Hotet från medvetna maskiner

Lars Bergström

Många framstående futurologer och dataloger varnar för att utvecklingen inom området artificiell intelligens (AI) ganska snart kan leda till superintelligenta maskiner (AGI), som eventuellt tar kontroll över världen och utrotar eller förslavar oss människor. Även datalogins fader Alan Turing tycks ha tänkt i de banorna. Om varningarna är befogade bör vi nog snarast försöka vidta motåtgärder. Den svenske matematikern Olle Häggström har i sin senaste bok *Tänkande maskiner* (2021) förslag på vad som bör göras. Häggström skriver bra, medryckande och engagerat, och boken förefaller mycket initierad. Litteraturlistan omfattar runt trehundra referenser, inklusive en del filosofi.

Förkortningen AGI står för ”artificiell generell intelligens”, närmare bestämt på minst mänsklig nivå. Hittills har något i den vägen inte synts till, de mest spektakulära AI har bara varit framgångsrika på mycket speciella områden, som t. ex. schackspel, bilkörning och ansiktsigenkänning. Det råder ingen enighet bland experterna om, och i så fall när, AGI kan förväntas uppstå, men somliga tror att det kan ske redan i vårt århundrade.

Som stöd för uppfattningen om AGI:s farlighet hänvisar Häggström huvudsakligen till det han kallar Omohundro–Bostrom-teorin (s. 139). Såvitt jag förstår består den i att varje AGI har ett övergripande ”slutmål” och att den är beredd att tillgripa vilka medel som helst för att uppnå detta mål – inklusive sådant som innebär mänskligheten utplåning eller förslavning. Jag ska återkomma till detta.

En förutsättning tycks alltså vara att det uppstår maskiner som har uppnått en intelligens som motsvarar eller överskrider människans. Man kan undra om en sådan AGI också är medveten. Kan den förstå vad den gör? Kan den inte bara tänka, utan också förstå vad den tänker? Eller hur man nu ska uttrycka det. Detta är något som också har intresserat filosofer.

Häggström diskuterar detta i kapitel 9 av *Tänkande maskiner*. Han har tidigare varit inne på ämnet i flera andra sammanhang, bland annat i sin bok *Here Be Dragons* (2016). Han är benägen att godta det som brukar

Filosofisk tidskrift 2021 nr 4, 3–11

kallas ”beräkningsteorin om medvetande” (*the computational theory of mind*, förkortat CTM), vilken lite förenklat innebär att medvetande består av beräkningar eller symbolmanipulation, något som även kan produceras digitalt i en dator.

En mycket känd invändning mot CTM har konstruerats av filosofen John Searle, i ett tankeexperiment som kallas ”Det kinesiska rummet”. Häggström invänder i sin tur mot Searles resonemang, och jag ska här försöka reda ut vem av dem som har rätt.

* * *

I det kinesiska rummet sitter Searle själv i ett rum där skriftliga meddelanden på kinesiska skickas in genom ett fönster och där Searle – som inte kan kinesiska – med hjälp av instruktioner i en regelbok, och ett stort förråd av kinesiska tecken, skriver ned andra kinesiska tecken på papperslappar som han sedan skickar ut genom fönstret som svar på de budskap som skickats in. Reglerna är konstruerade av någon som kan kinesiska, så utbytet av papperslappar utgör en helt begriplig konversation. Begriplig för den som kan kinesiska, alltså.

Det enda medvetande som finns i rummet tycks vara Searles, men han kan alltså inte kinesiska. Hur kan då konversationen fungera? Ja, det svar man omedelbart tänker på är förstås att det är rummet som helhet, med alla dess ingredienser, som kan kinesiska. Detta svar kallar Searle ”systemsvaret”. De mest kända företrädarna för systemsvaret är Douglas Hofstadter och Daniel Dennett, och de har formulerat svaret i sin bok *The Mind’s I* (1981), där också Searles artikel är omtryckt.

