EN KINESISK ROBOT

En C-uppsats om artificiell intelligens, med utgångspunkt i Det kinesiska rummet – ett argument av John R. Searle

Lunds Universitet
Teoretisk Filosofi
C-uppsats
Jens Engfors
Handledare: Lennart Karlsson

Innehållsförteckning

1. Inledning

Idag är det inte helt ovanligt att tala om datorer som att de förstår eller på ett eller annat sätt lever sitt eget liv. Idén om att datorer kanske kan tänka och vara medvetna har sina rötter i det arbete som Alan Turing utförde under 1900-talets första hälft. Han brukar anses ha lagt grunden till det som idag kallas *artificiell intelligens* (AI) (Preston, 2002). Man började fundera över om datorer kunde ha egenskaper som kan likställas med människors mentala tillstånd och därför användas för att förklara dessa. Ett sådant påstående förs fram i en teori kallad ”*datorfunktionalism*”. Tanken var - och är - dock inte okontroversiell och John R. Searle formulerade, 1980, ”*Det kinesiska rummet*”, som en reaktion på den. Det kinesiska rummet (stor bokstav betyder att jag syftar på Searles tankeexperiment och argumentet som det utgör, skilt från då jag syftar på rummet i tankeexperimentet och använder liten bokstav) är riktat mot påståendet att en dator kan förstå. Searle vill visa att datorer inte kan förstå och därmed inte verkar kunna vara medvetna. Han påstår också att man inte kan testa en artificiell intelligens’ förmåga på ett tillförlitligt sätt och att den därför inte kan jämföras med en människas egenskaper.

Searles argument har blivit väldigt uppmärksammat och diskussionen av det lever fortfarande kvar. Det har till och med hävdats att kognitionsforskning kan definieras som ett försök att vederlägga det (Cole, 2004). I den här uppsatsen så vill jag försöka visa hur den här diskussionen kan användas för att utveckla de påståenden som Searle försöker förkasta med sitt argument. Detta är inte helt självklart, eftersom man både kan anse att Det kinesiska rummet gör AI-projektets anspråk, på att kunna skapa artificiell medvetenhet, trivialt och att argumentet i sig är likaledes trivialt felaktigt. Mitt syfte är att visa att man kan finna argument för att en artificiell intelligens både kan vara medveten och förklara det mänskliga medvetandet i tre reaktioner på Det kinesiska rummet, som formulerats fristående av varandra. Daniel C. Dennett påstår i sin bok *Consiousness Explained* (Dennett, 1992) att Searle inte lyckas säga något om frågan *om* en artificiell intelligens kan vara medveten. Han hävdar dessutom att den fråga man bör fokusera på är *hur* en artificiell intelligens kan vara medveten. Jag tar, under 4.1 fasta på Dennetts resonemang och förändringen av frågeställningen. Det som då tycks återstå av Det kinesiska rummet är en aspekt av Dennetts nya frågeställning, nämligen hur en artificiell intelligens kan ta hänsyn till och hantera den semantiska aspekten av input; alltså dess meningsinnehåll. Detta problem kallar jag för ”*semantikproblemet*”. Semantikproblemet blir den koppling som jag ser mellan Dennetts

invändning och ”Minds, Machines and Searle” av Stevan Harnad (Harnad 1989), som jag behandlar under 4.2. Där påstår han att en robot har de förutsättningar som krävs för att lösa det som jag kallar för semantikproblemet, till skillnad från en rent symbolmanipulerande dator. Jag menar att man då genom att uppmärksamma semantikproblemet i Searles argumentation har utvecklat tesen om artificiella intelligensers medvetenhet, Searles ståndpunkt till trots. För att visa att AI därför också kan hjälpa oss att förklara mänsklig medvetenhet så tar jag, under 4.3, upp ”A Chinese Room that Understands” av Herbert A. Simon och Stuart A. Eisenstadt (Simon och Eisenstadt, 2002), där författarna diskuterar vad det innebär att testa om någon, människa eller maskin, förstår. Jag använder deras reaktion på Det kinesiska rummet för att visa att det finns förutsättningar för att jämföra en människas mentala förmågor med en artificiell intelligens’ egenskaper.

Utöver att dessa tre ståndpunkter är reaktioner på Det kinesiska rummet så finns det inget allmänt vedertaget samband mellan dem och de anses inte representera utvecklingen av diskussionen kring medvetenhet och AI, som en följd av Searles argument. Men jag menar att de kan användas för att försvara att en artificiell intelligens kan vara medveten och förklara mänskliga kognitiva tillstånd, utifrån en funktionalistisk teori om medvetandet. I uppsatsen vill jag alltså visa att man kan finna argument för att en robot kan förstå, bland reaktionerna på Det kinesiska rummet, därav dess titel. Den slutsats som jag drar av dem är alltså inte på något sätt allmänt vedertagen som resultatet av diskussionen kring Det kinesiska rummet. Men den är enligt mig ett konstruktivt sätt att tolka debatten och kan dessutom ses som ett intressant exempel på filosofiska resonemangs betydelse för en ung vetenskap, i det här fallet den som behandlar AI.

Innan jag går in på reaktionerna på Searles argument så försöker jag återge bakgrunden till de påståenden som det är riktat mot. Detta gör jag kortfattat utifrån hur Preston (Preston, 2002) och Searle (Searle, 1980, 1992 och 1994) beskrivit den. Därefter beskriver jag själva argumentet, hur Searle lägger fram det, vilka slutsatser han drar av det och något av hans motiv till dessa. Detta gör jag huvudsakligen utifrån ”Minds, Brains and Programs” (Searle 1980), där Searle presenterade Det kinesiska rummet i skriftlig form för första gången. De ståndpunkter som Searle lägger fram i ”Minds, Brains and Programs” kommer att vara de som jag behandlar också därefter, om inget annat framgår av sammanhanget eller referenser.

2. Bakgrunden till Det kinesiska rummet

För att förstå de slutsatser som Searle drar av Det kinesiska rummet så ska jag nu kortfattat redogöra för vad han drar dem om. Detta innefattar både den beskrivning som Searle ger av AI, för sina syften, och en mer fristående sådan**.** Jag försöker visa hur man kan resonera för att anta den syn på AI som Searle sedan förkastar.

2.1. Stark AI och svag AI

Searles kritik mot AI riktar sig inte mot allt vad detta kan innebära. Han väljer att göra en distinktion mellan vad han kallar ”*stark AI*” och ”*svag AI*”. Svag AI är ett namn på uppfattningen att datorer kan och bör användas som verktyg i forskning om det mänskliga medvetandet (när jag använder termen ”medvetande” så låter jag den vara liktydig med den engelska termen ”mind” och då ”consciousness” åsyftas så kommer jag att använda ”medvetenhet”). Som ett sådant verktyg kan en dator t.ex. användas för att ställa upp och testa hypoteser om kognitiva tillstånd genom att simulera dem. Searle har inget emot den synen på AI men han menar då inte att simulationer innebär duplikationer. I Det kinesiska rummet finns inget som är ämnat att användas för att kritisera svag AI. Argumentet och Searles kritik är endast riktat mot stark AI. Begreppet står för uppfattningen att en dator med rätt program har kognitiva tillstånd och ett medvetande, på precis samma sätt som en människa. Därmed så hjälper datorns program oss inte bara med att förklara vad som händer i en mänsklig hjärna utan det utgör också själva förklaringen. Dessa två påståenden är det som Searle explicit anger som mål för den kritik av AI som han vill framföra med Det kinesiska rummet (Searle, 1980).

Stark AI är enligt Searle resultatet av att AI-projektet stötte ihop med funktionalismen. Han kallar också tesen för ”datorfunktionalism” (Searle, 1992), vilket är liktydigt med de mer vanliga begreppen ”the computer model of mind” och ”computationalism” (Preston, 2002). Funktionalism är ett samlingsnamn för teorier om vad ett medvetande är och kommer att beskrivas nedan. Artificiell intelligens behöver inte göra anspråk på att förklara det mänskliga medvetandet, men stark AI innefattar alltså den ambitionen via funktionalismen. Distinktionen mellan stark- och svag AI är en förenkling som är Searles egen, vilket gjort att många ansett att de inte kunnat sälla sig till den ena eller andra uppfattningen. Stark AI kan anses ha sin grund i Alan Turings beskrivning av maskiner som intelligenta.

2.2. Turingmaskiner ochturingtestet

John Preston beskriver, i sin inledning till *Views into the Chinese Room* (Preston, 2002) hur Alan Turing anses ha lagt grunden till utvecklingen av AI i sin uppsats ”In Computatable Numbers, with an Application to the Entscheidungsproblem”, redan 1936. Turing lade fram en tes om hur det går till varje gång man beräknar något med en mekanisk metod, d.v.s. en algoritm. Han menade att sådana beräkningar alltid kan utföras genom att skriva symboler, från ett ändligt alfabet, på en pappersremsa. Remsan ska vara indelad i rutor med plats för en symbol i varje ruta. Hur den som utför beräkningen ska bete sig vid en viss tidpunkt bestäms av symbolerna som observeras för tillfället och hans eller hennes aktuella medvetandetillstånd (”state of mind”). Antalet symboler som kan observeras är begränsat och när man önskar få tillgång till fler så måste man göra en ny observation. Antalet aktuella och relevanta medvetandetillstånd är också begränsat. Beräkningen kan beskrivas som enkla operationer efter varandra, som var och en består i en förändring på pappersremsan och i den som utför beräkningen (jag förmodar att det som sker i beräknaren är förändringar av medvetandetillstånd och representationer). Det enda som har betydelse för processen och som en utomstående inte kan observera är medvetandetillståndet hos personen som arbetar. Detta löste Turing genom att hela tiden låta denne skriva ner hur nästa steg ska gå till. Den här beskrivningen av hur en beräkning utförs innebär alltså inget mer kontroversiellt än en beskrivning av en människa som räknar.

Turing fortsatte sedan med att hävda att man kan bygga en maskin som utför beräkningar på precis det här sättet; en *turingmaskin*. Allt som följer metoden som beskrivs ovan kan kallas för en turingmaskin, men detta innebär också att vissa delar (eller vissa motsvarande funktioner) måste finnas: en obestämbart lång pappersremsa som är indelad i lika stora rutor, vilka rymmer en symbol i taget, och en del som kan skriva, läsa och radera en symbol i taget. De symboler som turingmaskinen ska kunna behandla ska komma från ett ändligt alfabet. Turingmaskinen läser en symbol i taget och har ett visst antal villkor som den kan behandla denna symbol utefter. De villkor som finns, pappersremsans hela innehåll och den symbol som maskinen läser för tillfället bestämmer hur den kan bete sig. En sammanställning av vilka villkor som turingmaskinen kan behandla symboler utefter och vilka möjliga beteenden som detta kan ge kallar Preston, omsatt i dagens termer, för dess program.

