Tom Eriksson

Alan Turings två teser 1950

1. INLEDNING

Alan Turing (1912–1954) är en av 1900-talets stora tänkare. (Den som vill få en överblick över Turings storhet kan få ut mycket av samlingsverket Teuscher 2010.) Turing gjorde sina största insatser inom logik och matematik. Hans teori om s.k. Turing-maskiner är viktig inom logiken, och bildar en abstrakt grundval för vetenskapen om datorer. Under andra världskriget var han en av de viktigaste som var med och löste tyskarnas ”Enigma-kod” för kryptering, vilket bidrog till krigets utgång. Han borde alltså ha blivit betraktad som krigshjälte i England efter kriget, men så blev det inte. I stället blev han på grund av sin homosexualitet så skandalöst behandlad av landets rättsmaskineri att han tog livet av sig. Detta är en historia som bör berättas. Men nu skall jag ta upp en annan sak.

1950 publicerade han i tidskriften *Mind* en artikel som efterhand kom att bli en av de mest citerade filosofiska artiklarna överhuvudtaget. Den hette ”Computing Machinery and Intelligence”. Själv anser jag att denna artikel är viktig på ett negativt sätt. Den har bidragit till nutida oförståelse för medvetandets problem och det genom att formulera en falsk teori som mest är en tillämpning på datorerna av den s.k. behavioristiska traditionens synsätt. Men genom att vara en särskilt pregnant artikulering av detta synsätt är det en teori som är värd att titta närmare på.

Den *behavioristiska traditionen* inom psykologi och filosofi utmärks av en tendens att i någon bemärkelse identifiera ett psyke med ett visst beteende. Detta har preciserats i olika riktningar vilket gett en uppsättning olika behaviorismer. Traditionens historia började på 1910-talet inom ämnet psykologi. Ett ställningstagande till Turings artikel måste ta itu med den gamla stridsfrågan om behaviorismen, och det skall jag göra här via Turings sätt att formulera problemet.

Man kan alltså göra en ”*minimaltolkning*” av Turings idéer i artikeln: Att det bara är en variant av *behaviorism* tillämpad på datorer. Den tolkningen håller sig närmast texten. Men man kan också göra en ”*maximaltolkning*” som anknyter till nutida idéer – som brukar uppträda med namnet ”*funktionalism*”.

Filosofisk tidskrift 2012 nr 4, 29–42

Jag skall här inte gå in på detaljerna i artikeln utan plocka fram det som jag uppfattar som det väsentliga i den. Jag skall alltså inte lusläsa och analysera Turings formuleringar utan gå in på de *problem* han tar upp. Därför skall jag omformulera och generalisera hans tankar.

2. TURINGS PROBLEM

Turings artikel med dess spekulationer om datorers möjligheter, inledde den forskning som senare har kommit att kallas ”artificiell intelligens” (förkortat AI). Hans grundproblem var frågan ”Kan man genom lämplig programmering skapa ett mänskligt psyke i en dator?”, och hans svar var ”*Ja!*”. Frågan gällde inte vad som var tekniskt möjligt att åstadkomma 1950, utan vad som är *möjligt i princip*. Man kunde uppfatta det som ett problem som kunde lösas i en framtid på sciencefiction-avstånd från, inte bara 1950, utan också från vår nutid.

Jag måste nu förklara vad jag menar med ordet ”psyke”. Jag använder det här som en översättning av engelskans ”mind”, som är ett bra ord. I engelska kan man göra en distinktion mellan ”mind” och ”consciousness” – och det är inte så lätt att göra något motsvarande i svenska språket. Det senare engelska ordet betyder *medvetande* på svenska, men för det förra har vi inget etablerat ord i svenska. Jag skall alltså använda ”psyke”. Det avser allt som vi syftar på med psykologiska termer, rubb som stubb, både sådant som har med medvetandet att göra och sådant som inte har det.

Att jag använder ett särskilt ord, ”psyke”, för allt detta innebär inte att jag betraktar psyket som ett särskilt objekt – eller en ”entitet” som filosoferna säger. Våra psykologiska termer, både i vardagligt språk och i vetenskaplig psykologi, fungerar på mycket olikartade sätt i språket, och anger tillsammans ett begreppsligt heterogent område. Så är emellertid *inte* fallet med begreppet ”medvetande”, vilket jag skall återkomma till.

Man kan undra om Turing verkligen avsåg att ställa sin fråga om *allt* i psyket eller om han begränsade sig till en del av det, nämligen till *intellektet*. I hans överskrift ingick ordet ”intelligence” och i början av texten ställer han frågan om en maskin *kan tänka*. Det låter som om det skulle handla om intellektet. Men hans begrepp om ”tänkande” framstår i den följande texten som så generellt att det verkar överskrida intellektet. Hur som helst: den allmänna frågan ovan, alltså den om psyket i allmänhet, är den *mest intressanta*. Jag skall hålla mig till den.