Men Searle invänder redan i sin ursprungliga artikel mot systemsvaret. Han menar att rummet med dess hjälpmedel i princip kan elimineras genom att han själv memorerar regelboken och förrådet av tecken och därefter tillämpar reglerna med hjälp av sin egen hjärna. Då kan han alltså konversera på kinesiska, trots att han inte kan kinesiska.

På detta svarar Hofstadter och Dennett att Searle faktiskt kan kinesiska när regelboken är så bra att konversationen fungerar. Häggström ansluter sig också till systemsvaret, men han säger att han vill ge det en annan tolkning. Han hävdar att det uppstår ”en multipel personlighetsstörning” i Searle, när han memorerar regelboken.

> Måhända har Searle rätt i att han fortsatt inte kommer att begripa kinesiska, men i så fall talar han för bara den ena av de båda personligheter vi andra utifrån kan observera och samtala med och som tycks bebo hans

> kropp: å ena sidan Searle-engelsktalaren som (av allt att döma sanningsenligt) bedyrar att han bara begriper engelska och inte ett ord kinesiska, och å andra sida Searle-kinesisktalaren som obehindrat (fastän en smula långsamt) för konversationer på kinesiska. (2021, s. 235)

Ja, så kan man tänka, men jag har inte lyckats begripa hur detta är något annat än Hofstadter och Dennetts uppfattning. Även de skiljer på två personer, en som talar engelska och en kinesisktalande (som de till och med föreställer sig är en kvinna; jfr 1981, s. 377); det är alltså fråga om två personer i Searles kropp, inte en person med två personligheter. Från och med nu ska jag kalla dem Searle-E och Searle-K.

Faktum är för övrigt att något liknande redan finns hos Searle, som talar om att ”det finns i själva verket två delsystem i människan: det ena förstår engelska, det andra kinesiska” (ibid. 359, min övers.). I sin bok misstänker Häggström däremot, på grund av vad jag sagt i min recension av hans tidigare bok (Bergström 2016), att jag inte godtar existensen av Searle-K (2021, s. 236).

Den centrala frågan är förstås om Searle-K verkligen förstår vad som sägs i den kinesiska konversation som utspelar sig. Är det inte snarare så att hen bara manipulerar kinesiska tecken utan att veta vad de betyder? Den som utarbetade regelboken och kompilerade förrådet av kinesiska fraser förstår naturligtvis kinesiska, men den personen finns ju inte i Searles kropp, och inte heller i vare sig Searle-K eller Searle-E – eller ens i det ursprungliga rummet. Samma fråga kan man ställa sig när det gäller ett datorprogram som klarar ett Turingtest: förstår programmet (tillsammans med den hårdvara det körs på) det som sägs? Det är just det som Searle förnekar.

Hofstadter och Dennett tycks däremot anse att det är själva den enorma *mängden* av regler som skapar förståelse: ”nästan all förståelse måste ligga i miljarderna av tecken i regelboken” (1981, s. 375, min övers.; jfr Häggström, s. 234). Några få regler räcker förstås inte för att den som hanterar reglerna ska förstå vad tecknen betyder, men Hofstadter och Dennett tycks anse att när reglerna blir så många att konversationen fungerar, så uppstår på något sätt även en förståelse av de kinesiska teckens betydelser. Riktigt hur detta mirakel går till förklarar de dock inte.

Jag tycker det är betydligt rimligare att uppfatta situationen så att Searle-K visserligen ”kan” kinesiska i den meningen att hen kan föra en konversation på kinesiska (som i ett Turingtest), men att hen ändå

inte förstår vad konversationen handlar om eller vad de kinesiska tecknen betyder.

