JAN ERIC LARSSON

Kan Searle tänka?

Kan datorer åstadkomma medvetande? John Searle, professor i filosofi vid University of California, Berkeley, hävdar att så inte är fallet. Han angriper en ståndpunkt som kallas *stark AI*, medan den försvaras av många andra. Searles resonemang är centrerat kring ett exempel som kallas *det kinesiska rummet*, vilket har givit upphov till stor debatt inom filosofi- och AI-kretsar.

Min egen åsikt är att vare sig Searle eller företrädarna för stark AI kan veta att de har rätt. Vi kan för närvarande inte veta huruvida datorer kan åstadkomma medvetande, och med Searles medvetandeteori kan vi aldrig komma att veta det. Searle har emellertid klart deklarerat att han anser sig kunna visa att stark AI är principiellt omöjlig. Hans resonemang är emellertid bristfälligt och jag hoppas att i det följande kunna visa att det inte duger som argument mot stark AI.

1. Searles argumentation

Låt oss inleda med att beskriva Searles argumentation.

Stark och svag AI

Searles mål är att angripa en ståndpunkt hos vissa AI-forskare, som säger att man genom att skriva ett program för en dator *de facto* kan skapa ett medvetande.

> Ett stort antal forskare inom artificiell intelligens (AI) tror... att de genom att skriva rätt program, med rätt inmatning och utmatning, *de facto* skapar medvetanden, (1990, s 20).

Denna ståndpunkt kallar Searle för *stark AI*, i motsats till *svag AI*, som inte nödvändigtvis anser det möjligt att skapa medvetande, utan hävdar att AI-tekniken har andra värden. Vidare hävdar Searle att

Filosofisk tidskrift 1992 nr 4, 11-27

man inom stark AI anser sig ha ett test för att avgöra om man åstadkommit medvetande, nämligen *Turingtestet.*

> De tror vidare att de har ett vetenskapligt test för att avgöra sin framgång: Turingtestet, som beskrevs av Alan M Turing, den artificiella intelligensens fader. Turingtestet, i nuvarande tolkning, är helt enkelt: om en dator kan uppträda på så sätt att en expert inte kan skilja den från en människa som har en viss kognitiv förmåga – till exempel förmågan att addera tal eller förstå kinesiska – då har även datorn den förmågan, (1990, s 20).

Så långt Searles distinktion mellan stark och svag AI. Vi skall ha anledning att komma tillbaks till Searles uttalanden, men först skall vi ge en kort beskrivning av de två tänkta experiment som förekommer i argument och motargument.

Det kinesiska rummet

I korthet kan experimentet med det kinesiska rummet beskrivas enligt följande. Låt oss välja en försöksperson, i det aktuella fallet Searle själv, och ett språk som denne inte förstår, till exempel kinesiska. (Vi bortser med Searle från att kinesiska inte är *ett* språk. Den nitiske läsaren kan i det följande ersätta kinesiska med exempelvis mandarin.) Antag att vi stänger in Searle i ett rum i vilket finns en mängd lappar med kinesiska skrivtecken, ett hål genom vilket sådana lappar kan skickas in och ut, samt ett (mycket komplext) bokverk. Detta bokverk innehåller anvisningar om vilka skrivtecken som skall skickas ut till omvärlden, med hänsyn till vad för tecken denna omvärld skickat in. Dessa anvisningar kan mycket väl få innebära att Searle måste genomföra ett långvarigt och komplext arbete i flera led; det väsentliga är att de är fullständigt utförda, så att Searle kan klara jobbet utan att förstå språket.

Antag också att bokverket är så skrivet att, när en kinesisktalande testperson skickar in meddelanden skrivna på kinesiska och Searle genom att mekaniskt följa bokverkets anvisningar skickar ut svarsmeddelanden, testpersonen bedömer att det finns någon eller något i rummet som förstår kinesiska. Detta till trots, menar Searle, finns det inget *medvetande* i rummet som förstår kinesiska. Searle själv gör det inte, medan vare sig bokverket, lapparna eller någon annan fysisk del av rummet besitter medvetande.

Experimentet är, åter enligt Searle, ekvivalent med den tänkta situationen att vi har en dator som tack vare ett listigt skrivet program verkar förstå kinesiska. Bokverket motsvarar programmet, lapparna programmets data och Searle processorn. Inte i någondera fallet uppstår något medvetande till följd av den pågående symbolhanteringen, och alltså, sluter Searle, är stark AI inte korrekt.

Turingtestet

Alan Turing designade sitt test bland annat just för att undvika problem av den art som den här omnämnda debatten behandlar. I stället borde man, enligt Turing, anta ståndpunkter som i mycket liknar vad Searle kallar svag AI. I korthet kan Turings test beskrivas enligt det följande.