Bakom Turings fråga ovan finns en annan mer generell fråga: ”Kan man på konstgjord väg i en artefakt, en ”maskin”, skapa ett (ev. mänskligt) psyke?” Vi kan kalla detta ”*syntesproblemet*”.

Man kan hos Turing tolka in två förslag till lösning av syntesproblemet. *Minimaltolkningen* är det *behavioristiska* förslaget: maskinen kon-

strueras så att den *beter sig som om den hade* ett psyke. Då *har* den också – med begreppslig nödvändighet enligt Turing – ett psyke. Psyket sitter så att säga i *beteendet*. Denna tolkning håller sig nära texten.

Så har vi maximaltolkningen, som kan kallas ”*mjukvaraförslaget*”. Innebörden är: det är genom *programmeringen* av en dator som man kan syntetisera ett psyke. Psyket sitter så att säga i detta *program*, i *mjukvaran*. Olika program kan ge upphov till samma beteende, men enligt den här idén finns psyket i något unikt, eller åtminstone i någon snäv mängd av, program. Denna idé som numera kallas (vissa former av) ”funktionalism” eller ”stark AI” finns inte uttryckligen formulerad i texten, men kan kanske läsas in i den.

Om denna tolkning skall vara intressant, bör den kunna tillämpas också på oss människor, och då måste den ge en innebörd åt begreppet ”mjukvara” som kan tillämpas på våra hjärnor.

Båda de här förslagen har till konsekvens att *hårdvaran,* själva de ”hårda” tekniska detaljerna i maskinens uppbyggnad, är *begreppsligt oväsentlig,* vilket i sin tur betyder att psyket av båda uppfattas som en *abstrakt egenskap* hos hårdvaran. Detta är det väsentliga med båda.

Vad är det då som är intressant med Turings fråga och svar? De är ju intressanta för dem som är intresserade av datorer, och det finns många som har det intresset. Men de är också intressanta på ett annat sätt, nämligen som en uppfattning om *oss själva*. Turing framställde en teori om ”psykets natur i allmänhet, och om hans teori är riktig får det stora konsekvenser för vår uppfattning om oss själva.

För att se att denna teori är filosofiskt viktigt skall jag ge ännu en formulering av den problemställningen som är inblandad. Man kan se den som ett *variationsproblem* i filosofen Edmund Husserls mening: hur mycket kan man ändra på, variera, en människa utan att hon förlorar sitt psyke? Husserl varierade en företeelse för att bestämma dess *väsen* och kallade det ”eidetisk variation”. Han ville se hur mycket man kan variera något utan att det förlorar sitt väsen. Så kan man också se det i det här fallet. Turings ståndpunkt om oss människor, torde ha varit att man utan att psykets väsen går förlorat kan variera vår mänskliga hjärnas ”våtvara” till godtycklig annan hårdvara som har en viss abstrakt egenskap gemensam med vår hjärna

Nå, hade Turing rätt? Mitt eget – och många andras – svar är: *Nej!*

3. TURINGS TESER

Vi kan betrakta närmare hur Turing resonerade. Man kan urskilja två teser i hans resonemang. Den ena kan sägas vara tesen (ATT): Det är möjligt *att* en dator genom lämplig programmering kan bete sig så likt

en människa, att den kan *lura* en mänsklig observatör till att tro att den *är* en människa. När en dator lyckas med att luras på detta vis, sägs den ha klarat *Turings test*.

Den andra är tesen (OM): *Om* en dator klarar Turings test, så *har* den ett *mänskligt psyke*. Dvs. om en dator klarar testet så har man löst syntesproblemet.

Man får förgäves leta efter exakt dessa formuleringar i Turings text. Ändå går hela hans artikel ut på att hävda dessa teser. Teserna är ofta dåligt åtskiljda i debatten om Turings artikel, men de leder till mycket olikartade filosofiska problem.

Turing tänkte sig att vissa villkor måste vara uppfyllda för att en dator skulle kunna luras på det där viset: observatören, eller ”testaren” som jag skall säga, får inte ha möjlighet att se datorn i sitt skåp (eller vad det var 1950). Datorn måste vara dold för testaren. Kontakten mellan testare och dator får ske genom utbyte av textmeddelanden, dvs. genom det vi nu kallar *chattande*. En dator klarar alltså Turings test om den kan chatta som en människa.[1]

Man kan lätt se att denna Turings chattversion av problemen inte är något vidare bra. Den framstår som en onödig begränsning av det filosofiska problemet. Testobjektets beteenderepertoar blir för liten och testsituationen är alltför laboratorieartad. Redan den okroppsliga synen på psyket ger problem. Genom Turings chatt-test kan man avgöra t.ex. om testobjektet behärskar ett språk, kan skriva dikter och spela schack, men inte om den kan spela fiol, skala en potatis eller ens gå på två ben.

