

Gärdenfors, Peter, 2024. *Kan AI tänka? Om människor, djur och robotar*. Stockholm: Fri tanke.

Anmälan av Maria Hedlund

Nej, AI kan inte tänka. Det är det korta svaret på frågan som kognitionsforskaren Peter Gärdenfors ställer i titeln på sin bok *Kan AI tänka? Om människor, djur och robotar* (Fri Tanke, 2024). Men som Gärdenfors påpekar, så kan frågan om artificiella systems kapaciteter inte besvaras så entydigt. Det litet längre svaret utgår från den mer övergripande fråga som Gärdenfors ställer i boken, nämligen om var gränserna går för tänkandet hos människor, djur och artificiella system. För att besvara denna fråga, sätter Gärdenfors tänkandets evolutionära utveckling i relation till vad det är som händer i de "intelligenta" datorprogram och maskiner som omger våra liv på så gott som alla områden. Med utgångspunkt i hur tänkande har uppstått och vilka evolutionära funktioner tänkandets olika komponenter fyller, visar Gärdenfors hur människors (och djurs) förmågor att navigera i en komplex, föränderlig och ofta oförutsägbart materiell verklighet är kvalitativt annorlunda de artificiella systemens specialisering. Insikter om hur tänkande hänger ihop med att biologiska varelser har en fysisk kropp (förkroppsligad kognition) och är beroende av att interagera med den omgivande miljön och med andra individer (situationsbunden kognition) är här centralt. Hjärnan behöver alltså både kroppen och den omgivande världen för att fungera.

Detta skiljer sig radikalt från artificiella system, som existerar i avgränsade miljöer och bara reagerar på vissa typer av input. Robotar kan visserligen ses som AI-system med en kropp (förkroppsligad AI) som i vissa sammanhang interagerar med sin omgivning, men de flesta robotar utanför filmens värld interagerar inte med människor. Vissa aspekter av människors (och djurs) tänkande saknas helt hos AI-systemen, så som sinnesförmågor, känslor, motivation och inlevelseförmåga. Annat klarar AI-systemen snabbare och effektivare än människor.

Framför allt är de artificiella systemen överlägsna människan när det gäller vissa former av minneskapacitet och att identifiera mönster i stora datamängder. Gärdenfors går pedagogiskt igenom hur detta går till med olika metoder för maskininlärning. Numera är system baserade på så kallade neurala nätverk dominerande inom AI-utveckling. De artificiella "neuronerna" är tänkta att motsvara biologiska neuroner i hjärnan, vilket hänger ihop med att

AI-utvecklare inspireras av hur naturen har skapat intelligent beteende. Ett artificiellt neuralt nätverk ”lär” sig att klassificera data genom att exponeras för ett stort antal exempel. Om exemplen består av bilder på olika djur där varje pixel i bilden ger en input till nätverket, kan neuronerna i output representera till exempel ”katt”, ”hund” eller ”häst”. Men ett artificiellt neuralt nätverk lär sig inte på samma sätt som människor. Medan ett litet barn kan lära sig betydelsen av ett ord efter att ha hört det ett par gånger, behöver ett neuralt nätverk tränas med ett stort antal exempel. Denna skillnad hänger ihop med att människors (och djurs) kognitiva utveckling är kopplad till krav som evolutionen ställt. Till exempel är motivation och känslor viktiga för de kognitiva processerna hos biologiska varelser.

Kroppens strävan efter jämvikt (*homeostas*) ger upphov till motivation: när vi är hungriga vill vi äta, när vi är trötta vill vi sova. Motivationen gör att en individ kan välja beteende i olika situationer. Vi har olika behov och konkurrerande mål, och behöver därför kunna välja mellan olika handlingar och tillfälligt bortse från ett behov för att tillfredsställa ett annat. Att vilja något, skriver Gärdenfors, innebär att ha ”en medveten föreställning om ett mål och en motivation att sträva mot målet. Vilja förutsätter därför ett medvetande” (s. 40–41). AI-system och robotar kan göra val enligt vissa förutbestämda principer, men de har ingen vilja. De saknar både motivation och medvetande.