Anordna en kommunikationslänk som bara tillåter överföring av skrivna meddelanden, till exempel via två datorterminaler, placerade i var sitt rum. I den första versionen låter vi en kvinna och en man omväxlande sitta vid den ena terminalen, varpå testpersonen vid den andra terminalen ges möjlighet att föra en konversation via kommunikationslänken, i avsikt att utröna huruvida han talar med en kvinna eller en man. Både kvinnan och mannen har att försöka övertyga testpersonen om att de är kvinnan; mannen måste alltså försöka imitera en kvinna för att lyckas.

I den andra versionen har vi i stället en människa och en dator. Datorns uppgift är, på liknande sätt som tidigare, att försöka övertyga testpersonen om att den är människan. Lyckas den har den klarat Turingtestet.

Som synes är den beskrivning av Turingtestet som Searle tillskriver stark AI både felaktig och missvisande. Turing anser inte att testet stöder stark AI eller bevisar att datorn har ett medvetande. Tvärtom är hans avsikt att visa att man bör försöka gå runt de intrikata problemen; det finns intressanta resultat att nå i alla fall.

> Det kommer att underlätta för läsaren om jag förklarar min egen åsikt i frågan. Låt oss först ta den mer preciserade frågeställningen. Jag tror att det om cirka femtio år kommer att vara möjligt att programmera datorer... så att de kan spela imitationsspelet så väl att en genomsnittlig frågare inte har mer än 70 procents chans att göra rätt bedömning efter fem minuters utfrågning. Den

tidigare frågan 'Kan maskiner tänka?' anser jag vara alltför meningslös för att förtjäna diskussion, (1950, s 422).

Om mannen i Turingtestets första version skulle lyckas lura testpersonen bevisar detta naturligtvis inte att han är kvinnan, och på samma sätt: om ett program skulle klara testet bevisar detta vare sig att programmet är en människa eller att det har ett medvetande. Däremot kan vi tänkas vara intresserade av att använda programmets goda förmåga att härma en människa.

Searle klargör inte någonstans huruvida han själv omfattar den felaktiga beskrivning av Turingtestet han givit.

Searles axiomatiska argumentering.

Det av Searle föreslagna tankeexperimentet är endast ena halvan av den argumentation som framförs. För att motbevisa den starka AIns tes anför Searle även ett formellt resonemang, bestående av fyra axiom och fyra slutsatser. Så här ser axiomen ut:

1. Datorprogram är formella, (syntaktiska).
2. Mänskliga medvetanden har mentalt innehåll, (semantik).
3. Syntax i sig är varken grundläggande eller tillräckligt för semantik.
4. Hjärnor orsakar medvetanden.

Ur dessa premisser följer, enligt Searle, följande slutsatser:

1. Datorprogram är varken grundläggande eller tillräckliga för att orsaka medvetande.
2. Varje annat system som kan orsaka medvetande måste ha kausala förmågor (åtminstone) jämförbara med hjärnors förmåga.
3. Varje artefakt som producerar mentala fenomen, varje konstgjord hjärna, måste kunna duplicera den mänskliga hjärnans specifika kausala förmågor, och den kan inte göra detta enbart genom att exekvera ett formellt program.
4. Det sätt varpå den mänskliga hjärnan producerar mentala fenomen kan inte vara enbart genom exekverandet av ett program.

Vid första påseende kan axiom ett verka okontroversiellt. Searle menar att ett datorprogram är rent formellt till sin natur; dess implementation är i alla avseenden oväsentlig.

> Symboler och program är rent abstrakta begrepp: de definieras inte av några väsentliga fysiska egenskaper och kan implementeras i godtyckligt fysiskt medium, (1990, s 21).

Detta gör att det verkar rimligt att tolka det första axiomet som så, att det betyder att datorprogram är *rent* formella.

Det andra axiomet är svårtolkat, eftersom Searle normalt inte antyder vilket medvetandesbegrepp han använder sig av. Av (1984) framgår dock att han med *mentalt innehåll* avser ett subjektivt medvetande, alltså inre upplevelser som sinnesintryck, tankar, känslor, minne, etc. En företrädare för en radikal materialism eller funktionalism skulle troligtvis inte acceptera det andra axiomet.

Vidare är Searles användning av begreppet semantik förvirrande. Att döma av den kommentar som ges till det tredje axiomet är avsikten antagligen endast att säga att det är skillnad på fenomen med och utan mening, och att semantik i det andra axiomet mer eller mindre kan översättas med mentala tillstånd.

> Poängen är distinktionen mellan formella element, som inte har någon inneboende mening, och sådana fenomen som har ett inre innehåll, (1990, s 21).

