Frank Lorentzon

Intentionala maskiner

När John Searle publicerade uppsatsen "Minds, Brains and Programs" (*The Behavioral and Brain Sciences*, vol 3, 1980, ss 417-457) ville han visa att det är omöjligt att programmera ett datorsystem, oavsett komplexitetsgrad, till att ha mentala tillstånd. Datorn kommer med nödvändighet att sakna något mycket väsentligt hos det mentala, det vi brukar kalla dess intentionala karaktär. Den saknar vad Searle kallar "inneboende" eller "ursprunglig" intentionalitet, och då detta är en nödvändig förutsättning för att ha äkta mentala tillstånd så kan en dator aldrig ha sådana. Searle vill med ett tanke-experiment, det så kallade "kinesiska rummet", visa att det finns en principiell och oöverstiglig skillnad mellan människa och dator i det här avseendet. Jag tänker referera Searles argument, och därefter presentera en alternativ argumentering enligt vilket det tvärtom verkar rimligt att anta att det är principiellt möjligt att skapa system som har mentala tillstånd i samma bemärkelse som människor har sådana.

1. Det kinesiska rummet

Searle hävdar att det är på grund av kausala egenskaper hos hjärnan som människor och djur hyser medvetandetillstånd av intentional karaktär. Det är, påstår Searle, ett empiriskt faktum att vissa hjärnprocesser är tillräckliga för att producera intentionalitet, men att instantiera ett datorprogram räcker inte för att nå liknande resultat, utan det krävs att det i datorn finns kausala krafter motsvarande hjärnans. Därmed vänder sig Searle mot den del av forskningen inom artificiell intelligens (AI) som hävdar att rätt programmerad så inte bara simulerar datorn medvetande, utan har verkligen mentala tillstånd.

Searle har i artikeln "The Myth of the Computer" (*The New York Review of Books*, vol XXXIX, nr 7, april 1982) preciserat vilka teser han tar avstånd ifrån:

- Medvetanden är program av typen självuppdaterande och självdesignande representationssystem. Mentala tillstånd är tillstånd hos ett system (t ex en hjärna

Filosofisk tidskrift 1991 nr 4, 28-36

eller en dator) i vilket ett sådant program implementeras, och mentala processer är beräknings-processer hos detta system.

- Hjärnans neurofysiologi är irrelevant för att förstå vad medvetandet är, då varje system med rätt program har ett medvetande. Det finns alltså ingen nödvändig koppling mellan hjärnan och medvetandet.

- Turingtestet är ett kriterium på det mentala, dvs om ett system på ett så övertygande sätt kan imitera mänsklig tankeverksamhet att en kompetent bedömare inte kan skilja dess yttranden från en människas, så tänker datorn i lika hög grad som människan.

Searles tanke-experiment med det "kinesiska rummet" avser att övertyga oss om att dessa teser är felaktiga. Han föreställer sig att han är instängd i ett rum där det finns en korg full av papperslappar med olika krumelurer, samt regler på engelska för hur dessa skall korreleras med varandra. Då Searle inte kan kinesiska vet han inte att dessa krumelurer egentligen är kinesiska symboler (dvs att de står för något), och inte bara är meningslöst klotter. Han får nu ytterligare två korgar med krumelurer på papperslappar (även de kinesiska symboler), och regler för hur de skall korreleras till de andra, samt för hur han själv skall producera krumelurer (kinesiska symboler) som respons på krumelurerna i den tredje korgen.

Parallellen till hur många av de program fungerar som användes av AI-forskarna kring 1980 är uppenbar. Gemensamt för dessa program är att de försöker simulera någon del av det mänskliga medvetandet. Som exempel nämner Searle ett program av Roger Schank, vars uppgift är att utifrån given bakgrundskunskap om hamburgerbarer, samt små berättelser som utspelas på dessa, kunna svara på frågor även om sådant som inte explicit utsades i berättelserna. Om vi nu jämför det kinesiska rummet med Schanks program, så utgör den första korgen nödvändig bakgrundskunskap (det Schank kallar Script), den andra utgör berättelserna, den tredje frågorna, Searle är processorenheten och symbolerna han producerar utgör svaren på frågorna.

Parallellt med att han manipulerar krumelurer tar Searle även del av berättelser på engelska, samt frågor på dessa som han skall besvara. Om han nu lär sig manipulera symbolerna tillräckligt väl så kommer svaren givna på kinesiska att för en utomstående betraktare att te sig lika goda som svaren givna på engelska. Detta trots att Searle själv ingenting förstår av de symboler han manipulerar och lämnar ifrån sig. De förblir för honom otolkade symboler, och även om rummet som helhet skulle klara Turingtestet, så är det enda han vet att

"squiggle squiggle" följs av "squoggle squoggle". I det engelska fallet däremot vet han vad orden och satserna betyder.