Medvetande kan från ett evolutionärt perspektiv delas in i medvetande som upplevande av sinnesförnimmelser (som att se, höra, känna smärta), något som förmodligen alla djur med ett något så när utvecklat nervsystem har, och medvetande som reflexion (självmedvetande), det vill säga förmåga att kunna reflektera över sina egna tankar och beslut. Enbart människan och i viss mån vissa kråkfåglar och de stora aporna har självmedvetande, och antagligen är det bara vi människor som har förmåga att tänka på att vi är medvetna. I interaktion med andra är medvetandet evolutionärt viktigt för att underlätta samarbete eller undvika konflikt, och insikten att andra har medvetanden är sannolikt en viktig del i utvecklandet av ett självmedvetande.

Också inlevelseförmåga är central för alla sociala aktiviteter och grundläggande för människans tänkande. För att kunna samarbeta måste vi kunna skapa gemensamma avsikter och samordna våra mål, och för att kunna planera framåt krävs förmåga att föreställa sig något som ännu inte är. Djur har viss inlevelseförmåga, men människor är överlägsna när det gäller att förstå vad andra individer vet eller tror, och att avläsa andras avsikter. Detta är förmågor som är oerhört svårt att återskapa hos AI-system och robotar, men stora språkmodeller som ChatGPT kan *simulera* inlevelseförmåga.

Artificiella system har heller inga känslor. Detta framhålls ibland som en fördel, till exempel när det gäller automatiskt beslutsfattande. Men i kontrast till föreställningar om att känslor är irrationella och något som vi bör hålla borta från förnuftigt beslutsfattande, ställer Gärdenfors den evolutionärt

baserade uppfattningen att känslor är avgörande för tänkandet. Känslor, som till exempel rädsla, ger en kroppslig reaktion anpassad till situationen, och upplevelsen av känslan ger organismen motivation till att ta hänsyn till orsakerna till känslan, till exempel att fly från en hotande fara. Utan känslor och motivation skulle det inte finnas något lärande. Kunskap i sig är inte tillräckligt. Utan känslomässigt engagemang kan vi inte värdera kunskapen, och när det gäller de kognitiva förmågor som människor har, är det mycket som saknas i de artificiella systemen, som känslor, men också sådant som värderingar, omdöme och flexibilitet.

Att frågan om AI kan tänka överhuvudtaget behöver ställas hänger ihop med uttalade strävanden inom fältet allt sedan datorvetenskapens barndom på 1950-talet att skapa intelligenta maskiner som kan tänka och agera på liknande sätt som människor. Definitioner av AI innehåller ofta jämförelser med mänskliga förmågor, och företag som Open AI säger uttryckligen att de strävar efter att bygga så kallad generell artificiell intelligens (AGI), det vill säga ett artificiellt system som avsevärt överträffar mänsklig intelligens på alla områden. Centralt i AI-utveckling har också varit att hjärnan har liknats vid en dator (på motsvarande sätt som hjärnan vid telefonens genombrott liknades vid en telefonväxel (Shulman 2008)). Frågan om huruvida AI kan tänka stimuleras också av det pärlband av titlar som *Tänkande maskiner* (Häggström 2021), *Machines that think* (Walsh 2018) eller *Thinking machines* (Takano 2021) som ständigt ser dagens ljus, liksom av det antropocentriska språk med vilket vi talar om dessa artificiella system. Vi säger att datorn "tänker" när den bearbetar data, vilket är långt ifrån den komplexa process som mänskligt tänkande innebär. Datorer "tänker" inte när de bearbetar information, och generativa språkmodeller varken "svarar" eller "hallucinerar" när presenterar en (o)sannolik följd av ord på skärmen. Framför allt ställer ordet "intelligens" till problem för förståelsen av AI-systems och robotars kapacitet.

Det finns ingen generellt accepterad definition av intelligens, och forskningen har alltmer övergett tanken på att intelligens kan mätas med ett eller ett par mått, som IQ eller emotionell intelligens (EQ). Ett sätt att se på intelligens hos människor och djur är att det handlar om anpassningsförmåga, och Gärdenfors definierar intelligens som förmågan att lösa problem som uppstår i nya situationer. De artificiella systemen har ofta svårt att komma med lösningar på typer av fall som de inte tidigare har tränats på, medan människor och många djurarter har förmåga att hantera olika omständigheter, även sådana som är nya. Förnuft och omdöme kan vara mer relevanta begrepp än intelligens för att bedöma AI-systems och robotars förmåga att resonera och lösa problem.