Alonzo Church hade formulerat en tes som sa ungefär samma sak som Turings, ungefär samtidigt, och de skapade tillsammans ”Church-Turing-tesen”. Den säger att alla

funktioner som man kan utföra beräkningar över med en mekanisk metod kan beräknas av en turingmaskin. Tesen har dessutom tolkats som att alla funktioner kan beräknas med en mekanisk metod och av en turingmaskin (Preston påpekar att detta inte anses vara bevisat). Turing skapade senare det som kallas ”Turings teorem” och som av många, inklusive av Searle (Searle, 1992), anses vara det verkliga startskottet för AI. Det bevisar att det är möjligt att skapa en *universell turingmaskin* som kan utföra alla operationer som alla andra turingmaskiner kan utföra, d.v.s. emulera alla andra turingmaskiner. Den universella turingmaskinen är inget som man kan hoppas på att kunna bygga men den ses ändå som en modell till dagens datorer.

Turing visade alltså hur en människas sätt att använda en mekanisk metod för att beräkna något kan utföras på samma sätt av en turingmaskin. Church-Turing-tesen innebär att alla algoritmer kan följas av, implementeras av, en turingmaskin. Dessutom kan Church-Turing-tesen tillsammans med Turings teorem tolkas som att en universell turingmaskin därför kan hantera alla beräkningar. Detta inkluderar alla beräkningar som beskriver världen och betyder att alla beräkningar som den mänskliga hjärnan kan hantera också kan hanteras av en maskin. Resultaten har inspirerat till tanken att maskiner kan uppnå intelligens som kan jämföras med en människas; *artificiell intelligens.* Preston skriver att Turing själv insisterade på att hans resultat innebar att maskiner, t.ex. sådana som idag kallas datorer, kan tänka och ansåg att frågan som bör ställas är om de också kan vara intelligenta. För att avgöra en dators intelligens skapade Turing (1950) ”the imitation game”, som idag kallas för *turingtestet*. Testet går till så att en människa skriftligen ställer frågor till en dator (då en turingmaskin), utan att veta om det är en människa eller en dator. Datorn arbetar för att skicka tillbaka samma sorts svar som en människa skulle ha gjort. Om frågeställaren inte kan avslöja datorn när han får svaret så anser Turing att man måste tillskriva datorn intelligens. Detta ligger till grund för senare teoriers ambition att med ett likadant test avgöra om en dator också har funktioner som utgör ett medvetande.

2.3. Funktionalism

Funktionalism sammanfattas ofta med att ”medvetandet förhåller sig till hjärnan så som ett program förhåller sig till hårdvaran” (t.ex. Searle, 1992). Preston beskriver tesens drivkraft som att det mentala handlar om funktion och inte substans. Under funktionalism finns en mängd olika teorier men han menar att de flesta innehåller en metafysisk tes som säger att mentala fenomen individueras genom sina kausala roller. Detta ger alltså att de inte är bundna

till en substans (Preston, 2002). En dators funktioner skulle då kunna vara de samma som de som ger upphov till medvetande i en mänsklig hjärna, vilket är en grundsten i stark AI. Preston beskriver tidig funktionalism (ca 1960) som att den likställde ett medvetandes tillstånd med tillståndet i ett program. De mentala fenomenen skulle kunna bestämmas på samma sätt som när man studerar tillståndet i en dator via funktionerna hos ett datorprograms delar. Han menar att den ståndpunkten utvecklades till att man också måste ta hänsyn till funktionernas plats i den kausalkedja som också består av det studerade systemets input och output. Det räcker alltså inte att bara studera programmet i sig. Detta ser han som ett, idag, dominerande inslag i medvetandefilosofi (philosophy of mind) och en del i *datorfunktionalismen*, som Searle kritiserar med Det kinesiska rummet.

Datorfunktionalism beskriver medvetandet på ett sätt som är kopplat till Turings teori för hur beräkningar utförs. Preston återger Turings syn på medvetandet som ett system som behandlar mentala representationer av världen. Dessa representationer är symboler med både en *semantisk* och en *syntaktisk* aspekt, d.v.s. ett meningsinnehåll och form. Intelligens i ett system består då i att det kan hantera sådana symboler och samtidigt bevara den semantiska aspekten av dem, d.v.s. meningsinnehållet. Turing ansåg sig, enligt Preston, ha visat hur detta kan göras mekaniskt med sin turingmaskin. Turingmaskinen manipulerar de symboler som ges som input utefter deras syntaktiska aspekt och kan härleda en output. Han hävdade också att inputens semantiska aspekt bibehålls när en sådan manipulation sker. Att sanna premisser och en riktig härledning ger en sann slutsats är förmodligen inget kontroversiellt för den som är bekant med logik. Datorfunktionalismen innebär dock också att det arbete som en turingmaskin utför är samma arbete som en mänsklig hjärna utför, vilket inte är helt självklart. Searle karakteriserar tankegången som att eftersom åtminstone vissa av människans mentala processer kan uttryckas med en algoritm (en metod för att mekaniskt beräkna något), t.ex. då vi räknar, så finns det skäl att tro att också omedvetna processer kan beskrivas så (Searle, 1992). Samma algoritmer som beskriver funktionerna i den mänskliga hjärnan kan beskriva funktionerna i en dator, eller en turingmaskin, och tvärt om. För att testa detta så innebär det, främst av praktiska skäl, svårigheter att jämföra direkta observationer av en hjärna med en dators program. Man vänder sig då istället till funktionernas plats i en kausalkedja och observerar input och output. Det är då ett turingtest som används för att undersöka om datorfunktionalismen stämmer. Om en dator och en människa beter sig likvärdigt i en situation som kräver mentala processer så menar datorfunktionalisten att de ska anses besitta samma mentala egenskap.

Turings arbete har alltså varit grundläggande för hur en maskin, vanligtvis en dator, kan anses vara intelligent. Funktionalismen innebär att människans mentala fenomen inte är bundna till en mänsklig hjärna. Och datorfunktionalismen påstår att de beräkningar som en dator utför är samma arbete som en medveten hjärna utför. För att testa om t.ex. en dator har artificiell intelligens så ska man testa om den kan ge output som är korrekt härledd från input. Datorfunktionalismen innebär att processerna i en hjärna är av samma slag som beräkningarna som en dator utför. Ett turingtest används för att undersöka om funktionerna i en given dator är samma som de i ett mänskligt medvetande, genom att jämföra input-output-relationer. Det Searle inte vill gå med på är att den sortens funktioner som kan finnas i en dator är tillräckligt för att ge upphov till medvetenhet, oavsett resultatet av ett turingtest. Om datorfunktionalismen är sann så skulle ett system som manipulerar input utan hänsyn till dess semantiska aspekt också kunna förklara hur vi kan påstå att vi förstår världen omkring oss. Både turingmaskiner och dagens datorer manipulerar, enligt Searle, symboler på det sättet (Searle, 1980). I enlighet med stark AI så betyder detta att de kan vara medvetna på precis samma sätt som en människa, eftersom de verkar ha funktioner som kan ge output som inte kan skiljas från en människas. Detta vill Searle tillbakavisa med Det kinesiska rummet.

3. Searle i och om Det kinesiska rummet

Jag ska nu återge Det kinesiska rummet så som det ser ut i ”Minds, Brains and Programs” (Searle, 1980), där det för första gången togs upp i skriftlig form. Argumentet är ett tankeexperiment som är format för att likna och testa ett program som ska hantera språkförståelse. Det program som Searle efterliknar är baserat på arbete som Roger Schank och Robert Abelson utfört. Han går inte närmare in på vad detta består i men då Preston nämner det så framgår att det som Searle intresserade sig för handlar om hur en dator kan svar på frågor om en historia, då all nödvändig information inte ges explicit. Schanks och Abelsons resultat kan alltså användas i ett försök att programmera en dator för att fungera som ett mänskligt medvetande som förstår. Searle använder deras arbete för att skapa strukturen i det program som han vill simulera i Det kinesiska rummet (Preston, 2002).

Programmets uppgift är att simulera språkförståelse genom att kunna svara på frågor om en historia. Frågorna är utformade så att de inte handlar om information som faktiskt finns i den berättelse som tillhandahålls utan bygger på att tolka den, på samma sätt som en människa skulle ha gjort. T.ex. förutsätter vi att någon som grimaserar efter en spark

på smalbenet har upplevt smärta, utan att det behöver sägas explicit. För att kunna göra detta så ges datorn information om världen, som samlad i datorn benämns ett ”skript”. Programmet kan på det här sättet simulera mänsklig förmåga att förstå och testas i ett turingtest. Om datorns simulation är framgångsrik i testet så säger stark AI-tesen, tillsammans med datorfunktionalismen, både att den faktiskt förstår och att dess program dessutom förklarar hur en människa förstår. Med Det kinesiska rummet vill Searle visa att dessa två påståenden inte har stöd i ett program likt ett baserat på Schanks och Abelsons arbete eller på datorfunktionalistiska principer överhuvudtaget.

3.1. Det kinesiska rummet

I sitt tankeexperiment sätter sig Searle i ett rum tillsammans med en samling kinesiska symboler. Eftersom engelska är det enda språk han förstår så har de kinesiska tecknen ingen mening för honom. Sedan skickas det in en andra samling kinesiska tecken och regler, skrivna på engelska, för hur han kan sätta samman de kinesiska tecken han redan hade med de nya. En tredje samling kinesiska tecken skickas in i rummet, tillsammans med ytterligare regler. Dessa beskriver hur han ska sammanfoga de kinesiska tecken han har med de nya och skicka tillbaka vissa sammanställningar, ut ur rummet. De kinesiska tecknen är formella för Searle, i den meningen att han särskiljer dem utefter symbolernas form. De som befinner sig utanför rummet kallar den första samlingen kinesiska tecken för ”ett skript”, den andra för ”en berättelse” och den tredje för ”frågor”. Reglerna som Searle har att arbeta efter kallar de för ”programmet” och de kinesiska tecknen han skickar tillbaka benämns ”svar på frågorna”. Citationstecknen kring teckensamlingarnas benämningar återfinns i Searles egen text och understryker att han själv, sittandes i rummet, inte känner till dem eller att de överhuvudtaget kan tillskrivas de olika teckensamlingar han arbetar med. Samtidigt som han utför sitt arbete med de kinesiska symbolerna så får han också en berättelse och frågor på den, skrivna på engelska, och kan då svara på dem, på engelska. Trots att han inte förstår ett dugg av de kinesiska tecknen så är de regler som han manipulerar dem utefter, d.v.s. programmet, så välgjorda att hans svar på kinesiska är lika adekvata för personerna utanför rummet, som förstår kinesiska, som hans svar på engelska är för honom och andra som förstår engelska. Searle menar att i fallet med kinesiskan så har han implementerat ett program för att förstå kinesiska, precis som en dator skulle ha gjort, och det utan att förstå någonting.