Av detta följer enligt Searle att formell symbolhantering inte utgör ett tillräckligt villkor för förståelse (då hela processen kan simuleras av en människa utan att hon förstår något), samt att Turingtestet inte räcker för att avgöra om ett system tänker eller om det bara simulerar tänkande. (Inte ens om det kinesiska rummet på frågan "Förstår du kinesiska?" (på kinesiska) svarar "Ja" (på kinesiska) bevisar det något, då svaret är producerat utan att vara tolkat, dvs utan att betyda något för manipulatorn i rummet.) Med hjälp av analogin mellan det kinesiska rummet och datorns funktionssätt borde vi enligt Searle inse att inte heller datorn menar något med vad den säger. Den har bara syntax, ingen semantik.

Visserligen kan vi säga om t ex en termostat att den "vet" när den skall slå på värmen, men det beror inte på någon inneboende intentionalitet i termostaten, utan på att vi utsträcker vår egen intentionalitet till våra artefakter, och denna typ av blott *tillskriven* intentionalitet är ointressant i sammanhanget. I sig själv förstår en dator inget, och det hjälper inte att göra systemet mer komplext ; vi tillför bara mer av samma sort, dvs syntax, och förståelsen uteblir.

När AI-forskarna därför försöker simulera den formella strukturen hos hjärnan begår de enligt Searle ett stort misstag då det viktiga hos hjärnan är det faktiska *innehållet* i synapssekvenserna, inte den formella skugga dessa sekvenser kastar. AI tror felaktigt att denna skugga är det väsentliga, men intentionala tillstånd är intentionala på grund av sitt innehåll, inte på grund av sin form, och samma innehåll kan ges mycket skiftande syntaktisk form. Det är enligt Searle på grund av vår specifika biologiska (dvs fysiska och kemiska) struktur som vi är förmögna att producera intentionala tillstånd. Han menar sig se en sorts dualism i AI:s starka betonande av programmens vikt oberoende av deras realiseringar, och vänder sig själv mot alla sorters dualism. Enligt honom kan *bara* maskiner tänka, och då endast en mycket speciell sorts maskiner: mänskliga hjärnor och maskiner med samma kausala krafter som dessa. Intentionaliteten som vi känner den är ett biologiskt fenomen som troligen är lika beroende av sin specifika biokemi som fotosyntesen eller laktationsprocessen är av sina.

2. Searle och kritikerna

Redan i sin uppsats försvarade Searle sin position mot en rad invändningar, varav jag kort skall referera fyra.

(1) *Systemsvaret.* Manipulatorn i det kinesiska rummet utgör bara en del av systemet som helhet, och det är helheten som förstår kinesiska. Som svar på detta argument låter Searle manipulatorn internalisera alla element i systemet som är externa i förhållande till honom själv, dvs han memorerar alla symboler och regler, samt utför alla operationerna i huvudet. Det förändrar ingenting; manipulatorn förstår ingen kinesiska. Det kinesiska subsystemet i honom manipulerar som förut bara otolkade kinesiska symboler.

(2) *Robotsvaret.* Realiseras programmet i en robot med perceptuell och motorisk kontakt med omgivningen, så förstår roboten vad den gör och har mentala tillstånd. Searle svarar genom att helt enkelt låta en del av den kinesiska informationen komma från perceptuella anordningar och en del av responserna vara order till det motoriska systemet hos en jättelik robot i vilken hela rummet befinner sig. Manipulatorn förstår lika lite som förut, och roboten rör sig inte tack vare några intentioner vare sig hos honom eller hos systemet som helhet, utan i enlighet med sin konstruktion och sitt program.

(3) *Hjärnsimulatorsvaret.* Imitera verksamheten på neuronnivå i en hjärna som förstår kinesiska med exempelvis en parallellprocessande maskin. Om detta lyckas så förstår den, likaväl som hjärnan, kinesiska. Searle svarar genom att låta manipulatorn med hjälp av de kinesiska symbolerna kontrollera ett jättelikt system av vattenrör som simulerar hjärnans verksamhet, där varje enskilt vattenrör motsvarar en synapskoppling. Även nu uteblir förståelsen såväl hos manipulatorn som hos systemet som helhet. Det är fel sak som simuleras: hjärnans formella struktur, och inte dess förmåga att producera intentionala tillstånd.