I det tredje axiomet vill Searle antyda att syntax definitionsmässigt inte kan skapa semantik. Huruvida detta är sant i formell mening skall vi inte gå in på här, utan endast konstatera att i Searles tolkning, där semantik är ekvivalent med medvetande, gäller ingen sådan definitionsmässig sanning.

I det fjärde axiomet anger Searle en klassisk ståndpunkt inom kropp- och självproblematiken, nämligen att hjärnan *orsakar* medvetande. Denna ståndpunkt måste dock anses som allt annat än klar.

2. *Felaktigheter i Searles resonemang*

Searle hävdar att de ovan angivna premisserna är evident sanna och att slutsatserna logiskt följer ur dessa premisser. Vi skall emellertid se att dels undgår inte en enda av premisserna svårbemött kritik, dels följer inte självklart, på grund av Searles mångtydiga användning av flera begrepp, slutsatserna ur premisserna. Det är faktiskt anmärkningsvärt hur låg kvalitet Searles argumentation har.

Medvetande kontra intelligens

Den första svagheten i Searles resonemang är hans inriktning mot *medvetande*. AI-forskningens mål har så gott som alltid sagts vara att åstadkomma *intelligens*. Att detta skulle vara liktydigt med att åstadkomma medvetande är långt ifrån uppenbart; snarare verkar det rimligt att intelligens, eller åtminstone intelligent beteende, inte nödvändigtvis kräver ett bakomliggande medvetande.

Stark AI i Searles version har antagligen få försvarare som vet vad de talar om. Detta erkänner Searle själv i (1990, s 20). I artikeln konstaterar Searle att inte ens Churchlands förstått den starka AIns ståndpunkt. Min slutsats är att Searle till stor del skapat ett spöke som han sedan försöker bekämpa.

Redan denna inledande kritik anser jag vara förödande för Searles resonemang. Stark AI handlar om *intelligens*, inte om *medvetande*.

Det kinesiska rummet

Mycken kritik har riktats mot det tänkta experiment som Searle anför: det går inte att skriva ett bokverk med de angivna egenskaperna, Searle skulle vara för långsam för att klara uppgiften, han skulle inte kunna undvika att lära sig kinesiska, etc. All denna kritik är befogad, och kan sammanfattas i det enkla konstaterandet att i praktiken är hela förutsättningen absurd. Searle kan inte rimligtvis ersätta en dator på det sätt som beskrivs, han är alldeles för långsam, han skulle inte kunna följa de komplexa instruktioner som krävs, han skulle aldrig kunna överblicka en tillräckligt stor mängd data och han skulle oundvikligen begå alltför många fel; allt detta skulle med största säkerhet ligga långt bortom varje människas förmåga. Att, som Searle i ett kontraargument föreslår, anta att han skulle kunna memorera allt och klara sig helt utan yttre hjälpmedel är (om möjligt) än orimligare. Därtill kommer att tankeexperimentet har ytterligare en stor brist, nämligen att "processorn," det vill säga Searle, har ett eget medvetande. Detta medvetande förväxlar Searle (mer eller mindre medvetet) med det eventuella medvetande varom debatten rör sig. Hela experimentet är alltså synnerligen bristfälligt.

I stället för att fullfölja dessa argument skall vi emellertid konstatera att Searles rum *i princip* är en rimlig ekvivalent till ett datorsystem, vilket som helst. I samma ögonblick kan vi emellertid också konstatera att därmed försvinner exemplets specifika intresse.

Den principiella möjligheten för ett datorsystem att åstadkomma medvetande är troligtvis lika stor, oavsett om det rör sig om en dator i halvledarmaterial eller Searles föreslagna arrangemang. Vi har lika god eller dålig intuition i båda fallen och alltså kan vi lämna exemplet därhän. Dess praktiska orimlighet gör att vi lämnar det med glädje.

Searles medvetandesfilosofi

Innan vi tar oss an Searles axiomatiska argumentering skall vi undersöka hans teori om medvetandets natur. Denna fråga har notoriskt undvikits i vulgärdebatten.

Searle kallar ståndpunkten att allt som existerar i världen är fysiska partiklar för *naiv fysikalism*, och på samma sätt låter han *naiv mentalism* beteckna ståndpunkten att mentala fenomen verkligen existerar: att det finns mentala tillstånd, några är medvetna, många äger intentionalitet, alla är subjektiva och många har kausala funktioner i den fysiska världen.

Dessa två teorier är korrekta och förenliga, hävdar Searle. De mentala tillstånden är orsakade av processer i hjärnan.

> Mentala fenomen, medvetna såväl som omedvetna, visuella som audiella, smärta, kittlingar, klåda, tankar, ja hela vårt mentala liv, orsakas av processer i hjärnan, (1984, s 18).

Samtidigt som mentala tillstånd är orsakade av processer i hjärnan är de också egenskaper hos densamma.