3.2. Searles slutsatser

Genom att fungera som en dator som implementerar Schanks och Abelsons program så simulerar Searle, enligt honom själv, hur en människa förstår (naturligt språk). I tankeexperimentet är simulationen dessutom framgångsrik, på det sättet att han lyckas klara av ett turingtest där den kinesisktalande publiken godtar hans svar. Searle menar att enligt stark AI skulle detta betyda att hans program för att förstå faktiskt förstår. Dessutom så skulle det innebära att rummet därför förklarar ett mänskligt mentalt tillstånd. Han hävdar att han har samma input och output som någon som förstår kinesiska skulle ha haft i samma situation. Han verkar alltså mena att han simulerar en artificiell intelligens som, med en datorfunktionalistisk ståndpunkt, ska anses förstå på samma sätt som en människa förstår. Trots detta så hävdar Searle att tankeexperimentet visar att när han sitter i det kinesiska rummet så förstår han faktiskt ingenting. Eftersom han också menar att analogin mellan sitt och en dators arbete är giltig så blir slutsatsen att en datorsimulation av att förstå inte måste innebära att datorn faktiskt förstår. Han ser heller ingen anledning till att anta att en dator skulle förstå om den utför sitt arbete annorlunda. Searles sammanfattar sin ståndpunkt i Det kinesiska rummet med ”… vilka rent formella principer man än förser en dator med så kommer de inte att vara tillräckliga för att den ska förstå, eftersom en människa kan följa de formella principerna utan att förstå någonting”.

Searle ser alltså inte att programmet i Det kinesiska rummet varken kan användas för att påvisa någon tillräcklig eller någon nödvändig komponent i vad det innebär att förstå. Han menar att han har simulerat och testat ett medvetande enligt datorfunktionalismens mönster utan att få resultatet som tesen förutsäger. Med detta vill han inte säga att han har vederlagt den möjliga datorfunktionalistiska förklaringen att det bara krävs mer symbolmanipulation än vad han utförde för att förstå kinesiska i det kinesiska rummet. Men att bevisbördan ligger, i hans ögon, på datorfunktionalismen. Det kinesiska rummet ger också att turingtestet inte kan fungera som ett kriterium för mentalt innehåll. Detta eftersom det kinesiska rummet kunde genomgå ett turingtest med lyckat resultat utan att han upplevde det kognitiva tillstånd som testades. Datorfunktionalismens tes att mänskliga mentala funktioner är formella skulle beläggas med turingtestet men när Searle visat att det har klara brister så försvinner den möjligheten. Han anser då att inget verkar tala för teorin. Turingtestets brister kan anses beröra alla former av funktionalism, eftersom de, som Searle påstår, bara kan individuera mänskliga mentala fenomen genom att studera input-output-relationer. Då det har visat sig att en sådan metod inte stämmer för datorfunktionalismen så

skulle den lika gärna kunna ge felaktiga resultat tillsammans med andra funktionalistiska teorier.

Sitt arguments premisser ställer Searle själv upp som följande punkter (Searle, 1994): (1). Program är formella. (2). Ett medvetande har semantiskt innehåll. (3). Syntax är inte tillräckligt för att skapa semantiskt innehåll. Av detta följer att ett program, ett datorprogram eller andra samlingar av formella funktioner, inte kan vara förklaringen till medvetenhet. Han motsätter sig dock inte att vissa av våra mentala funktioner sker i enlighet med Turings metod för beräkning. Det jag uppfattar som att Det kinesiska rummet talar mot är datorfunktionalismens steg från att vissa av våra mentala processer är formella till att alla mentala fenomen är det. D.v.s. att de bara består i att manipulera symboler utefter deras syntaktiska aspekt. Att ett program är formellt, vilket ett datorprogram är, betyder att det består av godtyckliga regler för hur symboler ska manipuleras utefter deras form, som är deras syntaktiska egenskaper. Syntax är benämningen på sådana formella regler och Searle pekar på att en syntaktisk definition av symboler inte är tillräckligt för att förstå vad de betyder, d.v.s. deras semantiska innehåll, eller mening. Att Searle manipulerade de kinesiska tecknen utefter formella regler, som en dator skulle ha gjort, är alltså är hans förklaring till varför han inte förstår något i Det kinesiska rummet. Programmet kan implementeras utan hänsyn till om det överhuvudtaget finns ett semantiskt innehåll att förstå. Därför så kan andra kognitiva tillstånd, som också kräver medvetenhet om världen, inte heller infinna sig i en dator, enligt Searle. Vad som krävs för att något ska förstå meningsinnehåll, d.v.s. vara medveten om världen, är Searle inte tydlig med, men hans intuition är uppenbarligen att en dator inte kan ha en sådan förmåga. Han menar att hans positiva tes om medvetenhet inte har någon betydelse för Det kinesiska rummet. Trots Searles uppfattning så får den dock betydelse i diskussionen av reaktionerna på argumentet.

3.3. Biologisk naturalism

Att Searle inte tror att artificiell intelligens kan återskapa en mänsklig intelligens, d.v.s. att datorfunktionalism inte är en trolig tes, beror förmodligen inte bara på hans tankeexperiment. I samband med hans åsikt om att tillskriva datorer medvetenhet så ger han, i ”Minds, Brains and Programs”, uttryck för sin egen teori för villkoren för medvetenhet. I korthet så innebär den att man godtar att människor har mentalt innehåll och därför kan dra slutsatsen att allt som är någorlunda likt människan, främst med avseende på vad det består av, också har det. Preston kallar Searles teori för ”*biologisk naturalism*”, när han beskriver den (Preston, 2002).

Tesen säger, enligt Prestons sammanfattning, att hjärnans kapacitet att tillgodogöra sig och behandla representationer av världen, hjärnans *intentionalitet*, är en egenskap som kommer ur dess biologiska substans. Intentionalitet behövs för medvetenhet och att förstå symbolers relation till världen. Förhållandet mellan mentala egenskaper och hjärnans biologiska substans förklarar Preston med att ett tings delar, på mikronivå, har egenskaper som står i kausala relationer till tingets egenskaper, på makronivå. Mentala fenomen är alltså egenskaper på makronivå hos hjärnan, vilka orsakas av de egenskaper på mikronivå som är specifika för hjärnans biologiska substans. Eftersom hjärnan är ett fysiskt ting så kan den, med sin intentionala förmåga och medvetenhet, därför både orsaka handlingar i den fysiska världen och förstå den. Biologisk naturalism innebär alltså att mentalt innehåll och intentionalitet, som kan jämföras med en människas, aldrig kan uppstå i ett syntetiskt system.

3.4. Intentionalitet

Intentionalitet är alltså ett grundläggande begrepp för Searle. Det verkar vara uppenbart att förmågan att hålla sig med representationer av världen är nödvändigt för att förstå. Searle gör en distinktion mellan *inneboende* (intrinsic) och *härledd* (derived) intentionalitet. En inneboende intentionalitet innebär den sortens verkliga intentionalitet som en människa upplever. Härledd intentionalitet är den sortens egenskaper som vi metaforiskt tillskriver ting när vi t.ex. talar om att en termometer vet vilken temperatur det är. På så sätt förlänger vi vår egen intentionalitet genom tingets verkliga egenskaper. Man kan alltså använda yttre ting som en sorts förlängning av sinnesorganen och få kunskap om, i form av representationer av, verkligheten, som annars inte skulle vara möjlig. Searle anser att ett program implementerat i en dator bara kan innebära en härledd intentionalitet, eftersom ett program för att simulera ett skeende i världen bara tar hänsyn till syntaxen i dess input. Detta betyder att datorn inte har egna representationer av världen men att vi kan använda den för att förlänga vår egen intentionalitet. Han pekar på att informationsprocessandet i en dator kan likna mentala processer, mer än andra ickementala skeenden i världen. Men menar samtidigt att det inte finns anledning att tro att en simulation av den typen är mer verklig än t.ex. en simulation av en naturkatastrof. Om en dators arbete inte lyckas återskapa intentionaliteten så menar Searle att den inte kan anses duplicera utan endast imitera ett mänskligt medvetande. En datorsimulation av en mänsklig mental egenskap måste då ses på samma sätt som en datorsimulation av något ickementalt, d.v.s. inte som det faktiska skeendet. Man kan se det

som att Searle stöder den här uppfattningen med Det kinesiska rummet, eftersom det är just intentionaliteten som han går miste om när han arbetar som en dator.

Den potentiella skillnaden mellan en simulation av mentala processer och ickementala processer är datorfunktionalismens förmodan att mentala processer, precis som en datorsimulation, innebär manipulation av formella symboler. Den intuitionen anser Searle helt enkelt vara fel och vill med Det kinesiska rummet visa att den är helt ogrundad. Han ersätter dock inte datorfunktionalismen med belägg för biologisk naturalism. Han förklarar alltså inte varför en biologisk hjärna är medveten. Man får inte heller någon förklaring till vad det innebär att förstå och hur man skulle kunna testa detta. Han nöjer sig med att tycka sig kunna peka ut solklara fall där något förstår eller inte förstår och menar att hans argument inte behöver ett klarare begrepp. Att tala om ickebiologiska ting som om de förstår och har inneboende intentionalitet, istället för härledd intentionalitet, skyller Searle på att funktionalism har ett arv från behaviorismen. Den säger att mentalt innehåll ska individueras genom beteende (Preston, 2002), såsom i turingtestet. Detta anser han vara en felaktig uppfattning, som man kan undvika efter att han har visat turingtestets oförmåga att skilja mellan det kinesiska rummets produktion av svar på engelska och kinesiska. Samma typ av input ger då upphov till likvärdig output, men produceras på helt olika sätt. Behaviorismen ses allmänt som att vara utan förklaringskraft och samma tillstånd återfinner Searle i funktionalismen (Searle, 1992). Funktionalismens, särskilt datorfunktionalismens, fokus på beteende menar Searle alltså är ett misslyckat försök till att förklara vad som krävs för att förstå eller att ha inneboende intentionalitet och medvetenhet. Searle tog redan i ”Minds, Brains and programs” upp och bemötte svar på Det kinesiska rummet och de slutsatser som han ansåg sig kunna dra av det. Det har han dock fått fortsätta med sedan dess (Preston, 2002).