Än värre blir det om man kommer in på medvetandet, testobjektets upplevelser. Antag att objektet under chattandets gång blir ”het på gröten” som man säger och inleder ett veritabelt sex-chattande med testaren. Om det då visar sig att testobjektet är en dator som chattar på detta vis, kan man då dra slutsatsen att den *har* heta lustar? Ja, enligt OM kan man det.[2] Men det verkar inte rimligt att på fullt allvar anta att datorn har heta lustar. Det enda rimliga är att endast säga att datorn i detta fall *kan chatta som om den hade* det. (Om en filosofisk vän av ordning nu påpekar att jag borde ge ett *argument* för denna åsikt, så kan jag svara att jag längre fram skall göra det.)

[1]Turing la på ytterligare villkor på testet. Han tänkte sig att tre parter skulle vara inblandade, och att testet skulle bli mer formellt bestämt som ett slags spel, ”the imitation game”. Detta finner jag vara överflödigt och jag skall inte gå in på det.

[2]Man kanske kan invända att Turing aldrig avsåg att man skulle kunna dra en sådan slutsats eftersom han skulle ha begränsat sin tes till *intellektet* i någon vid bemärkelse. Jag kan svara, som ovan, att det filosofiska problemet blir mycket mer intressant om tesen OM syftar på ett fullblodspsyke, ty då avser Turings teori att säga mycket mer om oss människor.

När vi möter en människa, möter vi en kropp med mycket omfattande beteenderepertoar i en komplex situation. Den vi möter har ett kroppsspråk, ansiktsuttryck, och en röst. Och vi ingår båda i någon social situation med alla de relationer en sådan kan innefatta. Filosofen Sören Stenlund och sociologen Johan Asplund har, på olika vis, påpekat dessa sociala omständigheter som en kritik mot Turing.[3]

Turings idéer var redan sciencefictionartade 1950 och är det än idag. Ingen dator klarar Turings test idag, dvs. om chattandet får bre ut sig obehindrat. Varför då begränsa science fiction till chattsituationen och inte i stället dra på för fullt? Låt oss göra det senare!

Vi *generaliserar* alltså testet och de båda teserna till en *människoliknande robot* i *sociala situationer*. Roboten har en kropp som ser precis ut som en människas och den har ett avancerat människoliknande beteende. Roboten *klarar* det generaliserade testet om den *lurar* människorna i sociala situationer att den *är* en människa. Vi får tänka oss att den luras på ett avancerat sätt. Den luras även i nära relationer med människor. Och om den t.ex. lägger sig på schäslongen hos en psykoanalytiker och börjar associera fritt, så blir analytikern lurad att där ligger en människa.

Vidare: Enligt den generaliserade tesen ATT* är det genom lämplig programmering av datorn i en datorstyrd robot möjligt *att* få roboten att klara det generaliserade testet. Enligt den generaliserade tesen OM* *har* en datorstyrd robot ett mänskligt psyke *om* den klarar det generaliserade testet.

Det är så här man bör formulera Turings problem. På detta vis ställs de filosofiska problemen på sin spets. Vilka problem? Jo, följande.

För det första hävdar ATT* att en dator kan bete sig precis som en människa, dvs. att en dator kan styra en robot så att denna beter sig precis som en människa. Problemet kring denna tes handlar alltså om hur avancerat beteende som är möjligt för en dator.

Och vi kan se att OM* är en specialisering av följande generella *behavioristiska* tes till datorfallet (P): Om en varelse, levande eller konstgjord, *beter sig som om* den har ett (mänskligt) psyke, så *har* den ett sådant.

Det stora kruxet för behavioristiska teorier är medvetandet. Kruxet för P är alltså följdtesen (M): Om en varelse *beter sig som om* den har ett medvetande, så *har* den ett medvetande.

Den *behavioristiska traditionen* inom psykologi och filosofi utmärks, kan man säga, av att anta P och oftast även P:s omvändning. Många psykologer och filosofer har ända från början gett invändningar mot dessa idéer. Det hade alltså länge funnits en debatt om dessa saker före 1950.

Turing hävdade alltså inget nytt med sin tes OM* (eller OM). Turing

[3] Stenlund 1990, Asplund 2002.

ger emellertid i texten ett slags motivering för dessa teser via en motivering för M, nämligen genom att betrakta M som ”a polite convention” – en *artig konvention*. Turing tänkte sig att följande är byggt på en *konvention*: när vi möter en varelse som beter sig som om den har ett medvetande *bemöter vi den som en medveten varelse* – med allt vad det innebär, dvs. så gör vi i vanliga fall. Det är förvisso sant (och mycket rimligt) att vi i de flesta fall bemöter människor, hundar och katter och många andra djur på detta vis.[4] Frågan är dock om detta kan betraktas som byggt på en *konvention*. Har vi inte goda skäl att anta att dessa varelser *är* medvetna? Alltså ett antagande om fakta och inte en konvention. En del nutida filosofer skulle dock hålla med Turing i denna sak. Så kommer emellertid det stora problemet: Turing hävdar att vi kan (bör?) *utvidga* denna antagna konvention *till en dator* som klarar hans test. Men varför skulle man vara artig mot en dator på detta vis? Det tycks ju hänga på om datorn verkligen *är* medveten eller ej. Det framstår som fullt rimligt att vara artig mot människor och djur men oartig mot datorer. Turing går alls inte in på den saken. En naturlig tolkning är att han förutsatte det som borde ha bevisats, dvs. han förutsatte den behavioristiska ståndpunkten M och tillämpade den på datorerna.