Förnuft brukar förknippas med rationalitet, men människans förnuft följer inte alltid de rationella idealen. Med Gärdenfors ord, så är vi "förnuftiga i förhållande till de tidiga människornas ekologiska villkor, men inte enligt teorierna om perfekt rationalitet" (s. 64). Snarare verkar det vara så att vi litar

på tumregler och nöjer oss med det som är tillräckligt bra. Omdöme (*fronesis*) innebär att på ett klokt sätt lösa problem i unika situationer genom att väga samman värderingar och kunskap i ett samspel mellan förnuft och känslor. Ett gott omdöme behövs i många situationer och kan inte ersättas av de olika regelverk som alltmer styr våra arbeten. AI-systemen har inga känslor och ingen moral och saknar därmed omdöme, men robotar som ska interagera med människor behöver någon form av omdöme. Till exempel kan en självkörande bil behöva omdöme för att avgöra om en fotgängare är på väg att gå över gatan. För att skapa artificiella system med bredare kompetens behövs därför inte i första hand programmerare och tekniker, utan kunskap om människor, samhällen och värderingar.

På ett sätt kan Gärdenfors bok ses som en kommentar till debatten om att artificiella system kan komma utveckla en generell intelligens och "ta över" världen, som länge har varit ett tema i science fiction, och som fick ett ordentligt uppsving i den allmänna debatten efter den publika lanseringen av ChatGPT hösten 2022. Flera AI-forskare uttalade sig om dessa risker. En anställd på Google hävdade att programvaran till konversationsprogrammet Lamda hade blivit medvetet, ledande AI-utvecklare skrev under ett upprop om att pausa AI-utvecklingen, och AI-världens "gudfader" och senare Nobelpristagaren Geoffrey Hinton varnade för en utveckling där maskinerna skulle bli smartare än människor. Men en förutsättning för att AI och robotar skulle kunna "ta över" är att de har något som kan liknas vid en egen vilja. Gärdenfors återkommer här till betydelsen av motivation och den strävan efter jämvikt (*homeostas*) som motiverar biologiska varelsers handlande. Datorer och robotar har vissa homeostatiska behov, som exempelvis säkrad energiförsörjning och att temperaturen håller sig inom ett visst intervall. En sådan grundläggande motivation ska sedan kopplas ihop med systemets mål, alltså vilka uppgifter som roboten eller datorprogrammet ska utföra. En autonom bil skulle kunna ändra rutten för att nå sin destination, men frågan är om en robot kan ändra sitt grundmål. I Bostroms berömda exempel med gemtillverkning¹ förutsätts att grundmålet står fast, men som Gärdenfors påpekar, så måste roboten, för att kunna uppvisa någon form av vilja, kunna överge ett mål som den upprepade gånger misslyckats med att nå.

Detta är ett exempel på hur Gärdenfors resonerar kring (o)möjligheten att artificiella system kan bli smartare än människor, men det går också att se hela boken som en kommentar till denna debatt. Som författaren visar, uppstår mänsklig intelligens i ett samspel mellan hjärnan och kroppen och kroppens

1 Bostrom's gemexempel hänvisar till ett tankeexperiment där någon programmerar och slår på en AI med målet att producera gem. AI:n har fått förmågan att lära sig, och kan uppfinna metoder för att effektivare uppnå målet. Eftersom denna AI är superintelligent, kommer den att finna metoder för att omvandla allting till gem. Och världen kommer att bli översvämmad av gem (Bostrom 2014; Hedlund & Persson 2024).

interaktion med andra individer och det omgivande samhället. Detta är ett alltför komplext system för att kunna modelleras matematiskt och reproduceras i AI-system. Det är inte de artificiella systemens "intelligens" som utgör ett hot, utan hur deras integrering i stort sett allt vi gör påverkar oss människor. En nyansering till debatten om att en supersmart AI plötsligt skulle ta över världen är att vi under inverkan från olika AI-system istället gradvis förlorar självbestämmande och inflytande över våra samhällen (Kulveit m.fl. 2025). Gärdenfors ansluter sig till denna tankegång och varnar för att vi låter oss hjärntvättas av AI-systemen när vi låter företag och politiska system styra oss genom olika former av påverkan från de artificiella systemen. AI och robotar kommer inte att ta över världen, men risken finns "att vi gör oss själva omyndiga i förhållande till dem" (s. 268).