4. Svar på Det kinesiska rummet

I ”Minds, Brains and Programs” delar Searle upp de som han uppfattar som de mest direkta invändningarna mot hans argument i kategorierna ”*systemsvar*”, ”*robotsvar*”, ”kombinationssvar” och ”hjärnsimulatorsvar”. Han anser sig kunna avfärda var och en av dem och därmed att det inte finns några allvarliga invändningar mot Det kinesiska rummet. Ett hjärnsimulatorsvar utgörs av ståndpunkten att om man kopierar en hjärnas samtliga funktioner så måste det nya systemet ha egenskaper likvärdiga förlagans. Ett kombinationssvar går ut på att principen bakom flera av de andra kategorierna måste tillgodoses för att rädda stark AI.

Under 4.1 och 4.2 så kommer jag att förklara innebörden av ett systemsvar respektive ett robotsvar. Eftersom jag använder ett systemsvar och ett robotsvar som två delar i ett försvar av artificiella intelligensers potentiella medvetenhet så kan detta anses vara ett kombinationssvar. I den diskussion av stark AI som förs nedan så kommer jag dock inte att gå närmare in på varken ett hjärnsimulatorsvar eller eller kombinationssvar. Men jag har nämner dessa typer av reaktioner för att visa att diskussionen av Det kinesiska rummet redan från början var bredare och omfattande fler argument än de som jag vill visa är av betydelse för stark AI.

I den här delen av uppsatsen så använder jag tre reaktioner på Det kinesiska rummet för att visa hur man kan, och enligt mig, bör utveckla stark AI-begreppet. Detta visar att oavsett hur man ställer sig till argumentets giltighet så kan det anses ha betydelse för AI som vetenskap, genom att lyfta fram en aspekt av *hur* istället för *om* en artificiell intelligens kan vara medveten. De frågor som blir viktigast är hur funktionalismen kan förklara att ett system kan ta hänsyn till den semantiska aspekten av input (vilket jag kommer att referera till som ”semantikproblemet”) och hur turingtestets brister ska hanteras. Med andra ord så försöker jag visa på ett försvar av att en artificiell intelligens kan vara medveten på samma sätt som en människa och att den därför kan jämföras med och förklara mänskliga mentala tillstånd. Observera att de reaktioner på Det kinesiska rummet som jag tar upp varken allmänt anses utgöra ett enhetligt försvar av dessa påståenden eller motsvarar en historisk beskrivning av utvecklingen av AI, som vetenskap. Det kinesiska rummet används som utgångspunkt och stark AI-begreppet bevaras därför som fokus. I diskussionen så kommer det dock att ges en ny innebörd, efter det att Searles åsikt om datorfunktionalismen som en felaktig teori visas vara berättigad.

Jag börjar med att ta upp ett systemsvar av Daniel C. Dennett och sedan ett robotsvar av Stevan Harnad, vilka visar sig komplettera varandra. Jag menar att Dennetts invändningar klargör att den enda svårighet för funktionalismen som Searle påvisar med Det kinesiska rummet är semantikproblemet. Harnads svar bygger vidare på den funktionalistiska ståndpunkten genom att ta itu med just det problemet, dock inte genom att försvara datorfunktionalismen. Till sist ska jag ta upp en invändning, som Herbert A. Simon och Stuart A. Eisenstadt framfört, om Searles uppfattning av vad det innebär att förstå och betydelsen av detta för hans kritik av stark AI, samt för argumentationen som jag ser att Dennetts och Harnads svar utgör. De bemöter Searles påstående att turingtestets brister gör att funktionalistiska förklaringar saknar värde.

4.1. Systemsvaret

Systemsvaret är den första invändningen som Searle tar upp och bemöter i ”Minds, Brains and Programs”. Svaret går ut på att hela systemet som Det kinesiska rummet utgör, bestående av rum, papper, pennor, regeluppsättningar mm förstår kinesiska, trots att personen i rummet inte gör det. Den mänskliga komponenten liknas av Searle vid en dators processor, med funktionen att implementera ett program, d.v.s. bara en del av det system som ska tillskrivas ett kognitivt tillstånd. Searles uppfattning av systemsvaret är att det utgör ett mer eller mindre befängt försök att bemöta hans argument. Han ser ingen anledning att tro att ett formellt system plötsligt fylls av och behandlar mening bara för att det består av fler delar, som manipulerar formella symboler, än det kinesiska rummet. Han är dock medveten om att de slutsatser som han drar ur Det kinesiska rummet inte omöjliggör detta. För att visa på att hela systemet inte förstår mer än han själv gör, som en del av det kinesiska rummet, så modifierar han sitt tankeexperiment, genom att den som tidigare arbetade inuti rummet nu internaliserar det. Detta innebär att han, i en utveckling av tankeexperimentet, memorerar alla kinesiska rummets delar och deras funktioner, för att själv kunna utföra dem. I en sådan situation anser Searle att det är tydligare att varken någon del av systemet eller hela systemet förstår kinesiska. Trots att den som internaliserat rummet nu, genom introspektion, skulle kunna studera systemet på alla nivåer så tror Searle inte att denne kommer att uppleva någon sorts mening i de kinesiska symbolerna. Han menar också att en konsekvens av att sådana system plötsligt skulle anses som kognitiva är att intuitivt ickekognitiva system med input- och outputfunktioner skulle behöva omvärderas. Han exemplifierar med att våra magar skulle kunna sägas vara lika kognitiva som våra hjärnor.

John Preston skriver om systemsvaret att det egentligen inte är ett svar, utan är vad stark AI positionen innebär från början (Preston, 2002). Searles argumentation hänger därför, enligt Preston, på hans bemötande av systemsvaret**.** Daniel C. Dennett har formulerat ett svar som kan tolkas just så, i sin bok *Consiousness explained* (Dennett, 1992). Han argumenterar där för att de delar av ett system som är av betydelse inte är bundna till en viss sorts substans men ändå inte kan vara vilka som helst. Dessutom så menar han att det inte finns anledning att tro att personen i Det kinesiska rummet, eller någon annan komponent, bär egenskapen att förstå. Han ger alltså ett systemsvar och försöker visa att Searle inte gör någon som helst ansats att kritisera stark AI utifrån vad det faktiskt borde innebära utan baserar sitt argument på tveksamma intuitioner. Han bemöter Det kinesiska rummet på ett sätt som han anser kan avfärda Searles invändningar mot systemsvaret och stark AI. Efter att ha tagit del av

Dennetts reaktion på Det kinesiska rummet så tror jag att man kan se hur det intressanta i, och motivet bakom, Searles argument är semantikproblemet. Hur detta ska lösas berör Dennett inte, men att hans invändningar visar vad som är Det kinesiska rummets kärna och varför det är av betydelse för stark AI.

4.1.1. Rätt program

Dennett hävdar att det inte verkar finnas någon svårighet för dagens människor att föreställa sig en medveten maskin, t.ex. en av science fiction- litteraturens robotar. Problemet ligger istället i att föreställa sig *hur* en robot skulle kunna ha ett medvetande. Han beskriver Det kinesiska rummet som ett argument som försöker få oss att undvika att lösa detta problem, för att istället inrikta sig på att skapa en intuition om att dess lösning är omöjlig. Dennett vill visa att om man faktiskt strävar efter att testa stark AI i Det kinesiska rummet och inte hastar mot resultat baserat på en intuition om tesens felaktighet, så finns det ingen anledning att dela Searles pessimism gentemot tesen. Han utgår från en definition av stark AI, som han hämtar från Searle men som inte är formulerad på samma sätt som den jag nämnt under 2.1, vilken lyder: en rätt programmerad digital dator med rätt input och output har ett medvetande på precis samma sätt som människor har det. Den här formuleringen av stark AI har samma innebörd som den jag hanterat tidigare, men är mer explicit då den betonar rätt input-output-relation. Genom att uppge sig vilja testa den här definitionen så binder sig Searle till att simulera en dator som är programmerad på rätt sätt, vilket Dennett menar att han inte ens försöker göra. Searles arbete i rummet är ämnat att vara i analogi med en dators arbete utefter ett program, bestående i manipulation av ettor och nollor. I tankeexperimentet manipulerar han dock kinesiska tecken vilket Dennett uppmärksammar som ett sätt att försöka få läsaren att tro att tankeexperimentet också är i analogi med ett *fungerande* program, för språkförståelse. Det är, enligt Dennett, uppenbart att det program som Searle implementerar inte skulle fungera i verkligheten, vilket betyder att det inte skulle ge en godtagbar output. Men Searle vill ändå att vi ska godta rummet som om det var i analogi med en ”rätt programmerad” dator. Om man beskriver Det kinesiska rummet så som Searle så menar Dennett att det egentligen inte finns någon anledning att utsätta det för ett turingtest, eftersom det är uppenbart att programmet inte ens kan imitera någon som förstår kinesiska.

4.1.2. Dennett om medvetenhet

En liknande svaghet i Searles argumentation, som Dennett pekar ut, är att Det kinesiska rummet implicerar en tveksam teori om vad det innebär att vara medveten om något. I *Consiousness explained* argumenterar Dennett, fristående från sin diskussion av Det kinesiska rummet, mot det han kallar en cartesiansk teater- modell för medvetandet. Kortfattat så innebär den att det finns ett centrum i hjärnan där alla medvetna upplevelser sammanfaller och skapar den absoluta subjektiva upplevelsen. Dennett beskriver en sådan del av hjärnan som en teater där en medvetandeström visas för jaget. Han anser att en sådan teori om medvetenhet är allmänt ansedd som vederlagd. Searle antar en cartesiansk teater- modell genom att hävda att han borde förstå om det kinesiska rummet gör det. Då måste det dels skapas en sammanhängande upplevelse av att förstå och dels så måste han utgöra den del där upplevelsen uppstår. Om Dennett har rätt i detta så kan man ännu en gång dra slutsatsen att det inte är ett medvetande som Searles kinesiska rum simulerar; det är alltså inte programmerat på rätt sätt.