(4) *Kombinationssvaret.* Här förenas de tre övriga svaren: skapa en hjärnlik dator, sätt den i en robot och kör den med ett program som simulerar hjärnans verksamhet på ett sådant sätt att dess beteende är oskiljaktigt från en människas. Då har systemet som helhet mentala tillstånd i samma grad som en människa. Searle avfärdar dock även detta argument genom att påpeka att hur lockande det än kan vara att tillskriva systemet intentionalitet, så är dess beteende resultatet av verksamheten i ett mobilt kinesiskt rum där manipulatorn inte förstår vad den manipulerar. Den avser därmed inte att systemet skall göra vad det gör. Simulering är inte duplicering, hur övertygande den än kan vara.

3. Tanke-experiment som argument

Man kan angripa en ståndpunkt som Searles på många sätt. Ett har varit att konstruera tanke-experiment som väcker andra, motsatta, intuitioner än Searles. Se tex på Haugelands variant (som finns i hans kommentar till Searles uppsats, en av de 27 stycken som trycktes tillsammans med denna i BBS). Medan Searle vädjar till intuitionen att det är hos manipulatorn (en människa precis som vi) som vi måste leta efter förståelse, samt att det är absurt att tänka sig ett *rum* som tänker, så vädjar Haugelands exempel (med en supersnabb liten demon som manipulerar en defekt mänsklig hjärna i enlighet med dess normala

funktion) till intuitionen att det är absurt att kräva att manipulatorn (demonen) skall förstå de processer denne ger upphov till, utan att det är systemet som helhet (hjärnan—dvs det som motsvarar det kinesiska rummet) som har mentala tillstånd.

Själv är jag dock tveksam till värdet av tanke-experiment om de inte underbyggs av mer substantiell argumentering. Jag har en känsla av att man med välvalda tanke-experiment kan "bevisa" nästan vad som helst. Min kritik av Searle kommer därför bara indirekt att bemöta hans tanke-experiment i samband med att jag presenterar ett alternativ till de grundläggande antaganden utifrån vilka han konstruerade detta.

4. Program och formella system

Ett kännetecknande drag för program är att de är formella. Jag skall med utgångspunkt hos John Haugeland (*Artificial Intelligence, The very idea*, Cambridge Mass 1985) närmare precisera vad detta innebär.

I ett *formellt system* är det tecknens form, dvs deras rent syntaktiska egenskaper, som avgör deras funktion i systemet, inte den eventuella mening de kan ha för en utomstående betraktare. Systemet lämnar alltså den semantiska sidan därhän, och är sig självt nog såtillvida att det inte relaterar tecknen till någonting utanför sig självt. Tack vare att formella system är såväl finita som digitala till sin karaktär så kan de *automatiseras*, dvs de kan implementeras i någon hårdvara (tex i en dator—förutsatt att denna är kraftfull nog att härbärgera systemen ifråga) på ett sådant sätt att de kan fås att sköta sig själva.

Men ingenting i det formella systemet kräver någon särskild sorts realisering, och man kan därför tala om *mediaoberoende* hos formella system. På grund av mediaoberoendet får vi den intressanta konsekvensen att om ett formellt system fullständigt simulerar ett annat formellt system, så *är* det detta system, då alla egenskaper och förmågor det ena besitter även kommer att besittas av det andra. Searle menar att även om vi simulerar den formella strukturen hos hjärnans medvetna processer så har vi därigenom ändå inte duplicerat deras karaktär av medvetande. Men *om* det skulle vara så att det mänskliga medvetandet ur någon synvinkel fullständigt kan beskrivas som ett formellt system, en implementering av ett program, så borde detta innebära att vi, om vi lyckas simulera detta system, även har duplicerat det, och att det därför har medvetande i samma grad som vi själva har det. (Att det sedan handlar om artificiellt skapat

medvetande spelar ingen roll. Vi människor är också på sätt och vis artefakter, är designade av vår genetiska struktur—och vi har mentala tillstånd. Om det sedan finns en avsikt bakom designen eller inte saknar betydelse; ett systems tillkomsthistoria är inte grund nog för att bedöma dess mentala kapacitet.)

Searle gör en poäng av att ett program inte är beroende av någon särskild sorts fysisk realisering, och anklagar AI för dualism. Men ett program betraktat som något abstrakt och från sin realisering frigjort kan inte göra anspråk på att ha ett medvetande (vilket vore lika absurt som att påstå att de detaljerade specifikationerna över hur ett flygplan fungerar själva skulle kunna flyga). Det är först när vi implementerat programmet i en för ändamålet lämplig hårdvara som det kan ha mentala tillstånd. Programmet är då att betrakta som en detaljerad specifikation över detta systems interna processer. Searles anklagelse om dualism verkar därmed ogrundad.

