

Inför konstruktionen av en medveten dator

Medvetandets informationsinnehåll

Det mänskliga medvetandet består av en serie i hjärnan lokaliserade processer. Dessa processer har fem nödvändiga beståndsdelar, vilka vid skilda tidpunkter kan konstituera medvetandet var för sig, parallellt eller genom att avlösa varandra:

- (I) sinnesförmimmelser av omvärlden,
- (II) föreställningar om sinnesförmimmelser,
- (III) aktiverade minnen,
- (IV) känslolägen, och
- (V) en inre monolog, som *kan* vara språklig, och vilken inte kommuniceras till omvärlden.

Exempel på kategori (II) är drömmar, individers överväganden av alternativa vägar till ett bestämt mål och tyst nynnande på en melodi.

Ett avgörande kännetecken på det vi kallar medvetna upplevelser är att dessa vid någon tidpunkt har varit tillgängligt i det självbiografiska minnet, det vill säga det minne som samlar den information och upplevelser som utgör individens subjektivt ihågkomna livshistoria. För att definieras som medvetna måste därför serierna av processer i hjärnan enligt (I) – (V) koda in i det självbiografiska minnet en kortare eller längre tid (för en dement person är detta minne mycket kort). På grund av hjärnans begränsning med avseende på minneskapacitet registreras endast en liten del av den totala hjärnverksamheten och sinnesförmimmelserna i det självbiografiska minnet. Övrig hjärnverksamhet registreras i andra minnen – exempelvis arbetsminnet, det procedurella minnet eller det semantiska minnet – eller inte alls. Medvetandets informationsinnehåll kan således definieras på följande vis:

Medvetandets informationsinnehåll består av den serie av aktiverade neurala kopplingar i hjärnan som utgör individens självbiografiska minne.

Medvetandet som upplevelse

Nästa fråga är hur själva *upplevelsen* av att vara medveten uppstår. För att besvara denna fråga krävs en viktig ytterligare pusselbit, nämligen insikten att hjärnverksamhet av typ (I) – (V) inte existerar vid en enskild tidpunkt, eftersom den är uppbyggd av nervimpulser som med nödvändighet har en utsträckning i tiden. Av detta följer att hjärnans medvetna processer inte är

momentana utan har en temporal utsträckning. Detta medför i sin tur att de måste kodas in i minnet, innan de är tillgängliga för introspektion (detta begrepp preciseras nedan). Således upplever individen sitt medvetande i *efterhand* genom sitt minne.

Att den medvetna upplevelsen är fördröjd på detta vis kan tyckas vara kontraintuitivt, eftersom medvetandet kan framstå som människans mest omedelbara upplevelse. Motsägelsen upplöses emellertid genom insikten att integrationen av förnimmelser i hjärnan *kräver* att upplevelser är efterhandskonstruktioner i form av minnen som kodas in och sedan medvetandegörs med en viss fördröjning. Ett typexempel är den medvetna upplevelsen hos en individ, som samtidigt låter vidröra sin näsa och sin fot. Detta ger upphov till upplevelse av två *samtidiga* sinnesförnimmelser, trots att hjärnan registrerar nervimpulserna från foten avsevärt senare än nervimpulserna från näsan. Hjärnan förenande av dessa båda förnimmelser till något som upplevs som samtidigt kan endast ske genom en efterhandskonstruktion. Ett annat exempel är upplevelsen av en vacker melodi (eller en behaglig dans), vilken uppenbarligen kräver, att ett antal toner (eller rörelser) över flera sekunders tidsintervall lagras in i minnet för att därefter plockas fram och värderas som en helhet.

I enlighet med ovanstående resonemang kan följande hypotes formuleras:

Upplevelsen av medvetandet kan definieras, som att självbiografiska minnen av tidigare hjärnverksamhet av typ (I) – (V) från tidpunkten t_1 associeras med ny hjärnverksamhet av typ (I) – (V) vid den senare tidpunkten t_2 , varefter delar av denna integrerade information på nytt lagras i det självbiografiska minnet. Därvid har individen vid tidpunkten t_2 ett minne av att ha upplevt sig medveten vid tidpunkten t_1 .

Medvetandet kan således endast upplevas i efterhand – om än ibland med en fördröjning på endast bråkdelar av en sekund. Ett iögonfallande exempel på sådan fördröjning uppstår, när en person blir medveten om det störande ljudet från ett kylskåp eller en ventilationsanläggning först när oljudet stängs av.

