menneskelige hjerne er analog, hvor computeren fungerer på et digitalt niveau, endnu.

Searles Chinese Room argument

Vi vil i dette afsnit behandle *Searles Chinese Room argument*, ved at kontrastere det med Dennetts syn på argumentet.

Argumentet fra Searle

Argumentet kommer af følgende tankeeksperiment. Forestil dig at Searle, en engelsktalende person, bliver låst inde i et rum fyldt med bokse med kinesiske symboler (en database) og med en bog med instruktioner til at manipulere symbolerne (et program). Forestil dig så at personerne udenfor boksen sender andre kinesiske symboler ind, som, ukendt for personen, er spørgsmål på kinesisk (input). Ved at følge instruktionerne bogen er det så muligt for personen give symboler tilbage, som er de rigtige svar på disse spørgsmål (output). Programmet gør det således muligt for personen i rummet – eller input/output systemet – at gennemføre Turingtesten på kinesisk, uden at personen på nogen måde forstår kinesisk. (Searle 1980, s. 284)

Pointen fra Searle er således, at fordi personen ikke forstår kinesisk er der ikke nogen semantisk forståelse, og derfor er det blot en simulering af forståelse, hvilket betyder at når man ser eksperimentet som en lignelse, vil denne aldrig kunne være en Stærk AI i Searles brug af ordet, som tidligere beskrevet. Han uddyber forskellen mellem at have forståelse og ikke at have det ved at lave samme eksempel med engelsk input:

> "From the outside my answers to the English and the Chinese questions are equally good. I pass the Turing test for both. But from the inside, there is a tremendous difference. (…) In English, I understand what the words mean; in Chinese, I understand nothing. In Chinese, I am just a computer." (Searle 2004, s. 63)

Der er således en indlejret forståelse af ordene, når inputtet er på engelsk, hvor det er ren symbolmanipulation på kinesisk for Searle.

Dennetts modargument

Dennett mener at Chinese Room argumentet er en *intuitionspumpe*[12], der er afhængig af at vildlede opmærksomheden. Han bemærker at:

> "while philosophers and others have always found flaws in his thought experiment when it is considered as a logical argument, it is undeniable that its "conclusion" continues to seem "obvious" to many people. Why? Because people don't actually imagine the case in the detail it requires." (Dennett 1991, s. 436)

Searle antager rummet kan gennemføre en Turingtest, men dette indebærer mere end de fleste overvejer. Dennett fremviser et længere eksempel på en samtale mellem det kinesiske rum og dommeren af Turingtesten (Dennett 1991, s. 436-437), hvor en joke bliver forklaret af rummet[13]. At et så sofistikeret svar, som der bliver givet, kan produceres af Searles flittige arbejde er straks mere vanskeligt, end antagelsen at man blot skal *forestille sig* at rummet kan gennemføre en Turingtest. (Dennett 1991, s. 437)

Dennett siger endvidere, at for at tankeeksperimentet skal virke er det vigtigt at Searle inviterer os til at forestille os, at han arbejder med uforståelige kinesiske tegn i stedet for nuller og ettaller, da det i hans ord måske kan:

> "(…) lull us into the (unwarranted) supposition that the giant program would work by somehow simply "matching up" the input Chinese characters with some output Chinese characters. No such program would work, of course" (Dennett 1991, s. 437)

Dennetts argument er, at et program der skulle kunne gengive et fornuftigt svar, som det eksempel han viser, er nødt til at være et:

> "extraordinarily supple, sophisticated, and multilayered system, brimming with "world knowledge" and meta-knowledge and meta-meta-knowledge about its own responses, the likely responses of its interlocutor, its own "motivations" and the motivations of the interlocutor, and much, much more" (Dennett 1991, s. 438)

Ifølge Dennett benægter Searle ikke at rummet har denne enorme struktur, men han afholder os blot fra at tænke på det. Hvis vi til gengæld også overvejer det, som vi er nødt til, for at have en ordentlig forståelse af tankeeksperimentet, så er det ikke længere så intuitivt, at der ikke nødvendigvis eksisterer en ægte forståelse af joken. Hvis man nu er i tvivl om hvorvidt der er forståelse, peger det i retning af at Searles tankeeksperiment afhænger af at man forestiller sig eksemplet for simpelt og udleder den åbenlyse konklusion.