Jag har i ett annat sammanhang (Bergström 2021) sagt att detta bevisas av att Searle-K inte kan översätta de kinesiska tecknen i regelboken till engelska. Detta argument underkänner Häggström (på Facebook) med motiveringen att cirka en miljard kineser ju förstår kinesiska, utan att kunna översätta det som sägs på kinesiska till engelska. Det har han förstås rätt i. Och han anser tydligen att Searle-K inte kan engelska. Men om Searle-K inte kan engelska, så har han ingen nytta av regelboken. Den är nämligen skriven på engelska, eftersom den ju ska kunna begripas av Searle, som bara kan engelska (jfr Häggström 2021, s. 232).

Häggström invänder kanske att det inte är Searle-K, utan Searle-E, som tillämpar regelboken. Och att Searle-K därför inte behöver kunna engelska. Det låter i och för sig rimligt. Men om man utesluter Searle-E och dennes kunskaper i engelska från Searle-K, så finns det väl inget annat kvar av Searle-K än själva regelboken (och förrådet av fraser). Därmed är hen inte en person över huvud taget. Och en regelbok kan väl inte förstå något alls, inte ens de regler den innehåller. Det skulle nog inte ens Häggström påstå. Men han tycks vara övertygad om att Searle-K är medveten och förstår kinesiska (s. 238–41).

Searle själv använder inte det översättningsargument jag här föreslår. Han säger bara att systemet (eller Searle-K) inte vet vad konversationen handlar om, trots att den flyter på utan problem. Searle-E vet vad engelska ord refererar till, men Searle-K vet inte vad de kinesiska tecknen refererar till (1981, s. 359).

Någon skulle kanske vilja invända mot översättningsargumentet att Searle-K skulle kunna gå ut på nätet och där hitta ett översättningsprogram, med vars hjälp hen kan översätta de kinesiska tecknen till engelska. Men då har vi ju lämnat Searles tankeexperiment. Med ett sådant kriterium på förståelse, så förstår ju även Searle själv kinesiska. Det kan inte gärna vara relevant.

Över huvud taget är det väldigt konstigt att påstå att det inom en och samma kropp, i Searles tankeexperiment, finns två olika personer. Är det inte betydligt mer korrekt att säga att det bara finns *en* person, nämligen Searle, som utan regelboken endast kan engelska, men som med hjälp av regelboken kan simulera en konversation på kinesiska – dock utan att förstå vad konversationen handlar om och vad de kinesiska tecknen betyder? Searle säger själv att ”så som situationen har beskrivits är det kinesiska delsystemet helt enkelt en avdelning av det engelska

delsystemet, en avdelning som ägnar sig åt meningslös manipulation av tecken i enlighet med regler på engelska" (1981, s. 360, min övers.).

Men Häggström anser tvärtom att vi "utifrån kan observera" två olika personer som bebor Searles kropp (2021, s. 235). Jag tror inte att andra observatörer i allmänhet skulle hålla med om det. Vid första anblicken verkar det kanske som att Searle kan både engelska och kinesiska. Men när han sedan försäkrar att han inte förstår kinesiska, utan bara konverserar på kinesiska med hjälp av en regelbok, så skulle man väl godta detta. Man skulle revidera sitt första intryck, snarare än att tro att Searles kropp även rymmer en helt annan person, förutom Searle-E. Och om man å andra sidan tror att Searle ljuger och att han faktiskt förstår kinesiska, så skulle man nog fortfarande bara observera en person, nämligen Searle.

* * *

Att det i alla händelser finns *en* person, en medveten varelse, i det kinesiska rummet gör ju att det inte är ett exempel på artificiell intelligens. Hur kan det då användas som ett argument mot CTM, som implicerar att medvetande kan produceras digitalt i en dator? Ja, tanken är väl att Searle i det kinesiska rummet i princip kan ersättas med en icke-mänsklig mekanism som tillämpar regelboken – som i så fall får antas vara skriven på någon lämplig maskinkod. En sådan mekanism i kombination med regelboken och frasförrådet skulle då i någon mening förstå kinesiska, eller bete sig som om den gjorde det. Skulle den också vara medveten?