5. Mening och representation

Ett program är alltså att betrakta som ett automatiserat formellt system, dvs som en specifikation över de interna processerna i ett system vars verksamhet består i beräkningsprocesser över formellt definierade element. Men *vi* kan avgöra vilken *semantisk* mening de tecken systemet manipulerar har, samt avgöra vilken *avsikten* är med systemets verksamhet; om schackdatorn tex säger vi att den väljer att offra en bonde för att kunna slå tornet, naturligtvis med avsikten att vinna partiet.

Men vad är det vi gör när vi knyter mening till formella processer? Det verkar vara så att det formella systemets verksamhet får mening för oss genom att vi upprättar någon form av relation mellan systemet och världen. Systemets interna tillstånd ses som representationer av externa förhållanden, och vi avgör sedan de enskilda tecknens och processernas mening utifrån deras funktionella roll i denna större struktur av representationer.

Men olika system har olika rika relationer till sin omgivning, och ju fattigare dessa relationer är, desto osäkrare är det vad de inre tillstånden representerar, något som poängteras av Daniel Dennett i *The Intentional Stance* (The MIT Press, Cambridge, Mass 1987, kap 2). Ett enkelt formellt system som saknar rik interaktion med sin omgivning kan tack vare sin generalitet användas till en rad olika ändamål. Samma interna tillstånd kommer då att representera mycket olikartade företeelser hos omgivningen. En termostatliknande anordning kan tex användas som temperaturreglerare—eller för att

styra hastigheten hos ett tåg, utan att anordningen själv märker någon skillnad. Men ju rikare kontakterna med omgivningen blir, desto färre blir tolkningsmöjligheterna, ända tills vi når en nivå där de är så starkt begränsade att om något förändras i systemets omgivning, så kommer systemet sannolikt att notera detta samt korrigera sina tillstånd i enlighet med förändringen. (Detta förutsätter bland annat en viss snabbhet hos systemet ifråga, annars skulle det inte "hinna med" i interaktionen med omgivningen.) Först nu kan systemets inre sägas spegla eller representera sin omgivning i någon egentlig mening.

Att ett sådant system går att beskriva som ett *intentionalt* system beror enligt Dennett just på att en del av dess interna tillstånd går att beskriva som representationer av omvärlden vilka används för att reglera dess beteende i förhållande till denna.

Searle hävdar att det är ett *objektivt* faktum om världen att vissa system har intentionala tillstånd, och att det alltså inte bara är en fråga om *tolkning* utifrån systemets beteende. Dennett däremot menar att när en person hyser åsikter eller har mentala tillstånd så är dessa åsikter och tillstånd visserligen *objektivt* existerande fenomen, men, tillägger han, dessa kan bara urskiljas från den synvinkel där vi antar att personen ifråga är en rationell agent med åsikter, begär och andra mentala tillstånd av intentional karaktär. Att ett system är intentionalt innebär att dess beteende är någorlunda tillförlitligt och utförligt förutsägbart för en betraktare som intar vad Dennett kallar en *intentional synvinkel* i förhållande till systemet. Men vi får inte förledas att tro att vad som helst går att tolka intentionalt, utan det är ett *objektivt existerande faktum* hurvida denna metod tillämpad på något system kommer att ha framgång eller inte. (Molnrörelser tex kan svårligen förutsägas med hjälp av den intentionala strategin, utan här lämpar sig andra beskrivningsnivåer bättre.) Antingen finns de mönster som kan tolkas intentionalt eller så finns de inte. Men, påpekar Dennett, dessa mönster går att beskriva *utan* användning av intentionala termer, om vi i stället för att inta en intentional synvinkel håller oss på en fysikalisk beskrivningsnivå.

Av detta resonomang följer det enligt honom att den mänskliga intentionaliteten, som *inifrån* sett är nog så verklig, *utifrån* sett, på en fysikalisk nivå, kan beskrivas fullständigt utan användning av intentionala termer. Men det skulle ske till en sådan kostnad i uttryckskraft att vi svårligen skulle kunna upptäcka de intentionala mönstren i den stora mängd information som krävs för att beskriva något på den nivån, och vi skulle därför inte uppleva den mänskliga

intentionaliteten just som intentionalitet utifrån den fysikaliska synvinkeln.