Searles vildledning af læseren sker i det, at han antager at *han* ikke forstår kinesisk, selvom Turingtesten bliver gennemført. Dennett påpeger at Searle kun er en lille del af maskineriet og derfor er det kun naturligt, at *han* ikke forstår noget, på samme måde som at der ikke er forståelse

[12] Se en nærmere forklaring af intuitionspumper, et term opfundet af Dennett på følgende webpublicerede artikel http://www.edge.org/documents/ThirdCulture/r-Ch.10.html, fra *Third Culture: Beyond the Scientific Revolution*.
[13] Se bilag 1 for Dennetts fulde eksempel, som vi ikke vil indsætte direkte i teksten.

for noget i en lille del af programmering, da det blot er symbolmanipulation. Her afslører Dennett Searles underliggende præmis:

> "Surely, *more of the same*, no matter how much more, would ever add up to genuine understanding." (Dennett 1991, s. 438)

En cartesiansk dualist ville helt og holdent fastholde dette synspunkt, da der kræves en sjæl for forståelse. Hvis man til gengæld accepterer, at hjernen uden mirakuløs hjælp, selv skaber forståelsen må man, i henhold til Dennett, indrømme at forståelse opstår som en proces af interaktioner mellem mange undersystemer, som slet ikke forstår noget på egen hånd. Hvis man følger argumentet, at "denne lille del af hjernen (x) forstår ikke kinesisk, og denne lille del, hvoraf x er en del, heller ikke forstår kinesisk" kommer man frem til den uheldige konklusion at hele hjernen ikke er nok til at kinesisk.

> "It is *hard to imagine* how "just more of the same" could add up to understanding, but we have very good reason to believe that it does, so in this case, we should try harder, not give up." (Dennett 1991, s. 439)

Det er således dybt anti-intuitivt at forestille sig Dennetts synspunkt, men dette i sig selv invalidere ikke teorien, da intuitivhed kan snydes ligeså let som sanserne. Dennett ser det kinesiske rum, som et sæt af subsystemer således, at Searle, som en lille del af rummet, ikke forstår kinesisk i sig selv, men i forbindelse med regel bogen og alle de eventuelle hjælpemidler der eksisterer i rummet fremstår der et *selv*, som er hele systemet, og det er dette selv vi skal tilskrive forståelse. Hvis vi skal acceptere, at Searles kinesiske rum fungerer og kan gennemføre en Turingtest, er det således systemet vi skal tilskrive forståelse. Searle inviterer os til at forestille os, at et simpelt genkendelsestabelsystem udgør hele det enorme program, der skal til for at kunne gennemføre en Turingtest, og til det siger Dennett at:

> "We have no business imagining such a simple program and assuming that *it* is the program Searle is simulating, since no such program would pass the Turing test, as advertised." (Dennett 1991, s. 439)

Vi er nødt til at indregne kompleksitet i systemet ikke bare lade som om vi gør det, hvilket er hvad Searle indirekte opfordrer til, ifølge Dennett.

> "Searle's thought experiment yields a strong, clear conviction only when we fail to follow instructions. These intuition pumps are defective; they do not enhance but mislead out imaginations." (Dennett 1991, s. 440)

Dennett afviser herved Chinese Room argumentet på baggrund af, at det bevidst vildleder læseren for at kunne fungere. Der er en forståelse på system-niveau, således at Searle, regel bogen og alle de andre hjælpemidler, der eksisterer i rummet, hvis der kan opnås fornuftige svar, har en form for forståelse, selvom hver del ikke har nogen individuel forståelse; hvis denne individuelle forståelse skal forestilles, falder man endnu en gang tilbage på et homunculus-argument.

Argumenter mod Chinese Room

Kompleksitet

Dennett henleder allerede til et kompleksitetsproblem med det kinesiske rum. Fordi regel bogen skal inkludere *alle mulige svar*, skyder eksperimentet sig selv i foden, da det skal indeholde svar, der ikke er mulige at have før diskussionen er i gang. Den meta-meta-viden, som kræves for at kunne svare på et spørgsmål, der refererer til et spørgsmål tidligere i samme samtale, *kan ikke eksistere* som et forud programmeret svar, om det så er i en computer eller i en regel bog. Der er derfor nødt til at være en semantisk forståelse for at kunne generere et svar på dette spørgsmål. På denne baggrund kan vi afvise Searles kinesiske rum, da det end ikke er logisk at forestille sig.

Hastighed

Hvis man tilsidesætter al logik, og alligevel accepterer at rummet kan gennemføre en Turingtest, er der et hastighedsproblem. Det er estimeret at vores hjerner opererer med, i hvert fald 100.000.000.000 operationer i sekundet og hvis dette skulle eftergøres i det kinesiske rum, skulle den ellers arbejdsomme Searle kunne søge igennem sin regel bog hurtigere end vi kan forestille os, og end hvad der er fysisk muligt. På denne baggrund kan vi afvise Searles kinesiske rum, da det fysisk på ingen måde kan lade sig gøre. Vi er opmærksomme på, at som et tankeeksperiment kan man medgive visse ting, såsom at noget kun kan virke teoretisk og aldrig praktisk, men dette kan end ikke foregå på teoretisk plan. På samme måde, som man ikke kan lave et tankeeksperiment, hvor man som basis skal forestille sig at 2 + 2 = 5, for så at finde frem til at 2 < 2.