Detta gör också att Searles försök att förkasta systemsvaret genom att internalisera hela systemet verkar stå på en grund av tveksamma intuitioner, eftersom den cartesianska teatern då bara förflyttas till en högre nivå. Dennett argumenterar utifrån åsikten att ett program som skulle vara motsvarande det som krävs för att förstå måste vara enormt. Det skulle t.ex. bestå av flera nivåer för olika slags metakunskap, som kan behövas för att besvara frågor i ett turingtest. Han medger att det är svårt att föreställa sig ett sådant program men menar alltså att det är uppenbart att Det kinesiska rummet varken är komplext eller omfattande nog. Vad som då inte längre är uppenbart är att ett kinesiskt rum, som är programmerat på rätt sätt, inte skulle förstå. Dennett beskyller Searle för att försöka skynda förbi detta faktum, som kan omkullkasta betydelsen av tankeexperimentet, och därmed bygga sin slutsats på en felaktig premiss.

4.1.3. Mer av samma

Även om man behandlar Det kinesiska rummet som att det innehåller rätt funktioner så skulle Searle inte vilja påstå att det förstår. Att det kinesiska rummet inte innebär ett tillräckligt komplext program går Searle visserligen med på. Men han hävdar, förutom att stark AI inte bevisar att ett större program förstår, också att han har anledning att vara pessimistisk gällande att det skulle kunna visa sig vara så. Searles belägg för sin pessimism är, som jag

uppfattar det, att eftersom han inte förstår något i sitt kinesiska rum så förstår inte detta program något och därför kan inte heller ett större program förstå något. Dennett beskriver denna ståndpunkt som att ”mer av samma” inte ger ett nytt resultat, vilket han inte finner stöd för i Det kinesiska rummet. Han har visat att Searles argumentation både brister gällande en teori för medvetenhet och hur stark AI ska omsättas i praktiken (d.v.s. att hans program uppenbart inte skulle fungera) och därmed inte medför giltiga slutsatser. Därför anser Dennett att en skepticism gällande andra program än det i Det kinesiska rummet är helt obefogad. Det enda som Dennett kan se är att Searle underbygger sitt påstående om formella programs otillräcklighet med är att hävda att eftersom en liten bit hjärna inte kan förstå så kan inte heller hela hjärnan förstå. Här menar Dennett att Searle utnyttjar det faktum att det inte finns någon slutgiltig förklaring för hur en dator, eller en människa, är medveten.

4.1.4. Semantikproblemet

Dennett visar att Searle inte kan belägga några slutsatser angående *om* en artificiell intelligens kan förstå, varande ett formellt system eller ej, och hävdar att vi istället borde intressera oss för problemet om *hur* ett system kan förstå. Oavsett Det kinesiska rummets giltighet så menar jag att Searle berör just detta då han ifrågasätter om en artificiell intelligens kan ta hänsyn till och hantera den semantiska aspekten av input. Detta är vad jag kallar semantikproblemet, vars lösning är en aspekt av hur en artificiell intelligens kan vara medveten. Det är också det enda som återstår av Det kinesiska rummet efter Dennets behandling av det. Varför det är, eller inte är, möjligt att lösa semantikproblemet svarar Dennett inte på. Han argumenterar bara för att Det kinesiska rummet inte ger anledning att vara skeptisk gentemot stark AI-tesen, och därmed en artificiells intelligens’ möjlighet att hantera det. Att semantiskt innehåll har betydelse för ett formellt system är en förutsättning för stark AI, eller ett systemsvar, men detta verkar man varken kunna finna argument för eller emot i Det kinesiska rummet. Oavsett Searles positiva tes så utmålas hans pessimistiska intuitioner om stark AI därför, i mina ögon med rätta, som ogrundade av Dennett. Om Det kinesiska rummet ska ge stöd för att ett formellt system inte kan ha semantiskt innehåll så måste man gå med på följande: en cartesiansk teater- modell för medvetandet, att Det kinesiska rummet är rätt programmerat (om detta nu går utan föregående påstående) och att Searle skulle ha tillgång till den förmodade absoluta medvetandeströmmen.

Dennett har alltså visat att Det kinesiska rummet inte ger anledning att vara pessimistisk gällande stark AI, men han vill dessutom hävda att det finns anledning till att

vara optimistisk. Svårigheterna med vår förmåga att föreställa oss hur en artificiell intelligens skulle kunna se ut tror han att man kan övervinna med datorers hjälp, alltså den del av AI-projektet som Searle skulle kalla svag AI. Gällande svag AI som ett användbart hjälpmedel i medvetandeforskning så är Searle och Dennett alltså överens. Men de skiljer sig när det gäller om målet för sådan forskning kan, och bör, vara stark AI. Det som Dennett visar på som rätt riktning för stark AI är ett komplext system med processer som kan vara parallella, påverka varandra ömsesidigt och strukturerade i många olika nivåer. Ett sådant medvetande, artificiellt eller ej, har ingen enskild del som kan uppleva en absolut medvetandeström. Han argumenterar för ett systemsvar, där det är funktionerna hos systemets delar som tillsammans skapar medvetenhet och förmågan att förstå. Searle skulle som en del av ett sådant system inte förstå något i Det kinesiska rummet. Att internalisera hela systemet beskriver Dennett som omöjligt. Men om Searle skulle lyckas göra detta, med rätt program, så har han inget belägg för att systemet inte skulle förstå, eller att han ens skulle kunna avgöra detta.

Anhängaren av stark AI har alltså, enligt Dennett, inga hinder att ta itu med semantikproblemet. Hur något är medvetet är en öppen fråga. Searle hävdar i ”Minds, Brains and Programs” att bevisbördan, i denna fråga, ligger på datorfunktionalismen. Detta påstående verkar, efter Dennetts diskussion, fortfarande vara berättigat (lika berättigat som om det skulle ha riktats mot någon annan teori). Dennett beskriver ibland det system som ska bära ett artificiellt medvetande som en robot, men han avviker inte uttryckligen från Searles inriktning på datorfunktionalismen och en dators program som det som ger upphov till att den förstår. När man vill gå vidare med stark AI och ta itu med semantikproblemet så verkar det dock som om just robotar blir av betydelse. Prestons åsikt att systemsvaret egentligen är vad stark AI innebär sammanfaller med Dennetts syn på Det kinesiska rummet. Searles diskussion om huruvida han förstår eller ej blir irrelevant men samtidigt så har fokus hamnat på frågan om hur meningsinnehåll kommer in i bilden. Robotsvaret, i Stevan Harnads version, kan ses som ett försök att lösa detta. Lösningen är fortfarande funktionalistisk och systemsvaret lever vidare, men samtidigt så verkar Searles pessimism gentemot datorfunktionalismen bekräftas.

4.2. Robotsvaret

Searle beskriver, i ”Minds, Brains and Programs”, robotsvaret som att man skriver ett nytt program och placerar en dator med detta program i en robot, där det kan få robotens sensoriska och motoriska förmågor att fungera. Robotsvaret består, enligt honom, i att en robot kan ha mentala tillstånd på grund av sin förmåga att t.ex. få input från en kamera och

förstå hur den ska röra sig mellan två punkter. Preston sammanfattar robotsvaret som en reaktion på att det inte föreligger rätt kausala relationer mellan innehållet i symbolerna i t.ex. en persondator och världen (Preston, 2002). Rätt kausal relation uppnås då med sensoriska förmågor. I och med detta så ser jag det som att man har frångått den definition av stark AI som Searle likställer med datorfunktionalism, d.v.s. att endast det formella datorprogrammet är det som ska ge upphov till ett medvetande och förmågan att förstå. Searle verkar dock diskutera invändningen som ett system, manipulerande formell input, som ska förstå plus en funktion som står emellan det och världen. Han behåller alltså datorfunktionalismen i någon mån. Men stark AI kan tolkas enligt andra former av funktionalism, såsom visas nedan, vilket Searle dock menar är lika missvisande. Då han fortsätter att använda Det kinesiska rummet i sin argumentation mot robotsvaret så menar jag att man kan se det som att han riktar det mot funktionalism i allmänhet. Det som framträdde som Searles starkaste påstående ovan; att ett formellt system inte är tillräckligt för meningsinnehåll, är alltså det drivande i robotsvaret men används också i hans kritik av det.

Searle kritiserar robotsvaret genom att modifiera det kinesiska rummet så att det är datorn som fungerar som en robots hjärna. Rummet fungerar på precis samma sätt som tidigare med den skillnaden att nu kommer dess input från t.ex. videokameror och dess output tar sig uttryck genom robotens motoriska kapaciteter. Dessa skillnader är dock, enligt Searle helt omöjliga för honom att upptäcka när han manipulerar symboler inne i robotens ”hjärna”. Han menar att den information som kommer från robotens mekaniska sinnen varken kan ge upphov till intentionalitet eller ett tillstånd av att förstå. Detta eftersom att han kan förstå inputen i det här fallet lika lite som kinesiska tecken och inte heller förstå dem som ett resultat av sitt arbete. Resultatet är alltså det samma som i det ursprungliga tankeexperimentet. Om systemets input kommer från en kamera eller en person, består i kinesiska tecken eller andra symboler, har ingen betydelse enligt Searle. Det kinesiska rummets relevans kan ifrågasättas på samma sätt när det appliceras på robotsvaret som i samband med systemsvaret, ovan. Hans tveksamma påståenden om medvetenhet och utformningen av programmet har fortfarande betydelse. Men Dennetts svar på Det kinesiska rummet har flyttat fokus från *om* ett artificiellt medvetande kan skapas till *hur* t.ex. en robot skulle kunna vara medveten. I och med detta så måste man behandla semantikproblemet. När Stevan Harnad gör det, i ”Minds, Machines and Searle” (Harnad, 1989), så verkar det som att ett större program, i enlighet med datorfunktionalismen, inte är tillräckligt.