Den tes Alan Turing driver i sin berömda artikel från 1950 tycks vara att datorer är intelligenta och kan tänka i den mån de kan konversera som människor i Turingtest. Han bemöter diverse invändningar mot denna tes, bland annat invändningen att datorer inte är medvetna. De kan, enligt denna invändning, inte känna något, som t. ex. smärta, stolthet, ilska, depression osv.

Eftersom Turing tar upp och bemöter denna invändning, som rör medvetenhet, så menar han tydligen att datorer inte bara kan tänka, utan att de också kan vara medvetna. Kanske anser han att tänkande förutsätter medvetenhet – något som i och för sig kan låta helt rimligt.

Den som hävdar att maskiner inte är medvetna kan enligt Turing inte gärna godta den "extrema" tesen att det enda sättet att *veta* att någon annan aktör – människa eller maskin – är medveten är att *vara* den aktören. Ty den tesen leder till solipsism, dvs. åsikten att endast man själv

är medveten, vilket är en orimlig åsikt. Så den som hävdar att maskiner inte är medvetna skulle säkerligen inte godta den ”extrema och solipsistiska” tesen, utan antagligen godta Turings test för att en maskin är medveten (jfr 1950, s. 60–61). Simsalabim!

Felet med Turings argument är att han inte skiljer mellan den ”extrema” tesen (som är rimlig) och solipsismen (som är orimlig). Att vi inte kan *veta* att någon annan är medveten, är ju fullt förenligt med att vi kan *tro* det. Vi tror att andra människor är medvetna, men vi tror inte att maskiner är medvetna. Och Turings test visar inte att det är något fel på den åsikten.

Men det framgår faktiskt inte helt klart om Turing verkligen tror att datorer kan vara medvetna, eller att de måste vara medvetna för att klara ett Turingtest. Searle förnekar för sin del inte att en maskin skulle kunna tänka. Han påpekar att han själv är en sorts maskin som kan tänka, men att han är inte en dator (s. 367). Poängen med Turings maskiner är däremot att de inte är människor.

Att vi inte kan *veta* att andra människor är medvetna, är i alla händelser fullt förenligt med att vi *tror* att de är det. Att vi inte vet att andra är medvetna tvingar oss alltså inte att godta solipsismen. Så Turings resonemang verkar på denna punkt helt misslyckat.[1]

* * *

Jag har försökt visa att Häggströms argument mot Searles tankeexperiment inte är övertygande. Searles tankeexperiment tycks visa att en dator i princip skulle kunna konversera på kinesiska – och alltså klara ett Turingtest – utan att förstå kinesiska och utan att vara medveten. Däremot är detta inget argument mot CTM. Det är nämligen fullt förenligt med att en dator kan vara medveten.

En mänsklig programmerare kan knappast konstruera ett program som klarar ett utdraget Turingtest, men ett neuralt nätverk skulle kanske genom maskininlärning tillämpad på massor av konversationsdata kunna göra det. Och någon kunde kanske misstänka att en sådan inlärningsprocess på något sätt skulle dra med sig medvetande och förståelse som en sorts bieffekt.

Det verkar dock inte troligt att något sådant skulle kunna inträffa om inte datorn dessutom har någon sorts perceptionsapparat, som tillåter

1. I en redaktionell kommentar till denna text har dock Olle Risberg hävdat att det är något fel med att tro något som man tror att man inte vet. Det är jag inte beredd att hålla med om.

den att uppleva andra intryck från omvärlden. Det räcker då inte med kameror, mikrofoner och andra sensorer, som ersättning för sinnesorgan. Vad som krävs är dessutom sinnesintryck, och jag har väldigt svårt att tro att sådana skulle kunna åstadkommas av programmering eller maskininlärning. Häggströms initierade redogörelse för AI-forskning innehåller inte heller någon hypotes om hur verklig varseblivning – eller ens någon digital beräkningsstruktur med motsvarande funktion – skulle kunna uppstå i en dator. Själv lutar Häggström åt att datormedvetande är möjligt, men han medger också att frågan är ”fortsatt öppen” (s. 228 och 241).