Jag skall strax ta upp problemen kring OM* på ett annat sätt.[5] En sak som man dock direkt kan se är att om man kan *förkasta* OM*, och det skall jag visa att man kan, så behöver man inte ta ställning till ATT*. Ty *även om* en dator skulle klara testet, så följer när OM* förkastats ingenting om datorns eventuella psyke.

4. ALGORITMER

Jag måste säga något om ATT*. Jag skall först säga något om vad en *dator* är. Jag skall betrakta datorn mycket allmänt, nämligen som ett system vars uppförande bestäms av att den *programmerats* att följa någon *algoritm*. Det sistnämnda begreppet ovan kan exemplifieras av de räknesätt för heltal som vi lär oss tidigt i skolan, dvs. addition, subtraktion, osv. De bestäms av *regler*. Med en regel kan man ta itu med en mängd olika fall. Regeln för addition skall tala om hur man kan räkna ut 1 + 1, 1 + 2, 2 + 3, osv. Additionsregeln täcker alltså *oändligt* många enskilda fall, och så är det också för de tre andra räknesätten och många andra algoritmer.

[4]Det är ju välkänt att det också finns hemska undantag då människor och djur blir behandlade som om de saknade känslor och förnimmelser. Alltså koncentrationsläger, djurfabriker, osv.

[5]Jag skall i denna artikel säga en del om tesen ATT*, dock kan jag av utrymmesskäl inte säga så mycket som den förtjänar. Denna tes är i och för sig mycket intressant och förtjänar en egen artikel.

Redan på dessa elementära räknesätt kan vi se vad som är väsentligt för en algoritm.

Jag skall bara nämna en sak: algoritmens regel kan uttryckas genom en *ändlig* uppsättning *instruktioner* som har en särskild karaktär. De är ”*mekaniska*” i följande mening: instruktionerna måste kunna följas endast genom att man slaviskt, ”som en idiot”, följer själva instruktionerna, utan tillämpning av egen insikt eller kreativitet. Sådana ”mekaniska” instruktioner lämpar sig alltså för maskiner. Algoritmernas mekaniska karaktär är deras utmärkande, och mest intressanta, egenskap.

Algoritmer behöver inte vara matematiska, de kan handla om vad som helst. Alla algoritmiska regler kan reduceras till beräkning av *funktioner.* I sådan ”får man ut” ett entydigt resultat när man ”stoppar in” något. De ovan nämnda räknesätten anger funktioner. Funktioner behöver inte handla om tal. Men alla algoritmiska funktioner måste vara ”*diskreta*”, dvs. de handlar om objekt som kan *numreras av heltal.*

Begreppet algoritm leder till djupa problem i en mängd områden: logik, matematik, fysik, människans intellekt, osv. För att kunna ta itu med dessa problem måste man emellertid precisera begreppet, och den stora svårigheten ligger i att precisera begreppet ”mekanisk”. Det gjorde Turing på 30-talet med sin mest berömda uppsats (Turing 1936) som gav en definition av den s.k. *Turing-maskinen.* Denna är en standardform som han antog att man kunde ge alla algoritmer. Med begreppet Turingmaskin bevisade han i denna artikel att alla diskreta funktioner inte kan beräknas med algoritmer – med Turingmaskiner. Jag skall säga att sådana funktioner är ”*överalgoritmiska*”.[6]

Begreppet ”överalgoritmisk” kan uppfattas som ett *matematiskt begrepp om kreativitet* – eftersom det handlar om överskridande av det mekaniska. Detta kan inte direkt identifieras med mänsklig kreativitet – dock kan man fundera på eventuella samband.

Begreppet ”överalgoritmisk funktion” har en viktig underavdelning, nämligen de överalgoritmiska funktioner som i någon mening kan ”beräknas” med specificerbara procedurer. Sådana funktioner har kommit att kallas ”*hyperberäkningsbara*” (eng. *hypercomputable)* och de har på senare tid bildat ett snabbt växande och kontroversiellt forskningsområde.[7]

[6] I analogi med mängdteorins ”överuppräknelig”.

[7] Till särskilt kända inslag i debatten på området hör Roger Penroses *The Emperor's New Mind* (1989) med efterföljare. Syropoulos 2008 ger en samlad framställning av området. En viktig kritiker av dessa teorier är den framstående logikern Martin Davis, se t.ex. hans ”The Myth of Hypercomputation” i Teuscher 2010.

