Är artificiell intelligens intelligent?

På sista tiden har nyheter om ett genombrott inom artificiell intelligens (AI) förekommit allt tätare, ofta i kombination med varningar utfärdade av experter och toppforskare om vilka risker som AI kan innebära för mänskligheten. Tack vare en nästan ofattbar beräknings- och lagringskapacitet har AI nått en nivå som får en del av science fiction-genren att verka urmodig. AI förstår vårt tal, den talar själv och översätter vid behov nästan felfritt. Den styr drönare och robotfordon, känner igen musik, fågelläten och ansikten. Möjligheterna som AI erbjuder tycks vara gränslösa, på gott och på ont. Utan övervakning sägs AI kunna vara lika ödesdiger som pandemier eller ett kärnvapenkrig.

Datorer och sökmotorer som Google har hört till vardagen en längre tid, men nya chattrobotar eller ”digitala assistenter” som ChatGPT, som letar fram information och framställer den på ett för människan naturligt sätt, tycks ha överskridit en tröskel. Det har blivit allt svårare att skilja en chattrobots text från en människas, eller för den delen ett AI-producerat musikstycke från en mänsklig komposition. Finns det skäl till optimism eller oro? Förmodligen gäller både och, men i verkligheten är det ingen som vet.

Riskerna med AI är uppenbara om vi låter den välja för oss och besluta över våra liv. Bedragare på internet försöker ständigt dra nytta av vår oförmåga att skilja mellan det som är sant och äkta från det som endast imiterar sanningen. Dessa bedrägeriförsök blir allt mer avancerade och allvarligare. Bildbaserad igenkänningsteknik, också den ett slag av AI, kan ge totalitära regimer nästan obegränsad kontroll. Samtidigt står det redan klart, att det moderna samhället inte längre kan fungera utan hjälp av datorer och informationsteknik. Digitaliseringen har genomförts för att underlätta vår vardag och minska behovet av mekaniskt tankearbete, och det har skett i den grad att hela vårt tänkande nu håller på att förändras. En hel generation har vuxit upp, om inte flera, som knappt kan räkna utan en räknedosa eller minnas dikter, ramsor eller regentlängder, för att inte tala om årtal och telefonnummer. Människan är på väg att ge bort en väsentlig del av sin mänskliga förmåga att erinra, reflektera och återupptäcka information. Vi nöjer oss med att söka information med våra elektroniska apparater när vi behöver den, men därmed sätter vi oss frivilligt i en sårbar position.

Det är naturligtvis sant att data inte är meningsfulla i sig, utan ett sammanhang. Varför memorera nummerserier som telefonnummer? När allt kommer omkring är det inte att minnas i sig som är det väsentligaste för oss, utan innebörden och förståelsen, som minnet tjänar. Faran är dock att tekniken som var tänkt att befria oss från mödan att minnas i själva verket fördummar oss och gör oss beroende av ständigt föränderlig teknologi. Det som datorerna troligen inte ännu har uppnått är generell förståelse av den typ som liknar mänskligt medvetande. Och frågan är, om det av principiella skäl ens är möjligt.

År 1997 kom nyheten om att superdatorn Deep Blue vunnit över schackmästaren Garry Kasparov. Deep Blue vann tack vare sin enorma beräkningskapacitet. För varje möjligt drag vid en given situation kunde datorn förutse motståndarens alla tänkbara drag, sina egna svar på dessa drag, och så vidare åtminstone sex steg framåt. Med en snabb utvärderingsfunktion kunde den beräkna nyttan med varje

möjlig situation och sedan göra det drag som ledde till det bästa resultatet. Den utvärderade flera miljoner situationer i sekunden, medan Kasparov bara kunde utvärdera några tiotal innan han behövde fatta ett beslut. Deep Blue var uppenbarligen mycket bra på att spela schack, men någon förståelse om schackspelets natur hade den inte. Den kände inte till klassiska schackstrategier utan använde sig av rå beräkningsstyrka. Att kalla Deep Blue intelligent är därför en grov överdrift. För dagens datorer skulle den troligen vara en lätt motståndare. Men vad är intelligens egentligen? Kan den nivå av artificiell intelligens som vi för närvarande har uppnått, ett kvartssekel efter att Deep Blue besegrat en människa i schack, kallas intelligent?