De flesta av oss tror väl att medvetande är orsakat av fysikaliska processer, vi vet bara inte hur. Det är fortfarande ett mysterium. Kanske uppstår medvetandet som en bieffekt när vi lär oss tala och på annat sätt interagerar med vår omgivning. Vi kan i så fall knappast utesluta att maskininlärning i kombination med varseblivning skulle kunna ge samma effekt, alltså medvetande. Men vi kan aldrig veta om det är så, eftersom vi inte har något objektivt test på om medvetande föreligger. Turingtest duger exempelvis inte. Vi vet inte om det känns på något särskilt sätt att vara en dator som klarar ett Turingtest.

* * *

Men spelar det egentligen någon roll om en AGI är medveten eller inte för problemet om den kan vara farlig för oss människor? Det finns ju mycket som är farligt, utan att vara medvetet. Exempelvis massförstörelsevapen av olika slag.

De som tror att en AGI kan bli farlig uttrycker sig ofta som om den har fri vilja och som om den kan vilja göra något som kan skada oss. Det kan tyckas kräva medvetande. Häggström säger att man bland olika AI kan skilja mellan *agenter* och *verktyg* (s. 157–59). Man får intrycket att han tror att agenter kan bli farligare än verktyg.

Det är klart att det finns verktyg som inte är agenter, t. ex. hammare och bilar. Men när det gäller anordningar som beter sig på något lämpligt sätt för att uppnå ett givet mål är nog distinktionen svårare att upprätthålla. Är en termostat en agent eller ett verktyg? Är en självkörande bil, ett mästarprogram i schack, en autopilot i ett flygplan eller en självgående gräsklippare en agent eller ett verktyg?

Anhängare av Omohundro–Bostrom-teorin tycks anse att en AGI som har som slutmål att tillverka gem är en agent, och potentiellt en mycket farlig agent. Den kan nämligen få för sig att optimal gemproduktion

kräver att mänskligheten röjs ur vägen. Men jag har svårt att inse varför den inte lika gärna är ett verktyg. Kan det bero på om den är medveten eller inte? Eller på om den har en ”egen vilja”, vad nu det kan betyda?

En AGI som har som slutmål att tillverka gem är väl just ett verktyg för att tillverka gem – om någon annan har givit den detta slutmål. Har den däremot på egen hand bestämt slutmålet, så kan den betraktas som en agent. Men kan verkligen en AGI bestämma sitt eget slutmål? Såvitt jag förstår innebär Omohundro–Bostrom-teorin att AGI:n inte gör det, men att den givet sitt slutmål bestämmer sig för de mest effektiva medlen att uppnå det. Då är den ingen agent, utan ett verktyg – på samma sätt som t. ex. en autopilot eller en självgående gräsklippare.

I själva verket är det kanske bara verktyg som kan ha specifika slutmål. Häggström säger själv att han inte har något slutmål och han hänvisar också till Patrik Lindenfors, som tycks ha samma åsikt (s. 148). Jag är benägen att hålla med – men det hänger till stor del på om en aktör bara har ett slutmål, vilket framställningen hos Häggström och Bostrom ger vid handen, eller om en aktör kan ha flera slutmål (som i så fall på något hittills oförklarat sätt måste balanseras mot varandra). Men även om en AGI endast är ett verktyg, så kan den förstås ändå vara väldigt farlig. Farliga verktyg är ju ingen nyhet i en värld full av bilar, droger, kärnkraftverk och vapen av olika slag.