5. HÅRDVARANS ABSTRAKTA EGENSKAP

Tillbaka till datorn: den kan alltså programmeras med en algoritm. Det betyder att instruktionerna lagras i datorn. Men nu kommer något viktigt: det har ingen betydelse vad de lagras *i*. Mediet, *hårdvaran*, *är oväsentlig*. I nutida teknik finns magnetisk lagring, men också optisk och elektronisk av olika slag. Det kommer säkert nya i framtiden. Det väsentliga i lagringen är att den består i att hårdvaran ges en *abstrakt strukturell egenskap*. Tänk på t.ex. hur ett program överförs, ”laddas ner”, från först en optisk disk ner i datorns hårddisk, som är magnetisk, och sedan i ett USB-minne som består av chips. *Samma* abstrakta egenskap har uppträtt i *olika* konkreta bärare.

Detta resonemang gäller inte bara datorns minnen utan kan också tillämpas på dess processor. Samma beräkningsegenskaper kan realiseras av olika hårdvara, olika konkret uppbyggnad, men med samma abstrakta egenskap. Allt det här kan inspirera till den ovan nämnda ”maximaltolkningen”.

6. DATORERS FÖRMÅGOR

Nu kommer jag äntligen fram till tesen ATT*. En dator styrs alltså av algoritmer. Kan sålunda ett mänskligt beteende vara framställt av en algoritm? Nu gäller det inte bara chatt-beteende utan rubbet. Man gör det alldeles för enkelt för sig om man omedelbart svarar ”Nej!”. Vi skall tänka på tre saker.

För det första: tesen ATT* handlar om vad som är möjligt *i princip,* inte om vad som är mänskligt möjligt. Det gäller alltså inte vad som en mänsklig programmerare rimligen kan åstadkomma med en dator. Vi får i stället tänka oss att det handlar om vad en gudomlig programmerare kan åstadkomma.

För det andra ska vi tänka på *uppdelningen i nivåer*. Datorn är strängt och mekaniskt lagbunden och kan beskrivas med en överblickbar uppsättning matematiska och naturvetenskapliga begrepp. Men så är det inte för en människa på ”mänsklig nivå”, dvs. så som hon uppfattas av andra människor. Vi beskriver människor, och uppfattar dem, med en uppsättning begrepp som inte är naturvetenskapliga. De förra har andra logiska egenskaper än de senare. Det förra slaget beskrivningar kan i allmänhet inte formaliseras. Och även om det finns några lagar för människor på mänsklig nivå är det inte rimligt att hävda att människor är lag*bundna* på denna nivå.[8] Och vidare framstår människors egenskaper som otaliga och oöverblickbara. Det framstår ofta som orimligt att

[8]Detta är något som de flesta vanliga människor vet utan att formulera det, och som ibland uttryckligen har formulerats av filosofer, t.ex. av Donald Davidson.

försöka skapa fullständiga teorier om människor. Tänk bara på hur löjligt det skulle vara att försöka skapa sådana teorier om humor, konst, musik.

Men man kan inte av dessa uppräknade fakta *omedelbart* dra en slutsats att det hos människor finns något som överskrider alla algoritmer. Människan i vardagslivet befinner sig på en annan logisk nivå än datorn. Men detsamma kunde gälla *även en dator* med mycket avancerat program.

För det tredje: *om* människans beteende står i strid med ATT*, så kan det bara betyda att människan har *överalgoritmiskt* beteende. Det är en *extrem* egenskap, som implicerar att *naturlagarna* har överalgoritmisk karaktär. Det kan ju vara sant, och det diskuteras i nutida debatt av Roger Penrose och andra i hyperberäkningssvängen. Argumenten för det här bygger alltså på antagandet att vi människor i något avseende är kreativa i den *matematiska* meningen, och argumenten brukar utgå från några föregivna matematiska förmågor hos människor.[9] Men det är ännu osäkert hur det faktiskt förhåller sig med denna sak.

Sammantaget finner jag att det ännu inte finns några starka argument mot ATT*. Denna tes framstår, åtminstone än så länge, som rimlig.

7. PARALLELLITET OCH IDENTITET

Låt oss nu i tanken genomföra det generaliserade Turing-testet. Vi antar alltså att dessa robotar lurar oss. Det intressanta är hur vi förhåller oss till dem *efter* att de avslöjats. Detta experiment är inte möjligt i verkligheten, med det är lätt att föreställa sig. Vi kan ju föreställa oss människor, och dessa robotar är som människor när de betraktas utifrån. Men sådana robotar finns också på film, t.ex. i ”Alien”-serien. I den visas datorstyrda sociala robotar som först lurar människor och sedan avslöjas.

Vi kan i dessa filmer se hur en möjlig värld med sådana robotar kan vara, men också hur den *måste* vara – och det sistnämnda är viktigt: det blir klart att de som ingår i det sociala spelet med dessa robotar *måste* använda våra vanliga psykologiska och sociala ord, inte bara när det talar *om* robotarna, utan naturligtvis också när de samtalar *med* dem. Det gäller naturligtvis så länge de lurar oss. Men det viktiga är att det *också* gäller efter att de avslöjats.