> Smärta och andra mentala fenomen är egenskaper hos hjärnan (och kanske resten av det centrala nervsystemet), (1984, s 19).

Searles teori uppvisar stora likheter med en ontologisk teori som beskriver en vardaglig syn på medvetandesproblemet. Vi skall, obekymrade om att termen också används på annat sätt, kalla denna teori för *naiv realism.*

1. Det existerar en värld av fysiska föremål.
2. Det existerar subjektiva, mentala tillstånd.
3. Vi äger kunskap om den fysiska världen genom våra sinnesintryck.
4. Våra upplevelser av fysiska föremål är orsakade av dessa fysiska föremål.

Searle menar att en sådan teori inte behöver innebära dualism. De mentala tillstånden kan uppstå som makroegenskaper ur neurala processer i hjärnan, på samma sätt som till exempel en bordsytas hårdhet uppstår ur atomära gitterstrukturer på mikronivå. Vi noterar emellertid att detta är högst kontroversiellt. Den naiva realismen må vara intuitivt tilltalande, men inte någon av dess satser undgår allvarlig kritik, och teorin har få försvarare. Om man inte accepterar den faller Searles argumentation platt till marken.

Axiomens giltighet

Vi skall nu övergå till att analysera Searles axiomatiska argumentation. Vad det gäller tolkningen av de olika axiomen skall vi komma ihåg att Searle förutsätter den naiva realismens syn på världen. Det finns som bekant flera konkurrerande medvetandeteorier, men inte i någon tolkning är resonemanget hållbart. Låt oss se närmare på Searles egen naiva realism.

De axiom och slutsatser som Searle anför i sin artikel är ganska vaga och oklara. Låt oss därför göra följande tolkning.

1. Exekverande datorprogram är rent formella.
2. Medvetanden innehåller mentala tillstånd.
3. Rent formella system kan inte åstadkomma mentala tillstånd.
4. Hjärnor orsakar medvetanden.

Det första axiomets riktighet är avhängig definitionen av begreppet datorprogram. Ser vi ett program som enbart en instruktionssekvens är axiomet korrekt, men en sådan tolkning är knappast rimlig. En förnuftig definition av ett datorprogram måste omfatta en processorbeskrivning, samt en instruktionssekvens som *exekveras* på en *fysisk* processor. Med en sådan tolkning beror axiomets riktighet på om det är avsett att betyda att vissa representationer av datorprogram besitter syntaktiska egenskaper (vilket är korrekt) eller att datorprogram *enbart* besitter syntaktiska egenskaper (vilket är fel). Enligt den tidigare definitionen är ett program visserligen formellt, men inte enbart syntaktiskt, och detta är den enda rimliga ståndpunkten i en debatt. Ingen representant för stark AI skulle rimligen vilja påstå att en instruktionsföljd i sig (till exempel en programlistning) åstadkommer medvetande. Det är kombinationen av instruktioner och exekverande processor som kommer i fråga. Alltså kan vi konstatera

att Searles resonemang antingen inte gäller i det fall vi faktiskt exekverar ett program på en dator, eller så är det första axiomet felaktigt. Denna invändning är *helt* förödande för Searles resonemang.

Axiom två är avhängigt vilken medvetandeteori man omfattar. I den naiva realismen är det rimligtvis korrekt. Mentala tillstånd äger existens och är på något okänt men fungerande sätt kopplade till de olika processer som försiggår i den mänskliga hjärnan.

Det tredje axiomet kan, på grund av Searles oklara användning av begreppen syntax och semantik tolkas på två olika sätt. Dels kan det antas betyda ungefär att formell syntax är något annat än formell semantik, att syntaxregler inte räcker för att definiera en semantik. I denna tolkning är axiomet logiskt till sin natur. Antagligen är det inte korrekt, men det behöver vi inte diskutera här. (Frågan hänger på existensen av syntaxregler som samtidigt utgör en semantik; och visst bör det finnas sådana urartade fall.) Vi kan nöja oss med att konstatera att om axiomet ges denna tolkning går det inte att dra de slutsatser som Searle önskar, helt enkelt därför att de olika axiomen talar om disparata begrepp.

I den andra tolkningen är axiom tre i stället avsett att betyda att ren symbolbehandling inte kan åstadkomma medvetande. Denna tolkning är rimligtvis den Searle avsett. Här kan vi med ens konstatera att det inte, som Searle vill antyda, finns någon definitionsmässig motivering att stödja sig på, utan i stället innebär axiomet helt enkelt att Searle förutsätter den slutsats han vill bevisa. Accepterar man inte redan denna, faller det tredje axiomet. Även detta anser jag vara *helt* förödande för Searles resonemang.