Vi borde därför inte vänta oss att finna några intentionala tillstånd om vi studerar ett artificiellt intentionalt system från dess formella sida och ser det som en komplex samling beräkningsprocesser över formellt definierade element. Det är först om vi betraktar det ur den intentionala synvinkeln som dess intentionala karaktär framträder.

6. Semantisk aktivitet och självkorrigering

Vi människor kan själva knyta mening till de tecken vi manipulerar, men frågan är om datorn kan fås att göra detsamma. I sin kommentar till Searles uppsats använder Haugeland termerna *semantisk aktivitet* och *semantisk slöhet* för att peka på en viktig skillnad mellan olika system. De system som vi är beredda att tillskriva någon form av ursprunglig eller inneboende intentionalitet präglas alla av att de är semantiskt aktiva, vilket innebär att de har en ständigt pågående interaktion dels internt mellan olika interna strukturer, och dels externt med sin verksamhetsvärld. Det mänskliga tänkandet är således semantiskt aktivt, medan däremot fenomen med härledd intentionalitet (som tex nedskrivna meningar) är *semantiskt slöa* (en nedskriven mening aktiveras bara som meningsfull när den läses av någon och därmed interagerar med denne läsare). Denna distinktion motsvarar den Dennett gör mellan intelligent och ointelligent informationslagring (*Content and Consciousness*, Routledge & Kegan Paul, London, 1969, s 45ff). Intelligent informationslagring innebär att informationen är information för systemet självt, dvs att systemet har någon användning för informationen. Däremot är information som bara är information för någon extern användare, och alltså inte betyder något för systemet självt, ointelligent lagrad information.

Haugeland menar att förmågan till semantisk aktivitet är just den kausala kraft som Searle efterlyser som tillräcklig för att producera intentionalitet. Det borde därför inte vara principiellt omöjligt att skapa datorer med inneboende intentionalitet förutsatt att dessa ges möjlighet till en tillräckligt omfattande, flexibel och aktiv interaktion med sin omgivning, samt med sina egna interna tillstånd. Denna interna interaktion mellan olika inre strukturer innebär att systemet förmår *referera till sig självt*, och kan påverka sina egna processer i enlighet med den information systemet tar emot, detta för att ständigt kunna anpassa och modifiera sina egna mål och vägarna dit i enlighet med den faktiska situationen.

Sådana *självrefererande* system, som kan ses som formella system, borde principiellt sett vara möjliga att konstruera, även om den nödvändiga komplexitetsgraden överstiger kapaciteten hos dagens hårdvaror. (Mycket av arbetet inom AI på 80-talet, inom den så kallade "nya konnektionismen", har varit inriktat på att konstruera maskiner som arbetar parallellt (precis som den mänskliga hjärnan), i stället för de idag dominerande sekventiella maskinerna, detta för att öka såväl snabbhet som komplexitetsgrad.)

Om det är som Douglas Hofstadter och Daniel Dennett säger (*The Mind's I*, Basic Books, New York, 1981), vilket jag håller för troligt, att förmågan hos de högre nivåerna i ett program att referera till och påverka sina egna lägre nivåer ligger mycket nära "medvetandets kärna", så lär inte en manipulator av den typ som Searle placerar i det kinesiska rummet ha något med detta medvetande att göra. Manipulatorn agerar bara efter på förhand givna regler utan varje möjlighet att påverkas av systemet, och står därmed utanför de processer som kan uppfattas som intentionala. (Detta är i princip en variant av *systemsvaret*.)

Hjärnans verksamhet kan beskrivas som neurala processer eller, utifrån ett annat perspektiv, som avancerade processer för informationsbehandling. Om vi betraktar verksamheten på tex neuronnivå lär vi aldrig kunna urskilja några intentioner eller andra mentala tillstånd annat än i tillskriven bemärkelse. Det är först när vi betraktar systemet ur ett *intentionalt perspektiv*, som vi lär kunna upptäcka några intentioner.

Skillnaden mellan dagens datorsystem och de eventuella system som i framtiden kommer att vara förmögna att hysa mentala tillstånd verkar därmed inte vara en skillnad i kvalitet, utan snarare en skillnad i komplexitetsgrad och förmåga till rik, snabb och flexibel interaktion med omgivningen och sina egna interna tillstånd. En korrekt programmerad dator som står i en tillräckligt rik och aktiv relation till omgivningen borde därför ha ett medvetande i samma grad som vi människor kan sägas ha sådana. Mot detta verkar Searle inte ha givit några hållbara argument.