Och dessa ord måste, i en viss bemärkelse, användas på samma *sätt* som vi gör i våra vanliga mänskliga situationer. Den ”vissa bemärkelsen” är alla yttre, beteendemässiga och sociala, svar på ”när säger vi vad?”. Vi måste alltså använda samma ord, men det måste inte betyda att orden har samma innebörd i robotvärlden som i vår värld. Men det måste råda en

[9] Argumentationen har bl.a. av Pernrose anknutit till Kurt Gödels logiska s.k. ofullständighetsteorem från 1931. De flesta logiker anser emellertid att sådana argument bygger på en missuppfattning av Gödels teorem.

parallellitet. Detta innefattar även psykologiska och sociala *förklaringar*. Vi måste ge dessa robotars uppförande psykologiska och sociala förklaringar som är parallella med våra förklaringar av människor, och med *samma ord* som vi använder om människor. Det jag säger nu gäller inte bara ord utan också den icke-verbala inställningen, inklusive den känslomässiga attityden, till robotarna.

Den *nödvändighet* med vilken allt det här måste gälla är *praktisk*. Det är inte praktiskt möjligt för dem som är medspelare i dessa sociala situationer att förhålla sig till robotarna som till elektroniska eller datalogiska system. Det senare kräver en distansering som förutsätter att man ställer sig utanför dessa situationer.

Allt det jag nu påpekat kan kallas ”*parallellsatsen*”. Den säger alltså något om dessa robotars ”mänskliga nivå” som jag skrev om i föregående avsnitt. Det är inte mycket nytt i detta. Det är en generalisering av Daniel Dennetts teori om ”intentional stance”.[10] Jag skall inte gå närmare in på detta, bara nämna en sak.

Dennett höll sig huvudsakligen till *intentionala* språkliga uttryck och förklaringar. Det gäller uttryck som ”avse att”, ”önska att”, ”planera för”, etc., alltså uttryck som anger ett subjekts inriktning mot ett objekt. Sådana uttryck följer andra logiska lagar än de som gäller för termerna på naturvetenskaplig nivå, varför intentionala uttryck inte kan översättas till naturvetenskapliga. Ändå är det, påpekade Dennett, mycket praktiskt att tala i intentionala termer även om en robot. Men det innebär inte att den har någon mystisk dubbelnatur. Vi har bara att välja att tala olika språk om en och samma verklighet. På samma sätt kan man tänka sig förhållandet mellan den ”mänskliga” nivån i allmänhet och den naturvetenskapliga nivån.

Den viktiga frågan med det här tankeexperimentet är om man kan *förstärka parallellen till identitet*, alltså om man kan förstärka parallellsatsen till OM*. Alltså om de psykologiska orden om robotarna inte bara är paralleller till samma ord om människor utan har *samma innebörd* som de senare.

Wittgenstein hävdade en gång att det är alldeles meningslöst att tala om ett psyke hos en maskin – han menade att det vore lika meningslöst som att tala om färgen hos ett tal. Men det stora problemet med dessa robotar är att det verkar vara så svårt att *låta bli* att tala om dem som bärare av psyken. Det tycks som om Wittgenstein begick det misstag som han ofta anklagade andra för, nämligen det att hålla sig till alltför ensidiga exempel.

Vad är det egentligen för skillnad i psykologiskt avseende mellan dessa

[10]Dennett 1987.

robotar och oss. Vad är det egentligen som *fattas* hos dessa robotar? *Medvetandet* är naturligtvis problemet. Men vi måste använda samma vanliga psykologiska ord, och i samma sociala situationer, även om det *inre,* även om *upplevelserna* hos dessa robotar. Ja, vi skulle även bli tvungna att använda ordet ”smärta” om robotarna. Men problemet gäller inte användningen av dessa ord.

När kan det finnas en skillnad? Man kan räkna upp många fall som är relevanta. En viktig typ av fall är de *moraliska*. Kan man verkligen vara elak mot en sådan här robot? Det är elakt att tillfoga en människa smärta. Men skulle det vara elakt att göra det mot en robot när man måste använda ordet ”smärta” om den? Det tycks vara stor moralisk skillnad mellan att en robot *beter sig som om* den upplever smärta och att den *verkligen* upplever det. Visst kan man tänka sig en slags moral som gör det ”elakt” att bete sig på vissa sätt mot sådana här robotar i de fall som jag nu nämnt. I själva verket skulle det växa fram en sådan moral under de sociala sciencefiction-förhållanden som jag nu tankeexperimenterar om. Men det är inte säkert att en sådan morals problem är *allvarliga*. Skillnaden *konventionell moral/allvarlig moral* tycks bestämmas av frågan om dessa robotar verkligen har medvetande. Skulle en djupsinnig behavioristiskt lagd filosof kunna förneka det? Vad kunde Wittgenstein, som var behavioristiskt lagd och både djupsinnig och allvarlig, ha sagt? Jag vet inte. Antagligen skulle han ha sagt något om livsformer.