4.2.1. Från simulation till mekanism

Harnad håller med Searle om att ett rent symbolmanipulerande system, t.ex. ett datorprogram, inte kan ge upphov till kognitiva tillstånd på egen hand, men hävdar att en robot har andra förutsättningar. För att påvisa skillnaden mellan t.ex. ett datorprograms och en robots förmåga att förverkliga stark AI så tar han upp begreppet ”*simulation*”. Begreppet är viktigt eftersom Searle, enligt Harnad, tolkar det som imitation och hävdar att stark AI inte kan vara något annat på grund av en fundamental brist på intentionalitet (se biologisk naturalism ovan). ”Simulation” definierar Harnad som att man ger en dator all information man har om ett ting: dess delar, egenskaper och relationer, dem emellan samt till världen. Om man lyckas med detta så fungerar simulationen av tinget på samma sätt som tinget fungerar i världen och man har lyckats med att formalisera de kausala egenskaperna. Man har dock inte duplicerat tinget i sig. Jag tolkar det som att funktionerna som simulationen består av inte innebär de relevanta kausala egenskaperna, eftersom relationerna inte är mellan rätt entiteter utan mellan godtyckliga representationer. Symbolerna i simulationen är då inte representationer för systemet utan fungerar bara som sådana i medvetandet hos personen som skapade den, vilket innebär det som Searle kallar härledd intentionalitet. Om ett rent symbolmanipulerande system är målet för stark AI så kan det vara berättigat att kalla det för imitation. Som Searle visat så kan man då komma undan med att bara ge sken av att ha skapat ett system med de relevanta kausala egenskaperna, d.v.s. en imitation. Men när Harnad diskuterar formella system som simulationer så är det som ett steg på vägen mot ett system med inneboende, istället för härledd, intentionalitet. Han menar att man därför måste ha andra ambitioner än att skapa en imitation i simulationen i Det kinesiska rummet. Som nämnt så kritiserar Dennett Searle för att inte ha det.

Harnad håller med Searle om att en simulation inte är en duplikation, eftersom relationerna i en simulation inte är mellan rätt (relevanta) entiteter. Men simulation är, enligt Harnad, ett viktigt verktyg i forskning kring AI. Han menar att man bör använda simulationer som modeller, vilka när de bedöms innebära rätt funktioner kan omsättas i verkligheten. Att simulera ett medvetande innebär för honom ett steg på vägen mot ett robotsvar. Han anser att datorfunktionalismen aldrig tar just det steget. Det kinesiska rummet, i sin ursprungliga form, ser Harnad som en simulation och menar att Searle, med sitt tankeexperiment, egentligen ställer frågan: kan man hävda att en dator som simulerar beteenden som kan tolkas som att den förstår också faktiskt förstår? Inne i rummet påstår Searle att han inte kan komma att förstå kinesiska. Därför hävdar han att både en teori om medvetandet som ett datorprogram

och turingtestet som metod att avgöra om förståelse finns kan anses vara vederlagda (Searle, 1980). Harnads uppfattning är att Searle har rätt angående ett kinesiskt rum som bara manipulerar formella symboler och stark AI bundet till datorfunktionalism. Men Harnads svar fortsätter där Dennett slutade genom att han försöker visa vad som krävs för att gå från ett simulerat medvetande till ett verkligt medvetande, d.v.s. hur en robot kan vara medveten. En robot innebär då ett system som inte enbart består av formella funktioner.

En simulation tar ett steg närmre rätt kausala egenskaperna, men trots att rätt funktioner återges så de inte kallas relevanta förrän man *implementerar* dem. När Harnad använder begreppet ”implementera” så menar han alltså att bygga ett fysiskt ting vars egenskaper har formaliserats och testats i en dator. Generellt så menar man med ”implementera” att ett program körs på en dator. Det som Harnad benämner som att vara implementerat är inte verksam mjukvara utan konkreta fysiska tillstånd, verksamma genom kausala relationer till omvärlden. Han påpekar dock att implementerad mjukvara också innebär faktiska fysiska tillstånd och förändringar av dessa, i en digital dator, men poängen är att han inte vill begränsa sig till ett sådant specialfall. Genom att implementera en modell så skapar man det som Harnad kallar en ”*mekanism*”. ”Mekanism” definierar Harnads som: ”ett fysiskt system som fungerar enligt fysiska lagar”, vilket är en etikett som han sätter på såväl datorer och flygplan som människor och djur. Det är när man vill förstå en mekanism som en simulation kan vara användbar. Då strävar man efter att ta reda på vilka funktioner som ger en mekanism dess plats i en kausalkedja genom att testa den mot en formalisering av världen. När modellen av mekanismen fungerar i modellen av världen, på ett tillfredställande sätt, så kan man anta att man vet vilka funktioner som ska implementeras för att duplicera de relevanta kausala egenskaperna. Eftersom Harnad försöker svara på Det kinesiska rummet så avser han inte att redogöra för precis vilka funktioner detta innebär utan endast vilka funktioner som krävs för att ett system ska kunna ha inneboende intentionalitet.

4.2.2. Hårdvarumisstaget

Att blanda ihop implementering i en dator och implementering i Harnads vidare mening, när man försöker realisera stark AI, kallar han för ”*hårdvarumisstaget*”. Han ser att det finns en tendens att i tankeexperiment ta den mänskliga hjärnan, friställd resten av kroppen, för den mekanism som ger upphov till medvetenhet och likställa denna med en dator. Då skiljer man också på dess tillstånd, d.v.s. individuerar dess egenskaper, på samma sätt som olika tillstånd i ett program kan sägas vara en dators olika tillstånd. Harnad menar att en hjärna som är skild

från sin kropp bara är en del av det system som kan ha intentionala tillstånd. Detta på grund av att kroppens frånvaro innebär avsaknaden av input- och outputfunktioner. Harnad vill med hjälp av begreppen ”simulation”, ”implementering” och ”mekanism” komma fram till att: En simulation besitter inte de relevanta kausala egenskaperna som gör att en mekanism fungerar på ett visst sätt. Simulationer är modeller. En implementering av mjukvara på hårdvara ger inte upphov till en mekanism som kan vara medveten om omvärlden, utan bara en simulation av en sådan. Detta beror på funktioner för input saknas. Ett kinesiskt rum som förstår är därför inte möjligt så länge det är ett rent symbolmanipulerande system. Systemet i Det kinesiska rummet, som det såg ut från början, är enligt Harnad bara en datorsimulation. Den här simulationen menar han att Searle sedan simulerar genom att ämna utföra datorns arbete. Det kinesiska rummet handlar då om en simulation av en simulation. Om systemet som ska testas bara består av symbolmanipulerande funktioner så kan detta kanske vara berättigat. Men en simulation av ett robotsvar kan, som framgår ovan, inte innebära duplikation av samtliga funktioner.

4.2.3. Searle som robot

Searle och Harnad når samma slutsats gällande det ursprungliga kinesiska rummet; det förstår inte. Men när man modifierar detta i enlighet med ett robotsvar så går deras uppfattningar isär. Searle säger, i sin kritik av robotsvaret, att om man förser en mekanism med sensorer och möjlighet att röra sig så är det ändå inte möjligt för den att förstå input. Harnad ger ett robotsvar som också innebär ett systemsvar, alltså att det inte är en specifik del av systemet som ensam kan kallas medveten. Searles reaktion måste då, enligt Harnad, bli att, som i hans reaktion på systemsvaret, utföra alla det nya kinesiska rummets funktioner på egen hand. Rummet ser nu ut som så att det finns: En kamera som registrerar världen utanför rummet. En funktion som omvandlar kamerans information till kod. Ett program som behandlar rummets input och ger en output, vilket Searle implementerar. Samt motoriska funktioner som styrs av den output som Searle producerar, efter att ha omvandlats av en funktion lik den mellan kameran och Searle. Harnad invänder mot Searles försök att avfärda robotsvaret genom att visa hur han, trots allt, inte kan undgå systemsvaret. Om Searle bara sysslar med att manipulera den input som kameran samlar in så misslyckas han med att utföra själva kamerans uppgift; att tillgå informationen innan den blir något annat än det som passerade kamerans lins. I ett sådant fall så fungerar ett systemsvar. Detta berör Searles reaktion mot robotsvaret som det ser ut i ”Minds, Brains and Programs”. Om han däremot tittar på

omvärlden, och får input på samma sätt som en kamera, så utför Searle den funktion som behövs för att omvandla den input som kameran tar emot. Harnads poäng är att i så fall kommer Searle faktiskt att förstå symbolerna som han sedan ska manipulera. Om ett sådant försök att förkasta robotsvaret är av vikt ser jag som tveksamt. Dels så innebär Searles roll en cartesiansk teater-modell och dels så är det i hans hjärna som det kognitiva tillståndet uppstår. Harnad verkar dock ha visat att en robot kan tillgodogöra sig information som inte definieras formellt.

4.2.4. Symbolfunktionalism och robotfunktionalism

Harnad uppmärksammar att det finns många organismer som har sensoriska och motoriska förmågor men inte kan hantera språk, samtidigt som det inte finns exempel på det motsatta. Detta är ytterligare något som leder honom att tro att förmågan att förstå kinesiska, eller andra, symboler förmodligen har sin grund i ickesymboliska, d.v.s. ej formella, funktioner. Därför är datorfunktionalismen utesluten när man vill bygga ett kinesiskt rum. Harnads ståndpunkt är dock, trots avfärdandet av datorfunktionalismen, fortfarande funktionalistisk. Han delar upp funktionalism i ”*symbolfunktionalism*” och ”*robotfunktionalism*”. Symbolfunktionalism innebär att mentala funktioner endast är symboliska funktioner, likt det som händer i en dator då mjukvara används. Detta är alltså samma det samma som datorfunktionalism. Harnads egen ståndpunkt är robotfunktionalism, som innebär att ickesymboliska funktioner är nödvändiga för kognitiva tillstånd. Searle diskuterar inte en distinktion likt Harnads i samband med stark AI eftersom datorfunktionalism är liktydigt med stark AI för honom. Men Harnad anser sig alltså ha anledning att skilja på begreppen. Hans ståndpunkt är att stark AI kan förverkligas på bas av robotfunktionalism. Det är dock tydligt att Searle både vill förkasta robotsvaret och alla andra former av funktionalism, som kan ligga bakom stark AI, då han har hävdat att ”If you are tempted to functionalism, I believe you do not need refutation, you need help” (Searle, 1992).

4.2.5. Symbolisk och ickesymbolisk kod

Symbolfunktionalismen, d.v.s. datorfunktionalismen, klarar, enligt Harnad, inte av att lösa semantikproblemet eftersom symboler, i ett system byggt efter sådana principer, alltid definieras av andra symboler. Han menar att de då inte kan vara ”grundade” i hur världen, utanför det formella systemet, ser ut och att de därför blir meningslösa i förhållande till den.