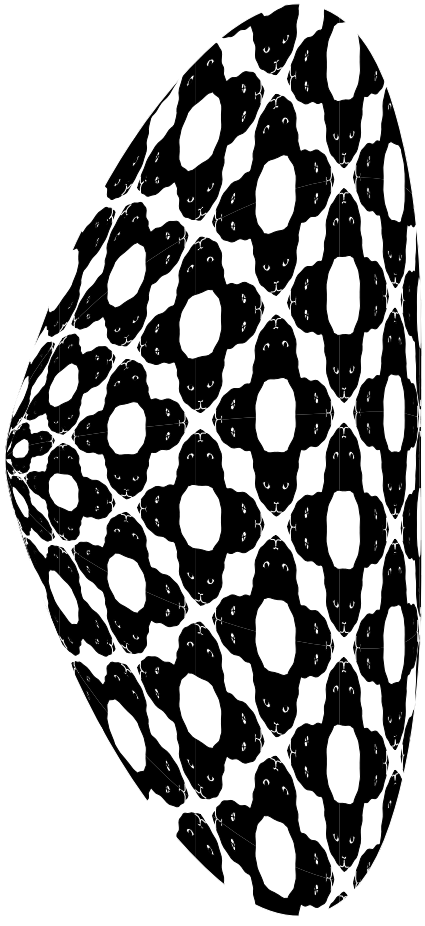
Det som Deep Blue har gemensamt med dagens superdatorer är att de kör datorprogram. De är med andra ord programmerade av en varelse som verkar på en annan ”meta-nivå” än de själva. Datorerna tänker i så motto inte, om vi med tänkande avser den komplexa neurofysiologiska aktivitet vi människor automatiskt ägnar oss åt som medvetna varelser. Datorer har inget ”själv”, inga avsikter, känslor eller föreställningar. De kläcker inga idéer, skapar inga nya begrepp eller teorier. Vackert eller fult, gott eller ont bekommer inte en dator. Man kan be en chattrobot skriva en dikt om kärlek, längtan och drömmar, men för datorn är de ord utan mening.

Desto bättre är datorer på att lösa specifika numeriska problem, kombinera och bearbeta enorma mängder data på ett sätt som vi människor lätt imponeras av och därför kallar intelligens. Datorer imiterar intelligens, men är skillnaden kvalitativ eller endast i grad? Frågan är svår att besvara. Till exempel strävar maskininlärning, som är en växande bransch inom AI, att med hjälp av data träna datorer att själva upptäcka och ”lära sig” regler för att lösa uppgifter. Tillämpningar finns inom datorseende och mönsterigenkänning, som redan innebär en viss typ av kognitiv förmåga.

För att orientera oss kan det vara bra att kasta en blick på hur vi kommit hit. Idén om intelligenta maskiner går tillbaka till år 1950, då den brittiske matematikern och datavetaren Alan Turing uppfann ett sätt att testa huruvida en maskin verkligen är intelligent. Han kallade sitt experiment för ”The Imitation Game”, vilket också råkar vara namnet på en film om Turing och hans tragiska liv (2014). Idag kallas det för Turing-testet.

Turing hade sedan 1936 arbetat på idén om en dator som med de rätta instruktionerna kunde programmeras till att göra nästan vad som helst. Även om den aldrig byggdes efter hans plan baserar sig dagens datorer till stor del på hans idéer. Och Turing trodde av allt att döma, att datorer en dag skulle bli sofistikerade nog att tänka. Vi kan betrakta det en människa säger eller gör – och samtidigt CT-scanna hennes hjärna – och anta, att det vi iakttar är ett uttryck för att hon tänker. Om en dator kan prestera detsamma som en människa måste väl även datorn tänka. Konklusionen är rimlig men är inte logiskt bindande.

Turing framlade sin idé om ett test i tidskriften *Mind* år 1950. Artikeln inleddes med orden: ”Jag föresätter mig att undersöka frågan: Kan maskiner tänka?” I testet sitter personen A ensam i ett rum och utväxlar meddelanden med två andra deltagare, kalla dem B och C. En av dessa deltagare är en människa, den andra är en dator. A utfrågar B och C inom ett begränsat ämnesområde med målet



att avgöra vilkendera av dem som är en människa. Testet upprepas många gånger. En dator som lyckas övertyga A om att den är en människa minst varannan gång anses ha klarat testet, eftersom datorn då verkar lika mänsklig som den mänskliga svararen. Ett omvänt Turing-test (så kallat captcha-test) förekommer när användaren anmodas övertyga en dator om att hon inte är en robot (till exempel genom att känna igen en visuellt förvrängd bokstav).