Upplevelsen av medvetandet är enligt ovanstående definition ett slags medvetande på metanivå, det vill säga medvetenhet om tidigare medvetna processer, vilket med ett annat ord kan kallas *introspektion*. Regress av detta slag kan ha så många nivåer som individen önskar, genom att individen gör sig medveten om sin medvetenhet om sin medvetna upplevelse, och så vidare tills hen riskerar att förlora sig i självreflektionens obegränsade djup.

Medvetandets metanivåer framträder ibland särskilt tydligt, exempelvis vid olika former av dissociativa tillstånd (bl.a. upplevelse av överklighet samt s.k. ”minnesförlust”) eller då en sovande individ plötsligt blir medveten om att hon drömmer.

Medvetandets illusoriska karaktär

Förmodligen är ovan nämnda metanivåer ett av skälen till att medvetandet är så svårt att studera och beskriva för oss människor, trots att vi alla säger oss uppleva detta fenomen. Ett annat troligt skäl är, att medvetandet har illusoriska inslag som evolutionen har gynnat, för att sådana illusioner innebär direkta fördelar i kampen för överlevnad, alternativt att illusionerna är en följd av andra funktioner som har sådana fördelar.

Övertygelsen att medvetandet är samtidigt med de registrerade processerna snarare än att vara en efterhandskonstruktion, är exempel på en sådan illusion. En annan illusion består i att medvetandet har en bestående komponent genom tiden – det så kallade ”jaget” – trots att någon sådan komponent inte kan identifieras (exempelvis vid jämförelse mellan spädbarnet och den åldrade versionen av samma individ). Upplevelsen av medvetandet liknar i detta avseende upplevelsen av att ha en ”fri” vilja eller upplevelsen av tidens gång, vilka likaså visar tydliga tecken på att bygga på illusioner.

Att konstruera medvetandets funktion i en dator

Går det då att konstruera medvetenhet hos en dator? Ja, det tycks i enlighet med ovanstående vara fullt möjligt genom att i en datorenhet programmera in följande beståndsdelar:

- (1) förmåga till sinnesförnimmelser, det vill säga processande av sensorisk data från omgivningen (som enklast erhålls genom språklig inmatning via tangentbordet),
- (2) förmåga till intern språklig monolog,
- (3) förmåga att associera (integrera) nya förnimmelser och monologer med tidigare minnen,
- (4) förmåga att attribuera (tillskriva) känslor (inledningsvis ett fåtal grundemotioner, exempelvis glad, ledsen och nyfiken) till upplevelser av förnimmelser, monologer och tidigare minnen,
- (5) ett självbiografiskt minne där åtminstone en del av de i punkt (1) – (4) nämnda förnimmelserna, monologerna, associationerna och känslotribueringarna registreras, och
- (6) förmåga att låta olika känslor påverka val enligt känslospecifika mönster, exempelvis vid val av ämne för inre monolog och framplockning av minne.

Därutöver är det lämpligt att förse den aktuella datorn med en förmåga att språkligt kommunicera ovanstående processer till omvärlden, för att på så vis i realtid kunna visa att det i datorn försiggår en medveten process.

En sådan konstruktion har primärt inget att göra med det så kallade Turingtestet, i vilket det provas om en dator genom språklig kommunikation kan få en bedömare att tro, att den är en människa. I ett senare skeende, när den ”medvetna” datorn har nått en tillräcklig komplexitetsnivå, kan emellertid ett sådant test vara av intresse som slutlig prövning.

Qualia

Det hävdas från en del håll, t.ex. av filosofen John Searle, att medvetandet skulle vara förknippad med en omedelbar känsla, så kallad *qualia*, som en dator aldrig kan erhålla på grund av sitt regelstyrda informationsprocessande. Vanliga exempel är känslan av färg, musik och smärta. Ingenting hindrar dock att en slags deklarativ subjektivitet åstadkoms i en dator genom associationer till andra företeelser av olika emotionell valör, t.ex. att röd associeras med värme och blod. Däremot är det omöjligt att vetenskapligt studera eller definiera den ”känsla” som därvid eventuellt uppstår, eftersom dess subjektiva karaktär gör den oåtkomlig för andra än subjektet självt. Denna begränsning är ofrånkomlig och gäller alla subjektiva upplevelser, exempelvis ifråga om hur det känns att vara en fladdermus (Nagel, 1974), att bli torterad eller att njuta av en bit choklad.

Medvetandets evolutionära funktion

Ledtrådar om medvetandet kan möjligtvis även erhållas genom att försöka räkna ut dess evolutionära överlevnadsvärde. Att dess innehåll hos människan kan definieras som identiskt med innehållet i det deklarativa minnet, ger en antydning om att överlevnadsvärdet ligger i att vissa processer i hjärnan görs tillgängliga för kommunikation med omvärlden, förmodligen främst med artfränder. Frågan är då varför inte *all* hjärnverksamhet görs tillgänglig för kommunikation med omgivningen. Ett rimligt svar på den frågan är att det skulle kräva allt för stora minnesresurser att över huvud taget minnas all hjärnverksamhet så länge, att den kan kombineras med andra samtidiga processer och associeras med tidigare minnen – det mesta av hjärnans processer glöms istället omedelbart bort.