Dubbeltydigheten i axiom tre utnyttjas av Searle på ett i högsta grad oacceptabelt sätt.

> På en nivå är denna princip definitionsmässigt sann. Man kan naturligtvis definiera begreppen syntax och semantik på andra sätt. Poängen är att det finns en skillnad mellan formella element, som inte har någon inneboende mening, och sådana fenomen som har inre innehåll, (1990, s 21).

Detta dubbla budskap är i bästa fall en monumental begreppsförvirring, och i värsta fall ett trick, avsett att ge Searles ställningstagande i en avgörande debattfråga en skenbar aura av logisk

nödvändighet. Det sistnämnda stöds tyvärr av att Searle senare hävdar att axiom tre är en logisk sanning.

Det fjärde axiomet bygger helt på att man accepterar den naiva realismen och därmed den lösning på kropp- och själproblemet som brukar kallas *interaktionism,* det vill säga att medvetandet orsakas av fysiska förlopp (och rimligtvis vice versa). Detta är dock inte okontroversiellt. Bland problemen med teorin i fråga brukar bland annat anges följande: kausalitetsrelationer förekommer normalt mellan *fysiska* händelser och det är inte oproblematiskt att anta att de även kan existera mellan fysiska och mentala händelser; en kausalitetsrelation från mentalt till fysiskt skulle dessutom komma att strida mot principen om energins bevarande, etc. Vi nöjer oss med att konstatera att Searle utan närmare motivering väljer en (långt ifrån problemfri) filosofisk teori av många. Accepterar man inte detta val, faller också det fjärde axiomet.

En välvilligare tolkning kan göra det första axiomet mindre kontroversiellt. Antar vi att det skall tolkas som att datorprogram är formellt beskrivbara, blir det plötsligt rimligare.

1. Exekverande datorprogram är formellt beskrivbara.
2. Medvetanden innehåller mentala tillstånd.
3. Formellt beskrivbara system kan inte åstadkomma mentala tillstånd.
4. Hjärnor orsakar medvetanden.

Här måste emellertid axiom tre anses betyda att inget formellt beskrivbart system kan åstadkomma medvetande, en absurd konsekvens som Searle knappast kan önska sig. I så fall försvinner ju allt hopp om att någon gång kunna beskriva ens den mänskliga hjärnan på ett vetenskapligt sätt. Att döma av artikeln i övrigt är det emellertid inte troligt att Searle har avsett denna välvilligare tolkning av det första axiomet.

Tillåter man sig att tänja tolkningen ännu ett steg kan det tredje axiomet anses betyda att formellt beskrivbara system inte kan åstadkomma mentala tillstånd genom sin egenskap att vara just formellt beskrivbara. Searle hävdar också att det är hjärnans specifika kausala egenskaper som orsakar medvetande. Om så skulle vara fallet (vilket är en helt öppen fråga) följer i vilket fall som helst inte att medvetande inte skulle kunna orsakas även på andra sätt, till exempel av datorer. Om detta säger Searle följande.

> Finns det något logiskt tvingande skäl för att de inte kunde avge medvetande? Nej. Vetenskapligt sett är det helt uteslutet, men detta är inget som det kinesiska rummet är avsett att bevisa,... Det viktiga är program, och program är rent formella, (1990, s 25).

Searle förnekar alltså inte att datorer i sin egenskap av fysiska föremål skulle kunna avge medvetande, eller, ekvivalent uttryckt, det är alltså enligt Searle fullt möjligt att formellt beskrivbara system i sin egenskap av att vara fysiska skulle kunna åstadkomma medvetande. Att hävda något annat vore knappast rimligt och detta är säkerligen inte den tolkning som Searle avsett.

Låt oss i förbifarten observera hur ohederligt Searle argumenterar. Lika lite som någon annan vet han något om vilka fysiska föremål som kan åstadkomma medvetande, men likväl framställer han möjligheten av att datorer skulle ha denna egenskap som "vetenskapligt sett helt utesluten."

Slutsatsernas giltighet

Vi har nu sett att axiomen inte är problemfria. Men även i det fall att vi antar deras riktighet kan mångtydigheten i de använda begreppen i vissa tolkningar komma att leda till att slutsatserna inte är giltiga. Låt oss återupprepa dessa och kommentera dem.

1. Datorprogram är varken grundläggande eller tillräckliga för att orsaka medvetande.
2. Varje annat system som kan orsaka medvetande måste ha kausala förmågor (åtminstone) jämförbara med hjärnors förmåga.
3. Varje artefakt som producerar mentala fenomen, varje konstgjord hjärna, måste kunna duplicera den mänskliga hjärnans specifika kausala förmågor, och den kan inte göra detta enbart genom att exekvera ett formellt program.
4. Det sätt varpå den mänskliga hjärnan producerar mentala fenomen kan inte vara enbart genom exekverandet av ett program.