Churchlands lysende rum

Dette argument bliver fremsagt af Churchland og Churchland (1990), som laver en parallel til et andet tankeeksperiment. Hvis en filosof ville afvise Maxwells bølgeligninger, som siger at elektromagnetiske bølger skaber lys, kunne han placere sig i et mørkt rum og bevæge en magnet op og ned. Når der ikke kommer lys, kan han afvise teorien. Problemet med argumentet er, at der ikke er overvejet hastigheden – det er korrekt at der ikke kommer lys i den hastighed der svinges med, men hvis han svinger magneten 450 mia. gange i sekundet, ville der netop opstå lys.

Andre sind

Searles tankeeksperiment bliver omgået af Turing, som vi har nævnt tidligere, da Turing ikke er specielt interesseret i at diskutere, om der eksisterer et sind i et andet væsen, hvis det kan klare Turingtesten. Han har personligt den høflige konvention, at ikke betvivle eksistensen af et sind i det andet væsen. Dette argument falder tilbage på et filosofisk problem, da vi faktisk aldrig kan vide at andre end os selv *har* et sind og en bevidsthed. Vi kan kun bedømme på opførsel og på neurologisk aktivitet.

Fra dette syn kan man, og Dennett har gjort dette, fremvise et evolutionært problem ved den nærmest dualistiske opfattelse af sindet, som Searle har. Antag, at der eksisterer filosofiske

zombier[14]. Fordi disse er simplere organismer, vil naturlig udvælgelse favorisere dem og i sidste ende vil hele den menneskelige race blive udskiftet med disse zombier. Hvordan kan vi vide, at vi alle ikke er disse zombier? Der er derfor ingen evolutionær basis for at udvikle bevidstheden på den måde Searle fremstiller den.

Konklusion om Chinese Room

Searles Chinese Room kan vi med ovenstående argumenter afvise. Argumentet viser ikke at en Stærk AI ikke kan lade sig gøre, da tankeeksperimentet ikke holder vand og ikke hører sig til i en videnskabelig sammenhæng, da den logisk ikke fungerer.

Konklusion

Vi har i nærværende tekst forklaret den Stærke Kunstige Intelligens og ved brug af Daniel C. Dennetts teorier fundet frem til at, fra det rent bevidsthedsmæssige perspektiv, kan det umiddelbart godt lade sig gøre at skabe en Stærk Kunstig Intelligens. Når vi via Dreyfus anskuer bevidstheden og hjernen som muligt beregnelig, finder vi at hans kritik primært retter sig imod den klassiske Kunstige Intelligens symbolparadigme, som han med rette afviser kan huse en stærk intelligens. Den moderne netværksbaserede teori om Kunstig Intelligens modsvarer størstedelen af Dreyfus' kritik, hvilket betyder, at hvis vi kan fremstille en praktisk computer baseret på dette paradigme, kan vi måske tage de første egentlige skridt mod en egentlig tænkende kunstig intelligens. Vores formodning i indledningen, vedrørende Searles Chinese Room tanke-eksperiments validitet viser sig at være korrekte, da Dennett på overbevisende manér maner eksperimentet i jorden. Vi kan således afvise at eksperimentet udelukker muligheden for Stærk Kunstig Intelligens, dog betyder dette ikke automatisk at den Stærke Kunstige Intelligens er mulig. Dennetts "Adskillige Udkast" teori om bevidstheden, hvis den er korrekt, peger dog i retning af, at hvis vi kan lave de individuelle funktionelle dele i hjernen, vil bevidstheden opstå på samme måde som den eksisterer hos os mennesker, da vi i så tilfælde vil have alle de forudsætninger i den Kunstige Intelligens, som vi selv har. Ved et længere heterofænomenologisk studie kan man finde ud af de enkelte dele af den fysiske hjerne, og derved begynde at simulere – og derved skabe – bevidsthed, og en Stærk Kunstig Intelligens.

[14] En filosofisk zombie er et væsen, der på alle måde opfører sig som et menneske, men netop *ikke* har en bevidsthed, og derfor ikke har kvalitative referencer til ting. (Om disse kan eksistere er et helt andet argument.)

Litteraturliste

Webpubliceringer blev indhentet den 22. januar 2008 og var tilgængelige på dette tidspunkt.

Brunak, Søren og Lautrup, Benny. 1990. *Neurale Netværk*. Munksgaard.

Chalmers, David J. 1992. *Subsymbolic Computation and the Chinese Room*. Fra "The Symbolic and Connectionist Paradigms".