Den lösning han ger på detta innefattar ickesymbolisk input från sensorer och en funktion för att förvandla sinnesinput till kod. Sådan kod kan en mekanism sedan behandla som information om verkligheten. Detta anser han beskriver en funktion hos både människor och robotar. I Harnads ögon så kan Searle alltså inte invända mot att en mekanism, byggd på robotfunktionalismens principer, kan lösa semantikproblemet utan att också invända mot en människas förmåga att göra detta. I vilken utsträckning som hjärnan sedan behandlar input i formella system är en öppen fråga, men att det förekommer, menar Harnad, är uppenbart. Detta kan beläggas med att det t.ex. uppenbarligen är en representation av en stol som man har i huvudet då man upplever en sådan framför sig. Att resonera om stolen är då i sin tur att manipulera symboler. Harnad gör därför ytterligare en viktig distinktion; mellan ”*symbolisk kod*” och ”*ickesymbolisk kod*”. Symbolisk kod förklarar han som fysiska markörer som manipuleras utefter ett godtyckligt formellt regelsystem. Betydelsen av sådana symboler ges av en konvention. Relationen mellan symbolen och ett objekt är därför avhängig intentionaliteten hos den som skapade regelsystemet, vilket innebär samma sak som Searles härledda intentionalitet. Ickesymbolisk kod får betydelse genom fysiska och kausala relationer mellan markören, eller symbolen, och objekt i världen, på ett icke godtyckligt sätt. Därför är symbolens relation till världen inte godtyckligt bestämd. Ickesymbolisk kod kan då ge upphov till relevanta kausala egenskaper och intentionalitet i ett system. Detta är den sorts kod som Harnad menar att en robots sensorer, precis som våra sinnesorgan, ger som input. Det är vad robotens sensoriska egenskaper bidrar med. När det gäller att testa dess förmåga att förstå världen, genom möjligheten till intentionalitet, så blir dess motoriska förmågor viktiga.

4.2.6. Det totala turingtestet

Turingtestet är, enligt Harnad, inte ett adekvat test av en robots förmåga att förstå. Han anser att om man både ger input i symbolisk kod och kräver output i symbolisk kod så är intuitionen att också mellanleden består av symbolisk kod berättigad. Att testa om en mekanism förstår symbolers mening bör alltså, enligt honom, inte göras genom språklig kommunikation. Searles skepticism gentemot utfallet av turingtestet kan då ses som berättigad eftersom det är utformat på ett sådant sätt att det inte tillåter någon annan slutsats. Att en simulation, som är ett rent formellt system, i princip kan lyckas genomgå ett turingtest utan att förstå något är både Searles och Harnads uppfattning. Harnad föreslår därför istället ”*det totala turingtestet*”. Detta ska mäta förmågor och avgöra egenskaper genom att observera alla beteenden hos en mekanism istället för bara dess förmåga att producera språk. Searle påstår att mätande av

beteenden inte kan bestämma om en mekanism förstår. Detta grundar han på att det ursprungliga kinesiska rummet klarade av turingtestet utan att förstå. Jag ser en möjlighet att också lura det totala turingtestet. Man kan tänka sig en robot med en fast uppsättning beteenden som direkt kopplas till olika input och därför ger intrycket att förstå sin omgivning. Detta skulle dock uppenbarligen inte innebära ett program som är rätt program, för att motsvara ett mänskligt medvetande. Harnad verkar dock ha visat hur semantikproblemet kan lösas utan hänsyn till substans. Därmed kan ett test som mäter beteende anses vara lika relevant för alla mekanismer. Han erkänner dock också att det totala turingtestet måste vara ”provisoriskt” och därmed bara kunna indikera kognitiva egenskaper, inte visa att de nödvändigtvis föreligger. Att testet ändå är användbart bygger han dels på att alla empiriska hypoteser kan falsifieras någon gång i framtiden och dels eftersom det inte verkar finnas någon annan metod att bestämma medvetenhet. Herbert A. Simon och Stuart A. Eisenstadt ger bland annat det argumentet när de diskuterar vad det innebär att testa om något förstår (Simon och Eisenstadt, 2002). Som jag ser det så kan deras svar på Det kinesiska rummet ses som belägg för att stark AI inte bara kan innebära medvetenhet utan också förklara det mänskliga medvetandet.

4.3. Stark AI som förklaring av det mänskliga medvetandet

I ”*Stanford Encyclopedia of Philosophy*” (Cole, 2004) nämns ett svar på Det kinesiska rummet kallat ”intuitionssvaret”. Det innefattar de reaktioner som utgår från Searles åsikt om vad som krävs för att påstå att någon förstår. Som det framgår av Dennetts diskussion så baserar han den åsikten på biologisk naturalism och därför så tror jag att det är berättigat att kalla detta för ett påstående utan belägg, eller en intuition. Intuitionssvaret är en kategori som Searle inte tar upp själv. Detta tror jag kan förklaras av att han inte anser det vara av vikt att formulera klara kriterier för vad det innebär att förstå. Han nöjer sig med att påstå att det finns fall då vi utan tvekan kan säga att något förstår respektive inte förstår. I motsats till Searles pessimism mot att en artificiell intelligens kan förstå så är Herbert A. Simon och Stuart A. Eisenstadt väldigt optimistiska när de diskuterar detta i ”A Chinese Room that Understands” (Simon och Eisenstadt, 2002). De går så långt som att hävda att det redan i skrivande stund finns syntetiska system som faktiskt förstår (de som nämns kallas EPAM, ZBIE och CaMeRa). Simon och Eisenstadt ger, i någon mån, belägg för sin optimism när de diskuterar vad det innebär påstå att något förstår och visar hur man kan testa stark AI. De erkänner att turingtestet inte fungerar men använder sig ändå av observationer av beteende och visar varför

detta inte behöver betyda samma mått av brist på förklaringskraft som gjort att behaviorismen vederlagts.

4.3.1. Att förstå

En grundsten i Simons och Eisenstadts argumentation är att en artificiell intelligens kan uppvisa beteenden som om de utfördes av en människa skulle ge oss anledning att hävda att denne förstår. I båda fallen måste man i normalfallet tillskriva egenskapen utifrån observation av beteende. För att invända mot Searles användning av begreppet så ger de följande analys: (1) Begreppet är operationellt. Vilket betyder att det ska definieras genom att ange de metoder som krävs för att det ska kunna användas. (2) Begreppet motsvarar den mening som vardagsspråkstalaren ger det. (3) Samma kriterier ska användas för att kunna tala om alla ting, människor eller ej, som att de förstår. De undviker, precis som Searle, att ge en definition av begreppet, eftersom de inte anger några kriterier för att applicera det. Poängen är dock att huruvida något förstår ska avgöras med ett test mot sådana kriterier. Kriterierna för att förstå baseras, enligt Simon och Eisenstadt på beteenden och det finns därför ingen anledning att på förhand sluta sig till ett resultat i vissa fall. Om man går med på Dennetts och Harnads försvaret av stark AI så håller jag med Simon och Eisenstadt om att Searle inte har någon anledning att vara skeptisk mot att en robot förstår. D.v.s. han har inte mer anledning att vara mer skeptisk mot att en robot förstår än att en människa gör det, om han observerar beteenden.

Searle vill, med Det kinesiska rummet, visa att stark AI inte kan förklara mänskliga kognitiva tillstånd. Då man, trots detta, söker en sådan förklaring via Searles tankeexperiment så verkar det som om man inte kan undgå att förlita sig på vardagsanvändningen av ”förstå”. Den kritik som Simon och Eisenstadt riktar mot Searle är alltså inte att han är vag när det gäller vad det innebär att förstå utan att han inte använder begreppet konsekvent. Man bör, enligt dem, använda samma test, med samma vagheter, som gör att vi kan säga att en människa förstår när vi avgör om en artificiell intelligens gör det. (De finner stöd för den här uppfattningen i att Quine har påpekat att den mänskliga förmågan att förstå är förknippad med den grad av osäkerhet som gäller all empirisk undersökning.)

4.3.2. Empirisk respektive logisk stark AI

Osäkerheten i att påstå att en artificiell intelligens förstår är, för Simon och Eisenstadt, något okontroversiellt. De menar att frågan om huruvida någon eller något förstår avgörs empiriskt och vill därför kalla sin ståndpunkt för ”*empirisk stark AI*”. De ställer detta mot den innebörd som stark AI hade när Searle först använde begreppet; att ett system som passerar turingtestet uppfyller de nödvändiga villkoren för att förstå. Detta kallar de för ”*logisk stark AI*”. Empirisk stark AI innebär att kognitiva tillstånd ska tillskrivas en mekanism utan anspråk på nödvändighet. Man kan formulera mer eller mindre säkra kriterier för sitt test men, som Harnad påpekade, det föreligger alltid en osäkerhet gällande resultatet. Logisk stark AI innebär motsatsen; att kognitiva tillstånd ska påvisas med nödvändiga och tillräckliga kriterier. Turingtestet verkar, som Searle framställer det, vara det enda sättet att testa kognitiva tillstånd för en datorfunktionalist och det måste då uppfattas som just ett nödvändigt och tillräckligt kriterium. Searle kan då påstå att datorfunktionalismens förklaringskraft är väldigt liten, eftersom man inte studerade kognitiva tillstånd i sig. Detta blir tydligt i Det kinesiska rummet då han visar hur han kan lura turingtestet. Simon och Eisenstadt delar Searles uppfattning om turingtestet och logiskt stark AI, som en felaktig tes respektive metod. De håller dock inte med om att stark AI måste innebära en medvetandemodell som saknar förklaringskraft. Jag tolkar deras ståndpunkt som att funktioner i ett medvetet system inte behöver definieras i termer av dess beteendemässiga input-output-relationer. Och att ambitionen för empirisk stark AI inte är att belägga en färdig tes utan att söka goda jämförelser mellan människa och syntetiskt system, i enlighet med de, vaga, uppfattningar vi har om medvetenhet och kognitiva tillstånd. Därmed kan man finna nya, bättre, förklaringar av dessa uppfattningar. Simon och Eisenstadt påpekar att empirisk stark AI liknar svag AI, men skiljer sig genom att hävda artificiella intelligensers möjlighet att vara medvetna. Detta påstående belägger de alltså med att inget nödvändigt och tillräckligt kriterium finns för vad som är medvetet. Att det inte finns några principiella hinder för att en artificiell intelligens kan likställas med en människa i en jämförelse, komna ur Det kinesiska rummet, anser jag att man kan hävda via argumenten under 4.1 och 4.2.