8. HUR DET ÄR ATT VARA EN ROBOT

Jag skall nu ta itu med problemet på ett annat sätt: hur skulle det vara om vi kunde *bli* sådana här robotar? *Medvetandet* i den filosofiskt relevanta mening som det nu gäller avser *hur det är att vara* ett subjekt, inte hur subjektet beter sig, och inte hur det talas om subjektet, varken av detta själv eller av andra. Medvetandet i denna bemärkelse är för det första ett *fenomenologiskt* begrepp. Det handlar om hur världen och subjektets inre *ter sig*, *finns till*, för subjektet. För det andra finns detta medvetande *i världen*, dvs. det är inte *endast* en ”transcendental förutsättning” *för* världen[11] – det är knutet till en kropp, närmast till en hjärna. Detta är det klassiska medvetandebegreppet som varit ett tema i västerländsk filosofi sedan början av 1600-talet. Det har under samma tid varit knutet till det s.k. psykofysiska problemet, dvs. *hur* detta medvetande är förbundet med materien, närmast materien i hjärnan.

Emellertid har det från 1900-talet varit en filosofisk stridsfråga om det alls finns något medvetande i denna bemärkelse. Denna strid är i

[11] Tyvärr har jag inte här utrymme att förklara vad det sistnämnda betyder. Jag får hänvisa intresserade läsare till Husserl 2004.

och för sig viktig, men jag skall inte här gå in närmare på dess turer. Ett par av de djupaste angreppen mot begreppet kommer nog från Husserl i hans sena artiklar om vetenskapens kris och från Wittgensteins senare filosofi. Jag skriver ”nog” eftersom det inte är lätt att tolka dessa filosofer. En bra allmän begreppsutredning av och försvar för det klassiska medvetandebegreppet finns i början av Nagel 1993. Jag har tagit upp i stort sett samma idé och kort relaterat den till ”transcendental” filosofi av Husserls typ i min artikel i *FT* 2003.

Nå, nu skall vi återvända till Turings test och göra ännu en generalisering av det. Vi skall generalisera det till *hjärnprotes*-problemet: det pågår nu en utveckling av *neurala proteser* – konstgjorda anordningar som sätts in i nervsystemet och skall ta över dess funktioner. En *hjärnprotes* är en neural protes för någon del av hjärnan. Om man vill skapa en hjärnprotes för någon del av hjärnan som är förbunden med medvetandet måste man ta itu med det psykofysiska problemet och därmed med det klassiska medvetandebegreppet. Detta kommer helt säkert att inträffa om man vill skapa en protes för *hela* hjärnan. Låt oss alltså tänka oss in i ett sådant fall. Hur skulle man göra om man då drar ut konsekvenserna från Turings tankar i artikeln från 1950?

Vi antar att det gäller en man, Kalle. Han har av något skäl, rimligen sjukdom, anledning att byta ut sin hjärna mot en konstgjord protes. Kalle lägger på ett naturligt villkor på protesen: hans liv skall fortsätta efter operationen i mänsklig form och anknyta kontinuerligt till livet före. Han skall vara samma gamla Kalle som förut efter operationen, åtminstone till att börja med. Han skall vara densamma både för andra och för sig själv.

Då prövar vi Turings idéer på detta fall. En gudom som är hårdvaratekniker, programmerare och kirurg ger ett förslag på protes: den skall vara en *dator,* med *elektronisk* hårdvara, alltså mycket olik en hjärna i konkreta avseenden. Gudomen erbjuder sig vidare att skapa ett *simuleringsprogram* för Kalles hela beteende – som det var strax före operationen – och att ladda ned detta program i datorn, sedan ersätta Kalles hjärna med denna och slutligen göra alla ihopkopplingar till utgående och ingående nerver. I det väsentliga betyder det här att Kalle *blir* en sådan robot som jag skrivit om ovan. Att han har kvar sin biologiska kropp utanför hjärnan är oväsentligt. Skall Kalle gå med på det? Vi kan för enkelhets skull anta att det verkligen går att skapa simuleringsprogrammet. ATT* är alltså uppfylld. Nu hänger allt på OM*.

Protesen klarar Turings allra mest generaliserade test. Den lurar alla, dvs. alla utom Kalle själv. Ingen som känner Kalle, men inte vet att han blivit opererad, kan misstänka att han blivit det. Och parallelltesen blir uppfylld: alla (utom Kalle själv) förhåller sig till Kalle efter operationen

som till en vanlig människa, både språkligt och icke-språkligt. Och Kalle själv talar naturligtvis om sig själv som en vanlig medveten människa. Är inte allt frid och fröjd då? Nej, *naturligtvis* inte. Kalle bör förkasta denna protes!