Lindenfors tycks för övrigt ha en uppfattning om ”slutmål” enligt vilken ”våra föregångare på savannen faktiskt hade ett, nämligen maximering av antalet avkomma” (s. 148). Men om slutmål tolkas på det sättet, har nog vem som helst slutmål, nämligen – förenklat uttryckt – en konsekvens av det man gör. En sådan konsekvens behöver ju inte vara avsedd eller ens medveten. Om slutmål i stället är något man avsiktligt eftersträvar, är det förstås ingalunda säkert att det också blir en konsekvens av det man gör.

Avsikter kan tyckas förutsätta medvetenhet, men det är nog inte nödvändigt. Flygande autonoma vapen har en sorts inbyggda avsikter, utan att – såvitt vi vet – vara medvetna. Huruvida Omohundro–Bostrom-teorin räknar med avsiktliga slutmål eller slutmål av konsekvenstyp har jag inte kunnat utläsa av Häggströms framställning. Och inte heller av Bostroms bok *Superintelligens.* Men jag antar att de ska vara avsiktliga. Eftersom en AGI är ett verktyg är det någon utomstående som tilldelar den ett slutmål. Slutmålet är förstås medvetet för den som inplanterar det i en AGI, men AGI själv behöver inte vara medveten om något alls.

För min del tror jag inte att människor har någon fri vilja. Jag tror att

det vi vill och gör bestäms av fysikaliska orsaker som i sin tur har fysikaliska orsaker och så vidare. Av liknande skäl tror jag inte heller att en AGI har någon fri vilja. Men jag utesluter inte att den kan vara medveten.

Frågan är då om dess farlighet är beroende av om den är medveten. Häggström citerar en författare som anser att om en AGI är farlig, så spelar det ingen roll ur just farlighetssynpunkt om den dessutom är medveten (s. 226). Det verkar riktigt. Men hur ska man se på risken för att en AGI blir farlig? Är den risken beroende av om den är medveten?

Om människor och andra kännande biologiska varelser utrotas av AGI och ersätts av datorer, så kan detta vara mindre katastrofalt om datorerna är medvetna än om de inte är det, nämligen om de skulle ha huvudsakligen behagliga upplevelser. Skulle datorerna däremot lida svåra plågor vore katastrofen å andra sidan större. Så kan man spekulera.

Men är själva *sannolikheten* för att en superintelligens ska utrota eller förslava mänskligheten beroende av om den är medveten? Varken Bostrom eller Häggström tycks ha någon särskild uppfattning om detta. Såvitt jag förstår har dock medvetenhet ingen som helst effekt i detta avseende, vare sig positiv eller negativ.

Detta är välkommet, eftersom vi nog aldrig kommer att ha anledning att tro att vissa datorer är medvetna. Det enda skäl vi har för att tro, eller kanske snarare orsaken till att vi tror, att andra människor – och vissa djur – är medvetna är att de liknar oss så mycket. Vi tycker inte att datorer liknar oss tillräckligt mycket. De liknar snarare andra maskiner och verktyg, som vi inte tillskriver medvetande.

Litteratur

Bergström, Lars. 2016. Recension av Olle Häggström, *Here Be Dragons*. *Filosofisk tidskrift* 37, nr 4.

Bergström, Lars. 2021. ”Läsarbrev”. *Sans* 10, nr 3, s. 97.

Bostrom, Nick. 2017. *Superintelligens: Vägar, faror, strategier*. Stockholm: Fri Tanke.

Hofstadter, Douglas R. och Daniel C. Dennett, red. 1981. *The Mind’s I*. New York: Basic Books.

Häggström, Olle. 2016. *Here Be Dragons: Science, Technology and the Future of Humanity*. Oxford: Oxford University Press.

Häggström, Olle. 2021. *Tänkande maskiner: Den artificiella intelligensens genombrott*. Stockholm: Fri Tanke.

Searle, John R. (1980) 1981. ”Minds, Brains, and Programs”. I Hofstadter och Dennett, s. 53–67.

Turing, Alan. (1950) 1981. ”Computing Machinery and Intelligence”. I Hofstadter och Dennett, s. 353–73.