Chalmers, David. J. 1995. *Facing Up to the Problem of Consciousness*. Fra "Journal of Consciousness Studies" 2, s. 200-219, også webpublicering (http://www.imprint.co.uk/chalmers.html)

Churchland, Paul & Churchland, Patricia. 1990. *Could a machine think?.* Fra "Scientific American" 262

Dennett, Daniel C. 1978. *Brainstorms.* Montgomery, VT: Bradford Books.

Dennett, Daniel C. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press/A Bradford Book.

Dennett, Daniel C. 1991. *Consciousness Explained*. USA: Little, Brown and Co.

Dennett, Daniel C. 1993. *Quining Qualia*. MIT Press, webpublicering. (http://ase.tufts.edu/cogstud/papers/quinqual.htm)

Dennett, Daniel C. 1995. *Facing Backwards on the Problem of Consciousness*. Webpublicering. (http://ase.tufts.edu/cogstud/papers/chalmers.htm)

Dennett, Daniel C. 2006. *Heterophenomenology reconsidered*. Udkast, webpublicering. (http://ase.tufts.edu/cogstud/papers/heteroreconsidered.pdf)

Dennett, Daniel C. 2007. *The Magic of Consciousness.* Forelæsning, webpublicering. (http://video.google.com/videoplay?docid=-8084768678469239623)

Dreyfus, Hubert L. 1972. *What computers can't do*. New York: Harper & Row, Publishers Inc.

Dreyfus, Hubert L. og Dreyfus, Stuart E. 1986. *Mind over Machine*. Oxford: Basil Blackwell Ltd.

Dreyfus, Hubert L. 1992. *What computers still can't do*. Cambridge, MA: MIT Press

Kurzweil, Raymond. 2005. *The Singularity is Near.* New York: Viking Press.

McCarthy, John. 2007. *What is artificial intelligence*. Webpublicering. (http://www-formal.stanford.edu/jmc/whatisai/whatisai.html)

Searle, John R. 1980. *Minds, Brains, and Programs*. USA: Cambridge University Press. (http://www.bbsonline.org/documents/a/00/00/04/84/bbs00000484-00/bbs.searle2.html)

Searle, John R. 1984. *Minds, Brains, and Science*. Cambridge: Harvard University Press.

Searle, John R. 2004. *Mind: a brief introduction*. New York: Oxford University Press, Inc.

Turing, Alan M. 1936. *On Computable Numbers, with an Application to the Entscheidungsproblem*. Webpublicering. (http://www.abelard.org/turpap2/tp2-ie.asp eller http://www.thocp.net/biographies/papers/turing_oncomputablenumbers_1936.pdf)

Turing, Alan M. 1950. *Computing Machinery and Intelligence.* Webpublicering. (http://www.abelard.org/turpap/turpap.htm)

Wackerhausen, Steen. 1989. *Mennesket i Computerens Billede*. Fra ”Philosophia” årg. 18 1-2.

Bilag 1 – Dennetts eksempel vedr. Chinese Room

Eksemplet er nedenstående citeret verbatim.

JUDGE: Did you hear about the Irishman who found a magic lamp? When he rubbed it a genie appeared and granted him three wishes. "I'll have a pint of Guinness!" the Irishman replied, and immediately it appeared. The Irishman eagerly set to sipping and then gulping, but the level of Guinness in the glass was always magically restored. After a while the genie became impatient. "Well, what about your second wish?" he asked. Replied the Irishman between gulps, "Oh well, I guess I'll have another one of these!"

CHINESE ROOM: Very funny. No, I hadn't heard it – but you know, I find ethnic jokes in bad taste. I laughed in spite of myself, but really, I think you should find other topics for us to discuss.

J: Fair enough, but I told you the joke because I want you to explain it to me.

CR: Boring! You should never explain jokes.

J: Nevertheless, this is my question. Can you explain to me how and why the joke "works"?

CR: if you insist. You see, it depends on the assumption that the magically refilling glass will go on refilling forever, so the Irishman has all the stout he can ever drink. So he hardly has a reason for wanting a duplicate, but he is so stupid (that's the part I objected to) or so besotted by the alcohol, that he doesn't recognize this, and so, unthinkingly endorsing his delight with his first wish come true, he asks for seconds. These background assumptions aren't true, of course, but just part of the ambient lore of joke-telling, in which we suspend our disbelief in magic, and so forth By the way, we could imagine a somewhat labored continuation in which the Irishman turned out to be "right" in his second wish after all – perhaps he's planning to throw a big party, and one glass won't refill fast enough to satisfy all his thirsty guests (and it's no use saving up in advance – we all know how stale stout loses its taste). We tend not to think of such complications, which is part of the explanation of why jokes work. Is that enough?