Den första slutsatsen skall, enligt Searle, följa ur axiom ett, två och tre. Detta förutsätter dels att axiom ett är betyder att datorprogram är *enbart* syntaktiska, vilket gör resonemanget så gott som ointressant, dels att begreppet semantik enligt axiom två och tre är detsamma. I axiom två betyder semantik ungefär detsamma som mentalt innehåll,

medan det i axiom tre eventuellt kan vara avsett att ha en språkanalytisk innebörd, (den normala). Den första slutsatsen är i så fall inte ens formellt riktig.

Slutsats två verkar, i en rimligt godvillig tolkning, korrekt. Det är väl närmast trivialt sant att varje system som kan orsaka medvetande måste ha förmågor som är i vid mening ekvivalenta med hjärnans. Möjligtvis skulle man kunna fälla den kritiken att slutsatsen är så gott som innehållslös.

I slutsats tre försöker Searle skärpa detta resultat, dock utan att lyckas. Av förutsättningen att hjärnor orsakar medvetande följer givetvis inte att medvetande inte kan orsakas av något annat än hjärnor.

Den fjärde slutsatsen avhänger den första och är på så sätt inte heller klar. Därtill kan den verka något besynnerlig: om man som Searle inte vill kalla den mänskliga hjärnan för en dator, följer trivialt att den mänskliga hjärnan inte producerar medvetande genom att köra ett datorprogram. Searle visar i denna tolkning en enkel sats på ett onödigt komplicerat vis. Detta är emellertid inget fel utan bara en klumpighet.

3. Ytterligare kritik

Det faktum att Searle noggrant begränsar sin argumentation hindrar honom inte från att sedan uttala sig betydligt mer generellt, och påstå att datorer inte kan åstadkomma medvetande, vilket man väl måste tolka som att det här rör sig om vanliga, fysiska datorer och subjektivt medvetande. Axiom ett är inte giltigt i detta fall, axiom två och fyra avhänger icke-triviala filosofiska ställningstaganden och axiom tre förutsätter helt enkelt det som skall bevisas. Följaktligen misslyckas Searle i sitt formella resonemang. Han försöker emellertid en ytterligare form av argument, i vilket exemplet med det kinesiska rummet spelar en viktig roll.

I rummet finns uppenbarligen, enligt Searle, inte något medvetande som förstår kinesiska, fastän Searle med bokverkets hjälp avger för ett Turingtest acceptabla svar. Denna enkla idé går igen i det tredje axiomet: något så formellt som symbolhantering kan inte rimligtvis orsaka medvetande. Detta anser sig Searle fullkomligt säker på, det är äger vardaglig evidens och betraktas av honom till och med som en logisk sanning. Låt oss nu undersöka denna Searles

intuitiva inställning till vad som kan och inte kan tänkas vara medvetet.

Vanligt folk och zombies

Dennett har angivit ett enkelt argument som visar hur svårt det är att uttala sig om ett medvetandes existens eller icke-existens, (1982, s 174). Låt oss anta, säger Dennett, att det finns två sorters människor. Båda har fungerande hjärnor, vilka genom livet tar emot intryck från omvärlden och på något för oss okänt men fungerande sätt lagrar, behandlar och omvandlar dessa intryck, via interna representationer, till olika aktioner. Det som skiljer de två sorternas människor är att hos den ena gruppen åtföljs hjärnaktiviteten av ett subjektivt medvetande, hos den andra saknas detta. Den ena sortens människor (vanligt folk) har medvetande, den andra (zombies) har det inte, och detta är *den enda* skillnaden mellan dem.

Vi kan, enligt Dennett, inte genom någon sorts observationer över huvud taget skilja en vanlig människa från en zombie; det enda medvetande vi kan observera är vårt eget. Världen kan vara full av zombies utan att vi vet om det eller kan veta om det. Inte ens när det gäller människor kan vi alltså veta vem som har ett subjektivt medvetande, utan måste nöja oss med att ställa upp en hypotes, vilken kan stärkas eller försvagas med stöd av observationer av människans beteende, något som mycket liknar Turingtestets idé. Dennetts enkla och, som jag anser, korrekta analys, visar att vi inte kan veta huruvida medvetande kan uppstå ur ett visst fenomen, i det aktuella fallet symbolhantering. Det bör tilläggas att Searle helt och fullt erkänner medvetandets subjektiva karaktär.

> Om vetenskapen försöker beskriva hur världen är kommer en av beskrivningens ingredienser att vara att mentala tillstånd är subjektiva, eftersom det är ett enkelt faktum om den biologiska evolutionen att den har åstadkommit vissa sorters biologiska system, nämligen människors och vissa djurs hjärnor, som har subjektiva egenskaper. Mitt nuvarande medvetandestillstånd är en egenskap hos min hjärna, men dess medvetna egenskaper är tillgängliga för mig på ett sätt som de inte är för dig, (1984, s 25).