När Turing skissade sitt imitationstest var det ännu ett hypotetiskt experiment. Den tidens datorer var långt ifrån tillräckligt avancerade för att testas. Transistorn, som mångfaldigade datorernas beräkningskapacitet, var under utveckling. När transistorer integrerades i allt tätare mikrochip, visade sig datorers kapacitet fördubblas ungefär vartannat år, vilket kallas Moores lag. I vår tid har datorerna utvecklats så långt att de redan klarar av Turing-testet. Resultatet är ändå omtvistat. En del datorer har rent av programmerats att begå små fel och misstag för att verka mer mänskliga. Hos människan är misstag alltid misstag, men man frågar sig vad ”misstag” som en maskin ”begår” så att säga med flit kan vara. Tecken på dumhet, kanske? Turing-testet är kanske inte hela sanningen om intelligens.

Alan Turing avled år 1954. Några år senare myntades begreppet artificiell intelligens av datavetaren John McCarthy, som var en av arrangörerna för ett berömt möte sommaren 1956 i Dartmouth, USA. Denna workshop anses allmänt utgöra början på forskningen kring AI. I en utdragen brainstormingssession som varade i ungefär åtta veckor deltog tidvis ett tiotal vetenskapsmän från olika områden. Man antog optimistiskt att man genom ett intensivt samarbete skulle kunna skapa maskiner med förmågan att använda språk, forma begrepp, lösa problem som hittills bara människor kunnat lösa, och att lära sig. Till slut måste gruppen ändå erkänna sitt misslyckande, eftersom AI visade sig långt svårare att skapa än de trott.

Forskningen kring AI, som börjat som en strävan att svara på Turings berömda fråga, ”Kan maskiner tänka?”, hade gått i stå. I stället hade en avart av AI, ett applikationsdrivet underområde av mjukvaruteknik, uppkommit i slutet av 1970-talet. Den amerikanska armén behövde simulatorer och navigationssystem för sina stridspiloter och konstruktörer av olika slags automatiserade expertsystem. Som en biprodukt uppstod persondatorn och dagens allt kraftigare superdatorer.

År 1979, ungefär samtidigt som mjukvarutekniken började utvecklas på allvar, utkom den amerikanske datavetaren Douglas Hofstadters bok *Gödel, Escher, Bach. An Eternal Golden Braid*. Boken var med sina nästan 800 sidor en omedelbar succé som belönades med Pulitzerpriset och småningom blev en klassiker inom AI. Tack vare sina lättsamma dialoger mellan Akilles och Sköldpaddan, ibland även Kräftan och Myrsloken, kunde den läsas och uppskattas, åtminstone delvis, som ett skönlitterärt verk. För att vara tillgänglig för den stora allmänheten var *GEB*, som boken kallades, ändå mycket utförlig och fuskade föga i detaljerna. Den kom ut på svenska år 1985 på Brombergs förlag i översättning av Jan Wahlén.

GEB är tvärvetenskaplig och rör sig i snittet mellan datavetenskap, kognitionsvetenskap, neurovetenskap och psykologi. Den handlar i grunden om Turings problem: vad tänkande är. Dess återkommande teman är datorprogram som refererar till sig själva, den genetiska koden, ”den perfekt återgivande skivspelaren”, så kallad *mise en abyme*[1] inom bildkonsten och en viss oändlig kanon inom musiken. Boken är inte bunden till existerande teknologi och kunde därför behålla sin aktualitet och sitt värde. Emellertid befann den sig snart utanför forskningens huvudfåra, som hade anammat ett nytt mål: att få datorer att prestera effektivt på alla tänkbara sätt.

För att illustrera sin syn på medvetandeproblemet från olika vinklar diskuterar Hofstadter tre protagonister: logikern Kurt Gödel, grafikern M.C. Escher och kompositören J.S. Bach, och vad de har gemensamt. År 1931 hade österrikaren Gödel visat hur man i ett matematiskt system kunde uttala sig inte bara om siffror utan om själva systemet (självreferens av typ ”detta påstående är falskt”). Medvetandet är enligt Hofstadter väsentligen en nivåöverskridande återkoppling av detta slag. Holländaren Escher utforskade självreferens, högre dimensioner och visuella paradoxer med sina hyperrealistiska bilder. Hans berömmelse fick sin begynnelse i Mar-

1 När en bild innehåller en kopia av sig själv i upprepade, allt mindre steg. Även känd som Drosteeffekten.

tin Gardners sakkunniga kolumner i *Scientific American*. Och Bach, slutligen, som lyckades inte bara med konststycket att byta tonart successivt i en kanon för att till slut landa i begynnelsetonarten, utan också att självharmonisera stycket med dess spegelbild. För Hofstadter utgör det här det första exemplet i historien av en sällsam slinga (*strange loop*).