En annan aspekt av debatten om AGI är att jämförelsen mellan artificiella system och människor (och djur) inte är helt adekvat. Befintliga specialiserade AI-system löser problem enligt helt andra principer än människor, och att försöka bygga generell artificiell intelligens som efterliknar mänskliga förmågor är fel strategi. Det är bättre, menar Gärdenfors, att fokusera på vad de artificiella systemen kan göra bättre än människor, och utveckla detta för att komplettera människornas förmåga. Det kan i detta sammanhang vara intressant att uppmärksamma att Russell och Norvigs lärobok *Artificial Intelligence: A Modern Approach* (2021), som betraktas som ett standardverk inom AI-undervisning, i sin fjärde utgåva har tagit bort jämförelsen med mänskligt tänkande ur sin definition av AI, och nu beskriver AI i termer av rationalitet utan att specificera hur denna rationalitet kommer till uttryck.

Kan AI tänka? har framför allt ett evolutionsperspektiv, men bygger på insikter från flera olika discipliner, såsom kognitionsvetenskap, psykologi och datorvetenskap, och det är inte möjligt att här göra denna mångfacetterade och tankeväckande bok rättvisa. Förhoppningen är ändå att kunna inspirera den som är intresserad av AI-systemens kapacitet i förhållande till tänkandets evolutionära utveckling att själv ta sig an denna lärda men samtidigt tillgängliga bok om "vad begrepp som 'tänka', 'intelligens' och 'medvetande' kan betyda om vi ska tillämpa dem inom AI-områdena" (s. 16).

Referenser

- Bostrom, Nick, 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.
- Hedlund, Maria & Erik Persson, 2024. "Who should obey the Laws of Robotics? A question of responsibility", s. 9-25 i Spyridon Stelios & Kostas Thelougou (red.), *The Ethics Gap in the Engineering of the Future: Moral Challenges for the Technology of Tomorrow*. Leeds: Emerald Publishing.
- Hägström, Olle, 2021. *Tänkande maskiner*. Stockholm: Fri Tanke.

- Kulveit, Jan, Douglas, Raymond, Ammann, Nora, Turan, Deger, Kreuger, David & Duvenaud, David, 2025. "Gradual disempowerment: Systemic existential risks from incremental AI development", *arXiv:2501.16946v2*.
- Shulman, Seth, 2008. *The Telephone Gambit: Chasing Alexander Graham Bell's Secret*. New York: W.W. Norton & Company.
- Takano, Shigeyuki, 2021. *Thinking Machines: Machine Learning and Its Hardware Implementations*. London: Academic Press Inc.
- Walsh, Toby, 2018. *Machines That Think: The Future of Artificial Intelligence*. New York: Prometheus.

Barrling, Katarina, 2024. *Världens mest protestantiska land. Sverige – det extrema landet lagom*. Stockholm: Bokförlaget Stolpe.

Anmälan av Antje Jackelén

Katarina Barrlings utgångspunkt är att summan av det religiösa tänkandet är konstant (218). Mot den bakgrunden är boken ett försök att förstå hur Sverige har påverkats av sin egen tolkning av kristendom och protestantism" (9). Den svenskprotestantism som vi ser idag definieras som "en form av religion där religiöst och sekulärt tänkande hänger nära samman, men där man inte ser de religiösa stråken, utan betraktar sig som fullt ut sekulär" (9–10).

World Value Surveys välkända diagram där Sverige ligger högst upp och längst ut när det gäller värdena för självförverkligande och sekulär rationalitet spelar roll, liksom David Thurfjells analyser av det gudlösa folket och granskogsfolket.¹ Men även Greta Thunberg presenteras som en perfekt företrädare för en sträng svenskprotestantism. Förenklat kan sägas att bokens tes är att en viss form av religion har präglat svensk kultur, att denna religion har präglat de sekulära politiska ideologierna och att de politiska ideologierna är vår tids religion.

Barrling menar att detta drama med nödvändighet skapar kättare, syndabocker och andra avvikare (kapitel 1) – en mekanism som förstärks av de nära

¹ David Thurfjell, *Det gudlösa folket: de postkristna svenskarna och religionen* (2015) och *Granskogsfolket: Hur naturen blev svenskarnas religion* (2020).