För den som accepterar existensen av subjektivt medvetande och att denna subjektivitet innebär att var och en endast kan komma i

kontakt med, observera, sitt eget medvetande, bör den rimliga ståndpunkten vara som följer. Vad avser det *egna* medvetandet är den naiva mentalismen en tillfredsställande teori, men vad gäller *andra* medvetanden måste vi nöja oss med en funktionalistisk ståndpunkt. Vi kan tro att andra är medvetna, men det enda sättet att stödja ett sådant antagande är genom iakttagelser av funktion och yttre beteende.

Searle anser sig av någon anledning kunna avgöra huruvida andra medvetanden existerar utan att förfalla till funktionalistiska betraktelsesätt, samtidigt som han accepterar medvetandets subjektiva karaktär. Då han emellertid inte gör någon anstalt att bevisa möjligheten av detta, skall vi temporärt lämna denna suspekta åsikt åt sitt öde.

Vad åstadkommer medvetande?

Vi kan ställa upp flera fysiologiska förklaringar till vad det är hos en hjärna som på något sätt producerar eller utgör en fysisk motsvarighet till medvetande. Bland annat kan nämnas följande:

1. Närvaron av materia över huvud taget, (animism).
2. Närvaron av vissa kemiska eller biologiska substanser.
3. Att vissa fysiska, kemiska eller biologiska reaktioner äger rum.
4. Närvaron av elektriska eller kemiska signaler.
5. Närvaron av ett informations- eller symbolflöde (i godtyckligt medium).
6. Närvaron av komplexa kopplingsstrukturer mellan nervcellerna.
7. Närvaron av parallellt pågående processer.
8. Etc.

Dessa fenomen kan, bland många andra och eventuellt i kombination, vara förklaringen till hjärnans förmåga att producera medvetande. Hur det faktiskt ligger till vet vi i dagens läge inte. Searles grundläggande fel är att han försöker bevisa att *en* möjlig förklaring, informationsflödet som sådant, inte duger. Det tredje axiomet är helt enkelt en alternativ formulering av detta ställningstagande, och längre än så når inte Searles argument.

Searles "bevis" för det tredje axiomet

I debatten har huvudsakligen det tredje axiomet angripits, framför allt i det som brukar kallas "the systems reply." Searle själv förstår

måhända inte kinesiska, men systemet bestående av Searle, rum, lappar och bokverk kan mycket väl tänkas ge upphov till ett annat medvetande, som förstår kinesiska.

Bortsett från tankeexemplets praktiska absurditet instämmer vi helt och fullt i detta försvar. Det innebär ett ifrågasättande av Searles tredje axiom. Detta har länge varit helt utan motivering, men i Scientific American försöker Searle faktiskt bevisa dess riktighet.

Searle hävdar att axiomet är logiskt nödvändigt, även då begreppet syntax tolkas som symbolhantering och semantik som medvetandesinnehåll. Detta kan tyckas överraskande, men Searle anger ett "motsägelsebevis".

> Som med alla logiska sanningar ser man enkelt att den är sann, eftersom man får inkonsistenser om man försöker anta motsatsen, (1990, s 25).

Låt oss anta, säger Searle, att det kinesiska rummet verkligen åstadkommer ett medvetande, M1, som är en följd av Searles mekaniska användande av bokverk och lappar. Detta medvetande är ett annat än Searles eget, M2, och förstår kinesiska och den konversation som försiggår. Så långt är allt väl, säger Searle, men antag nu att Searle själv, i sitt eget medvetande, beslutar sig för att tolka tecknen som om de beskrev dragen i ett schackparti. En ytterligare person tolkar i stället, i sitt medvetande, M3, symbolerna som förutsägelser om aktiemarknaden. Searle frågar nu vilken semantik som rummet avger, det vill säga hur tecknen tolkas i rummets medvetande, M1. Denna fråga har enligt Searle inget klart svar, utan här föreligger en logisk motsägelse.

> Vilken semantik avger systemet nu? Avger det en kinesisk semantik, en schacksemantik eller båda samtidigt? ... Det finns ingen gräns för antalet semantiska tolkningar som kan ges symbolerna, eftersom dessa, än en gång, är rent formella. De har ingen inre semantik, (1990, s 25).

Först konstaterar vi att det verkar ganska orimligt att, som Searle, tänka sig att längre följder av tecken skulle kunna ges oberoende men konsistenta tolkningar. Låt oss emellertid bortse från detta problem och anta att det rör sig om kortare teckensekvenser.