Sällsamma slingor är legio i dagligt tal, exempelvis i synbarligen motstridiga uttalanden som ”intoleransen för intolerans”. Däremot medför självreferens ofta ett problem för datorer. En följd av det så kallade stopp-problemet (*halting problem*) inom beräkningsteorin är att det finns uppgifter som en dator (Turing-maskin) inte kan utföra, däribland att resonera om sig själv. Stopp-problemet är en datateoretisk motsvarighet till Gödels första ofullständighetsteorem (1931), enligt vilken det i varje konsistent formellt system finns sanna men likafullt oavgörbara påståenden. Till exempel ett program för beräkning av decimalerna av π (3,14159…) blir i princip aldrig färdigt, men det kan datorn inte själv avgöra. Att pi är ett transcendent (och därmed irrationellt) tal bevisades år 1882.

I en intervjuartikel som publicerades i tidskriften *The Atlantic* år 2013 argumenterar Hofstadter för att kärnan i mänsklig intelligens är att ”förstå de mentala kategoriernas flytande väsen”, till exempel att förstå vad alla ”A” har gemensamt. ”Kognition är igenkännande” (*cognition is recognition*), säger han och beskriver ”se något som” som den väsentliga kognitiva handlingen: att se vissa linjer som ”ett A” och att uppleva en viss uppsättning av bräder som ett bord. Det är vad det innebär att förstå. Men hur fungerar förståelsen? Hofstadter och hans elever har i mer än tre decennier försökt ta reda på det genom att bygga datormodeller av tankens grundläggande mekanismer. Det kan nämnas att Hofstadter är språkkunnig och förstår svenska. Han har också nyligen samarbetat med författaren Christer Sturmark i en gemensam bok, *Konsten att tänka klart* (2021).

En annan nu levande originell tänkare inom AI är den Nobelprisbelönte fysikern Roger Penrose (känd även för Penrosetessellationen som anlagts bland annat på Centralgatan i Helsingfors). Han har länge intresserat sig för medvetandeproblemet och hävdat att vår tids fysik inte är tillräcklig avancerad för att förklara medvetandet. I de populärt hållna böckerna *The Emperor's New Mind* (1989) och *Shadows of the Mind* (1995) söker Penrose visa att tänkandet inte är en mental beräkning och att hjärnan inte väsentligen är en programmerad dator. Vissa analogier finns, förvisso, men hjärnan fungerar inte i grunden på samma sätt. Som fysiker söker han lösningen i en ny slags konsistent kvantmekanik, som är fri från paradoxer likt den kända ”Schrödingers katt”.[2] Hans senaste idé om att medvetandet är resultatet av kvantgravitationseffekter i hjärncellernas mikrotubuli har rönt kritik av den svensk-amerikanske fysikern Max Tegmark.

Samtidigt som horder av AI-experter försöker lära maskiner att imitera mänskligt tänkande, att känna igen och lära sig, vet vi inte innerst vad vårt medvetande är. På det filosofiska planet leder medvetandeproblemet till Descartes dualism: de två gåtfulla substanserna, det fysiska *res extensa* och det mentala *res cogitans,* som ingenting har gemensamt. Det förefaller som om vi bara har kommit en liten bit på vägen att förstå deras inbördes förhållande och vad liv är för något. En annan växande trend inom datavetenskapen är forskningen som siktar på en kvantdator, som avviker från den vanliga binära datorn i att den kan utföra flera beräkningar samtidigt. Tillstånden utgörs inte av traditionella bitar, rader av ettor och nollor, utan varierande koherenta superpositioner av dessa kallade kvantbitar (*qubit*). Frågan är vilka löften som kvantdatorn, då den en dag framtagits, förmår infria.

Alla dessa gigantiska multidisciplinära projekt är förbluffande. Försöken att bygga mekaniska kopior av oss själva kan antingen vara till stor nytta eller likna ett Frankensteins monster. Mänskligheten är besatt av sitt grundproblem, vad det är att vara människa och en medveten varelse, synbarligen ensam i ett oändligt universum. Osökt går min tanke till Harry Martinsons dystopiska ord i *Aniara:* vi rör oss aningslöst i ett öde hav, bisatta i vår stora sarkofag.

JOHAN C.-E. STÉN

2 Tankeexperimentet illustrerar en paradox inom kvantmekaniken. En katt är utsatt för en giftbägare som utlöses av ett radioaktivt preparat med viss sannolikhet. Enligt en tolkning kan katten vara *både* levande *och* död samtidigt. I verkligheten är dock katten antingen eller.