4.3.3. Ett nytt test

Turingtestet, som det används i Det kinesiska rummet, är i enlighet med empirisk stark AI: s principer, i den meningen att det inte behöver användas som av datorfunktionalister och

anhängare av en logisk stark AI. Det kan anses vara nödvändigt för ett system som förstår att klara av ett turingtest, men det måste inte tolkas som ett nödvändigt och tillräckligt kriterium. Det är ett test som också kan användas för att visa på medvetenhet i enlighet med empirisk stark AI, men som sådant så menar Simon och Eisenstadt att det är ett dåligt test. De anser att Det kinesiska rummet illustrerar möjligheten för dess inneboende, människa eller maskin, att lura en observatör till att missta sig angående om systemet förstår kinesiska eller ej. Deras invändning mot turingtestet är alltså att det behövs ytterligare kriterium att testa efter, för att på ett bra sätt återge innebörden av att förstå. Detta är samma anledning som Harnad hade till att formulera det totala turingtestet. Den som sitter i det kinesiska rummet uppvisar god förmåga att behandla input utefter syntax, men saknar förmågan att förstå kinesiska. Personen kan därför dölja att systemet inte kan hantera mening i vare sig svar eller frågor, p.g.a. oförmögenhet att relatera ett föremål i världen till en symbol. Författarna menar alltså att man måste konstruera ett test som fungerar bättre än turingtestet genom att minska möjligheten att det ger ett missvisande resultat.

För att kunna avslöja en simulation av att förstå och undanröja turingtestets svaghet i detta avseende så föreslår de att man bygger ett fönster i det kinesiska rummet. Personen i rummet kan då ta del av de potentiella denoteringarna som fönster mot omvärlden bjuder. Man kan då fråga efter namnet på en process, ett ting eller relationer i världen genom att peka. Simon och Eisenstadt beskriver sådana frågor som det vanligaste sättet att testa om en människa förstår. Att man kan ställa frågorna genom att peka gör att den som svarar måste förstå input direkt från världen istället för via omvägen om symbolisk kod. En sådan fråga vill de alltså ställa om något som är inom svararens synfält och pekas ut ostensivt. Ett riktigt svar, på ett naturligt språk, gör, enligt Simon och Eisenstadt, att vi bör dra slutsatsen att systemet i fråga förstår meningen i namnet som det har kommunicerat. Förutom att ombyggnaden av det kinesiska rummet representerar en möjlighet till intentionala tillstånd så innebär det också att man kan observera alla beteenden hos personen i rummet. Fönstret ger alltså också, då det kinesiska rummet är i analogi med en artificiell intelligens, möjligheten att kunna studera vilka funktioner som ligger bakom systemets output.

Författarna tar upp ytterligare en förbättring av ett test för stark AI: s förmåga att förstå. Man kan, som jag gjorde ovan, spekulera i möjligheten att en programmerare skapar de funktioner som krävs för att en robot ska kunna passera ett beteendebaserat test. Om programmeraren då skapar ett system med förutbestämda kopplingar mellan en viss input och en viss output så fungerar roboten som en förlängning av programmerarens intentionalitet, d.v.s. den instantierar härledd intentionalitet. För att undvika detta så menar de att man måste

lägga vikt vid en inlärningsprocess. När kopplingen mellan svar och fråga går via resultatet av en inlärningsprocess, inte en programmerare, så vill de hävda att en artificiell mekanism inte har ett direkt och mekaniskt stimuli - respons- förhållande till input.

Simon och Eisenstadt menar inte att man ska låsa sig vid ett test som det ovan. Det är enligt dem ett bra sätt att avgöra om något förstår, men det säger inte särskilt mycket om det kognitiva tillståndet som är förknippat med att förstå. För att få en sådan förklaring så menar de att studier av ett syntetiskt system är överlägset studier av en mänsklig hjärna. Det syntetiska systemets funktioner kan studeras på mycket mer detaljerade nivåer än den beteendemässiga. Det är något som man, i de flesta fall, inte kan göra när man studerar en människa direkt. Därför är syntetiska system av stort värde om de faktiskt kan fungera som en förklaring till mänskligt mentalt innehåll. Man kan utföra undersökningar av systemets mindre och inre delar, som kan ge kunskap om vad som implementerar de mentala tillstånden. Simon och Eisenstadt pekar på att man också har en fördel genom att man känner till det syntetiska systemets struktur. De menar att det finns goda möjligheter att, via empirisk stark AI, bygga funktionalistiska förklaringsmodeller för vad kognitiva tillstånd innebär, från botten upp. De formulerar den åsikten, i behavioristisk jargong, som att systemen ger svar på vad som finns i den svarta lådan, mellan input och output.

5. Sammanfattning

Det kinesiska rummet har uppenbart betydelse för både vad artificiell intelligens innebär och hur man bör förstå det mänskliga medvetandet. Med argumentet riktar Searle, explicit, kritik mot påståendena: (1) En riktigt programmerad dator kan vara medveten på precis samma sätt som en människa. (2) Datorns program utgör därför en förklaring av det mänskliga medvetandet. Dessa påståenden definierar det som Searle kallar ”stark AI”, vilket är en konsekvens av datorfunktionalismen. För att visa att det inte finns några belägg för stark AI så simulerar han det arbete som han hävdar att en dator utför för att förstå kinesiska, i enlighet med datorfunktionalismen. Resultatet av hans arbete är, enligt Searle, att han inte kommer att förstå kinesiska. Han drar därav slutsatserna att datorfunktionalismen står utan belägg, ett formellt system kan inte lösa semantikproblemet och att det inte finns en godtagbar metod att testa en funktionalistisk tes med.

Dennett invänder att Searle bygger sina slutsatser på tveksamma intuitioner och saknar en godtagbar argumentation. Han hävdar att Searle uppenbart inte testar ett system

som är i enlighet med stark AI och att om han gjorde det så finns det ingen anledning att dra några slutsatser av att han inte är förmögen att förstå kinesiska. Dennett flyttar fokus från om en artificiell intelligens kan vara medveten till hur den kan vara det. Den frågan berör också Searle, i Det kinesiska rummet, genom att han betonar vikten av inneboende intentionalitet för en lösning på semantikproblemet. Han visar också hur ett systemsvar ligger närmre en modern uppfattning av medvetandets struktur än Searles ståndpunkt. Jag använder alltså Dennetts argumentation för att visa på semantikproblemet och att dess lösning är en öppen fråga.

Harnad kan ses som att ta vid där Dennett slutade och ge en lösning på semantikproblemet. Han förkastar datorfunktionalism och framhåller robotfunktionalism. Robotfunktionalismen innebär att ett syntetiskt system med sensoriska och motoriska förmågor kan tillgodogöra sig ickesymbolisk kod och därmed ha förutsättningar för inre intentionalitet. Att tro att man kan duplicera ett medvetet system genom att simulera också dess ickeformella funktioner kallar Harnad för hårdvarumisstaget. För att undgå att låta sig luras av en skickligt utformad simulation så vill han testa ett systems medvetenhet i det totala turingtestet. Detta innebär att man observerar systemet i alla avseenden, istället för att koncentrera sig på formell output, som i turingtestet. Jag menar att ett robotfunktionalistiskt systemsvar har potential att bära ett medvetande, i ljuset av Dennetts och Harnads reaktioner på Det kinesiska rummet. Därmed så bör man också omdefiniera stark AI. En ny formulering skulle kunna vara att en robot (med nödvändiga sensoriska egenskaper) kan vara medveten på samma sätt som en människa och därmed så förklarar dess funktioner mänskligt medvetande. Den senare delen i denna definition kräver dock ytterligare belägg, vilket jag finner hos Simon och Eisenstadt.

Simon och Eisenstadt håller med Harnad om att ett test av relationen mellan formell input och formell output innebär att ett formellt system skulle kunna prestera på samma sätt som ett medvetet system. De utgår från hur vi tillskriver människor förmågan att förstå och menar att ett test med sådana kriterier också ska användas på syntetiska system. Den empiriska undersökningen är grundläggande för dem och de vill därför skilja på empirisk stark AI och logisk stark AI. De argumenterar för empirisk stark AI, som innebär att ett test av medvetenhet aldrig kan innehålla nödvändiga och tillräckliga kriterier. Osäkerheten beror varken på stark AI eller funktionalism. De påpekar fördelarna med att studera medvetenhet i ett syntetiskt system och att dessa verkar göra funktionalistiska förklaringar både testbara och värdefulla. Då system- och robotsvaren ger belägg för att en artificiell intelligens kan vara

medveten så ger Simon och Eisenstadt belägg för att denna medvetenhet kan jämföras med och därför förklara mänsklig medvetenhet.

Stark AI kan alltså försvaras, dock inte i sin ursprungliga form. Lösningen på semantikproblemet innebär att datorfunktionalism måste bytas ut mot robotfunktionalism. Då bör man också tala om robotens funktioner, snarare än dess program, som det som kan likställas med och förklara en människas medvetande. Min slutsats, av att sätta de tre skilda invändningarna under 4.1, 4.2 och 4.3 i ett sammanhang, är alltså att man kan använda Det kinesiska rummet för att driva en utveckling av stark AI-tesen, istället för att vederlägga den. Jag delar dock inte Simons och Eisenstadts optimism gällande medvetenheten hos de syntetiska system som finns idag, utan lutar mer åt Dennetts ståndpunkt att ett system med rätt struktur ligger framför oss. Jag menar att empirisk stark AI baserad på robotfunktionalism ger en god utgångspunkt för sådan forskning och att Det kinesiska rummet därför kan ses som ett fruktbart argument, på ett sätt motsatt dess upphovsmans intentioner.

6. Referenser

Cole, David (2004), "The Chinese Room Argument", *The Stanford Encyclopedia of Philosophy (Fall 2004 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2004/entries/chinese-room/>.

Dennett, Daniel C. (1992), *Consiousness Explained*, Allen Lane, Harmondsworth, 1992.

Harnad, Stevan (1989), "Minds, Machines and Searle". Journal of Theoretical and Experimental Artificial Intelligence 1: 5-25.

Preston, John (2002). (Red.) Preston, John och Bishop, Mark, *Views into the Chinese room*. Clarendon Press, Oxford, 2002.

Searle, John R. (1980), "Minds, brains and programs". Behavioral and Brain Sciences 3 (3): 417-457.

Searle, John R. (1992), *The rediscovery of mind*. Cambridge, Massachusetts, 1992.

Searle John R. (1994), "Searle, John R.". (Red.) Guttenplan Samuel, *A companion to the philosophy of mind*. Blackwell, 1994.

Simon, Herbert A. och Eisenstadt Stuart A. (2002), "A Chinese Room that Understands". (Red.) Preston, John och Bishop, Mark, *Views into the Chinese room*. Clarendon Press, Oxford, 2002.