I det beskrivna fallet föreligger ingen motsägelse och Searles

resonemang är grovt felaktigt. De tre inblandade medvetandena, M1, M2 och M3, har *var sin* tolkning av symbolerna. Searles initiala antagande om att M1 tolkar tecknen som kinesiska är fortfarande lika giltigt (eller ogiltigt). Att Searle själv, M2, tolkar symbolerna på ett sätt och den tredje personen, M3, på ett annat omöjliggör inte att rummet kan ha ett medvetande.

Searle tycks tro att i och med att han i sitt eget medvetande, M2, börjar tolka tecknen på ett visst sätt, denna tolkning skulle emanera ut till rummets medvetande, M1, eller på något sätt ersätta rummets tolkning, men naturligtvis finns det ingen anledning att tro att så måste vara fallet.

Återigen är det anmärkningsvärt att en filosof som Searle gör sig skyldig till så synnerligen grova misstag. Om resonemanget vore giltigt hade Searles egen tolkning av symbolerna omöjliggjort inte bara rummets medvetande, utan även den tredje personens medvetande och alla andra tänkbara, tolkande medvetanden över huvud taget.

4. Sammanfattning

Enligt det ovanstående kan vi konstatera att Searles argumentation lämnar mycket övrigt att önska. Sektion 3 illustrerar att Searle inte kan *veta* huruvida symbolhantering är tillräckligt för att åstadkomma medvetande eller ej. De axiom han anger som evidenta är inget annat än åsikter vars riktighet inte låter sig testas eller undersökas. De tekniska delarna av argumentationen har så stora brister att resonemanget i sin helhet knappast kan räddas.

Först och främst rör Searles argumentation begreppet *medvetande*. Det är på inget sätt klart att resultaten också gäller för *intelligens*. Vidare är det angivna tankeexperimentet inte rimligt verklighetstroget, utan är i alla avseenden sämre, oklarare och mindre intuitivt än att direkt resonera om en dator med program.

Searle förutsätter tyst sin egen medvetandeteori, *naiv mentalism*. Den är inte problemfri; rimligtvis kan Searle inte direkt tillämpa den på andra medvetanden än sitt eget. Omfattar man inte denna teori faller hans argument platt till marken.

I sitt första axiom överdriver Searle ett datorprograms formella egenskaper på ett felaktigt vis. Faktiskt finns det skäl att misstänka att han helt enkelt inte förstått vad ett program är. I det tredje

axiomet förutsätter Searle helt enkelt det som skall bevisas, och det motsägelsebevis han anger som stöd är felaktigt på ett anmärkningsvärt grovt sätt. Vidare använder Searle begreppen syntax och semantik på ett högst oacceptabelt sätt; det är närmast att betrakta som en sorts trick eller argumentationsknep.

Oavsett vilken medvetandeteori man omfattar och vilken tolkning man vill ge Searles axiom och slutsatser är resonemanget felaktigt. Det går alltså inte att motbevisa stark AI på det sätt Searle försöker. Accepterar man en naiv mentalism, följer rimligtvis att varken stark AI eller Searle har rätt. Den starka AIns företrädare kan inte *veta* att ett symbolbehandlande system verkligen besitter subjektivt medvetande, och Searle kan inte *veta* att det inte gör det.

Ända sedan Searle presenterade sina argument om det kinesiska rummet år 1980 har debatten rasat, framför allt inom AI-kretsar i USA. Hela tiden förbises så gott som alla grundläggande teorier och resultat som finns i de flesta grundkurser i teoretisk filosofi. Kontrahenterna anger till exempel sällan eller aldrig sina åsikter om medvetandets problem eller kropp- och själproblemet. Det naturliga vore givetvis att man började med att klargöra sina ställningstaganden vad gäller vanligt, mänskligt medvetande, och först när detta klarats av, kastade sig över det artificiella medvetandets problem. Debatten visar emellertid att inte ens välkända, amerikanska filosofer orkar vara så metodiska.

Litteratur

DENNETT, DC, *Brainstorms,* Bradford Books, Montgomery, Vermont, 1978.

DENNETT, DC, "How to Study Human Consciousness Empirically or Nothing Comes to Mind," in Hintikka, J, (utg.), *Synthese,* Vol 53, 1982.

SEARLE, JR, "Is the Brain's Mind a Computer Program?" *Scientific American,* Januari, 1990.

SEARLE, JR, "Minds, Brains, and Programs," in Hofstadter, DR and DENNETT, DC, *The Mind's I,* Bantam Books, Toronto, 1981.

SEARLE, JR, *Minds, Brains, and Science,* Harvard University Press, Cambridge, Massachusetts, 1984.

TURING, AM, "Computing Machinery and Intelligence," *Mind,* Vol 59, 1950.