Detta leder till hypotesen, att endast arter som har användning för ett deklarativt minne – det vill säga ett minne som samlar fakta om en del av hjärnans processer, som sedan kan kommuniceras till omgivningen – har ett medvetande. Låt oss säga att en chimpanzé, en delfin eller en kaja har sett en

annalkande fara och vill meddela detta till sina artfränder. Behöver den i så fall ett deklarativt minne för att klara denna uppgift (vilket enligt ovanstående evolutionära resonemang skulle tyda på förekomsten av medvetande hos arten)? Ja, om det förekommer en fördröjning mellan uppfattandet av faran och varningen till artfränderna, tycks det som om ett sådant eller ett liknande minne är nödvändigt. Således leder detta resonemang till hypotesen, att de arter – där individerna behåller fakta i minnet för att senare meddela dessa till artfränder – har ett medvetande.

Innehåll, funktion, lokalisering, upplevelse

Ett i mina ögon fruktbart förhållningssätt till medvetandet är att dela upp analysen i innehåll, funktion, lokalisering och upplevelse. Innehållet är lätt att fastställa eftersom allt som hamnar i det deklarativa minnet är medvetet. Funktionen är något svårare att precisera men innebär inga principiella svårigheter, eftersom vi helt enkelt kan fråga människor vilken sorts funktioner som utmärker deras medvetanden.

Medvetandets lokalisering i hjärnan är en fråga för neurologin och har inte nått särskilt långt, men det finns spekulationer om att *claustrum* som finns alldeles under hjärnbarken har en nyckelroll (Nikolenko m.fl., 2021). Lokaliseringen saknar dock i stort sett helt och hållet betydelse för frågan om en dator kan förse med medvetande.

Upplevelsen, slutligen, är en fråga av mestadels metafysisk karaktär, eftersom den är av så subjektiv karaktär, att en individ inte ens kan avgöra om någon annan än hen själv upplever sig medveten. Som regel grundar vi vår uppfattning om den eventuella förekomsten av medvetandepplevelse hos andra varelser på vår förmåga att leva oss in i dessa varelser, vilket är beroende av vår förmåga att tolka deras signaler, vilket i sin tur beror på hur lika de är oss själva. Således har vi starkast tendens att tillskriva andra människor medvetande, i andra hand andra primater, i tredje hand välkända husdjur, i fjärde hand däggdjur, i femte hand fåglar och så vidare.

Vilka arter har medvetande?

Ett sätt att ändå avgränsa vilka arter som kan möjligen kan uppleva medvetandet är att fråga sig vad en sådan upplevelse har för funktion för överlevnaden, vilket sedan ger ledtrådar i form av vilka levande varelser som delar den aktuella överlevnadsstrategin. Om exempelvis möjligheten att kunna kommunicera minnen (det vill säga ”deklarera” fakta ur minnet) med artsfränder är det som ger medvetandet överlevnadsvärde – vilket den starka kopplingen till det deklarativa minnet antyder –

bör upplevelsen av medvetandet inte sträcka sig bortom de arter som har förmåga till och fördelar av sådan kommunikation. Den enklaste och mest "primitiva" formen av sådan kommunikation är att varna artfränder för en fara, som individen har blivit varse *tidigare* (inte *samtidigt* eftersom det inte kräver minneslagring), och sådan kommunikation tycks förekomma hos åtminstone alla däggdjur och fåglar och förmodligen även hos reptiler och kanske till och med hos sociala insekter som bin och myror. Växter och encelliga djur tycks däremot enligt ett sådant resonemang inte ha några överlevnadsfördelar av ett medvetande, men att dra en *skarp* gräns mellan medvetna och icke medvetna arter utifrån ett evolutionärt perspektiv framstår som en omöjlig uppgift.

Öppen fråga

Bör theory of mind vara en del av en artificiell intelligens medvetande, något som människor tillägnar sig någonstans runt fem års ålder (och som ger oss förmågan att bedra andra varelser)?

REFERENSER

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, 83, 435–450.

Nikolenko, V. N. m.fl (2021). The mystery of claustral neural circuits and recent updates on its role in neurodegenerative pathology. *Behavioral and Brain Functions*, 17:8.

Searle, J. (1997). *The Mystery of Consciousness*. New York: New York Review.