Medvetandet kan identifieras med det totala, mest inklusiva, fenomenologiska begreppet: *livet i fenomenell mening.* För Kalle är det viktiga att hans *liv* skall fortsätta efter operationen, dvs. livet så som det är för honom när han lever det. Tycker du läsare att det finns något oklart med detta begrepp? Skärp dig då! Det ju är helt klart *vad* det är som Kalle kräver att det skall fortsätter efter operationen. Vi vet alla vad livet är i denna mening. Det är det där vanliga, vardagliga – ingenting annat. Det kan inte begreppsligt identifieras med livet som exakt bestämd biologisk process. Detta blir ju mycket förändrat av operationen, och det stora problemet är hur mycket man kan ändra det utan att livet ändras.

Enligt det här Turingska protesförslaget skall alltså livet i fenomenell mening identifieras med (ev. någon aspekt av) datorns mjukvara. Alltså med en *abstrakt strukturell egenskap* hos datorns hårdvara. Man kan då invända att livet *inte är abstrakt utan konkret* – och det går att precisera.

Men man kan påpeka något mycket enklare: livet är inte någon *egenskap* alls (men det *har* naturligtvis egenskaper). Livet är ett *självständigt föremål,* eller kanske man kan säga ”*substans*” i någon av ordets filosofiska meningar. Det där låter kanske väldigt konstigt. Livet är inte någon sådan vanlig pryl som det finns så många av i vårt konsumtionssamhälle. Vad jag tänker på är ett självständigt föremål *i filosofisk mening.*

Ett sådant kan för det första inte prediceras om något och för det andra är det inte för sin existens väsensmässigt beroende av att något annat existerar. Egenskaper kan prediceras om något. Man kan predicera egenskapen ”röd” om en boll, bollen *är* röd, men bollen själv kan inte prediceras om något. I *denna* mening kan inte heller ett fenomenellt liv prediceras om något. Det här gäller inte användningen av ord i vanlig mening, utan *fenomenologi*. Vårt språk är ganska missvisande när det gäller denna sak. Vi skall dock inte hänga upp oss så mycket på detta utan se på det andra som utmärker ett självständigt föremål.

Här sitter jag nu framför min dator och skriver. Det är en liten bit av mitt liv. Jag ser datorn, känner tangenterna, är inne i skrivandet som aktivitet, skymtar mitt vardagsrum, känner min kropp inifrån, har vissa tankar och känslor, osv. Alla dessa ord måste nu ges fenomenell innebörd, de handlar alltså om *hur det är för mig.* Nu är klockan 20.30. Ungefär så här fortsätter mitt liv fram till klockan 21. Då är följande *filosofiskt* möjligt: När klockan blivit 21 dyker det upp en granne som hojtar: ”Å, så fantastiskt att du är tillbaka Tom! Din kropp har varit obefintlig, har inte existerat, mellan 20.30 och 21! Detta har vi kunnat bevisa med vetenskapliga

instrument". Då måste jag svara: "Det var konstigt. Allt har varit som vanligt för mig under denna halvtimme. Jag har suttit och skrivit på min dator och ingenting särskilt har hänt!"

Jag skulle bli förvånad om något sådant skulle inträffa i verkligheten (men ibland blir man ju förvånad!). Det skulle strida mot mina föreställningar om naturen. Jag menar bara att det här är möjligt utifrån mitt fenomenella livs väsen, som visar sig för mig. Det betyder emellertid att mitt fenomenella liv är ett självständigt föremål, eller en substans. Likadant måste det vara för Kalle. Därför bör han förkasta den Turingska protesen som förutsätter att medvetandet är en (abstrakt) egenskap. Och därför är OM* falsk. Och följaktligen är Turings idéer i artikeln från 1950 falska.

Nu måste jag sluta denna artikel, men nu börjar de stora frågorna.

LITTERATUR

Asplund, Johan. 2002. *Genom huvudet: Problemlösningens socialpsykologi*. Göteborg: Korpen.

Davis, Martin. 2010. "The Myth of Hypercomputation". Ingår i Teuscher 2010.

Dennett, Daniel. 1987. *The Intentional Stance*. Cambridge, Mass.: MIT Press.

Eriksson, Tom. 2003. "Om det psykofysiska problemet". *Filosofisk tidskrift* nr 3.

Husserl, Edmund. 2004. *Idéer*. Stockholm: Thales.

Nagel, Thomas. 1993. *Utsikten från ingenstans.* Nora: Nya Doxa.

Penrose, Roger. 1990. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford: Oxford University Press.

Stenlund, Sören. 1990. *Language and Philosophical Problems*. London: Routledge.

Syropoulos, Apostolos. 2008. *Hypercomputation*. New York: Springer.

Teuscher, Christof, red. 2010. *Alan Turing: Life and Legacy of a Great Thinker.* Heidelberg: Springer.

Turing, Alan M. 1936. "On Computable Numbers, with an Application to the Entscheidungsproblem". *Proceedings of the London Mathematical Society*. Ser. 2, Vol. 42, pt. 3–4 (Nov.–Dec. 1936). Finns på nätet.

Turing, Alan M. 1950. "Computing Machinery and Intelligence". *Mind* 59, nr 236 (oktober), 433–60.