Recension

Here Be Dragons: Science, Technology and the Future of Humanity
Olle Häggström
Oxford University Press 2016, 278 s. ISBN 978–0–19–872354–7

En vanlig uppfattning är att framsteg inom vetenskap och teknologi är bra och ofta nyttiga. Olle Häggström, som är professor i matematisk statistik vid Chalmers i Göteborg, anser att denna uppfattning inte utan vidare bör godtas. Den kan vara farlig. Och vissa vetenskapliga framsteg kan vara mycket farliga. De drakar som förekommer i titeln på denna välskrivna och tänkvärda bok är ”existentiella risker”, dvs. möjliga faror som kan hota mänskligheten i framtiden. Inte vilka faror som helst, utan sådana som har sin grund i – och/eller förhoppningsvis kan motverkas av – vetenskapliga framsteg.

Bland de faror som Häggström tar upp finns dels sådana som kan uppstå helt oberoende vad vi gör (som t.ex. asteroider som kolliderar med jorden, dödliga pandemier och stora vulkanutbrott som leder till svavelföreningar i atmosfären som hindrar solstrålarna att falla in mot jorden) och dels sådana som beror på vad vi gör (t.ex. kärnvapenkrig, global uppvärmning, bioteknik som används till framställning av smittoämnen, genombrott för artificiell intelligens som leder till att framtida robotar förslavar eller utrotar oss, nanoteknik som kan utnyttjas till billig och okontrollerad vapenspridning, osv.) Eventuellt skulle också vissa vetenskapliga experiment, kanske liknande de i CERN, kunna ta död på oss, t.ex. genom att skapa ett svart hål som slukar hela jorden (s. 199). Han tar också upp vissa värderingsfrågor, t.ex. frågan om hur det positiva eller negativa värdet av framtida händelser eventuellt ska påverkas av hur avlägsna de är.

Att den vetenskapliga utvecklingen kan ge upphov till ”existentiella risker” har ju påpekats av andra och Häggström refererar också till en stor mängd litteratur på detta område. Hans litteraturförteckning omfattar över femhundra skrifter, huvudsakligen om naturvetenskap och teknik, men även en del filosofi. Numera finns också särskilda forskningsinstitut som specialiserar sig på detta, exempelvis ”Future of Humanity Institute” i Oxford och ”Machine Intelligence Research Institute” i Berkeley, vars ledande forskare Häggström ofta hänvisar till i boken. Men Häggström anser att vi behöver mycket mer av sådan forskning. Det är detta som är bokens huvudbudskap.

Häggström är ju statistiker och boken innehåller också flera lärorika avsnitt om statistik och statistiska resonemang. Ett kapitel innehåller allmänna synpunkter på vetenskap, teknologi och vetenskaplig metod, samt redogörelser för olika statistiska metoder som frekventism och Bayesianism. Här hävdar Häggström bl.a. att beslutsfattare behöver sannolikheter (s. 161–3); men även om detta kan vara önskvärt, så finns ju också en del idéer i den beslutsteoretiska litteraturen om hur man ska kunna fatta beslut under genuin osäkerhet, dvs. utan sannolikheter. En bra sak i sammanhanget är att han skiljer mellan det han kallar fullt rationell Bayesianism, som ställer orealistiskt och orimligt höga krav på beslutsfattare, och praktisk Bayesianism (s. 165), som är mer realistisk.

Det så kallade ”domedagsargumentet” (s. 171) är ett omdiskuterat statistiskt argument som antas visa att sannolikheten är mycket hög för att domedagen är nära. Det kan

förekomma i flera olika varianter, men alla har, som Häggström visar, ett antal oklarheter och brister. Han drar slutsatsen att man tills vidare kan strunta i detta argument. Och egentligen passar domedagsargumentet inte så bra in i denna bok, eftersom det inte handlar om någon specifik teknologisk utveckling, eller något naturligt händelseförlopp, som kan innebära risker för mänskligheten. Det är helt abstrakt-matematiskt. Men alltså tvivelaktigt.

En olycka som skulle kunna drabba mänskligheten är att utomjordiska varelser, som är betydligt mer avancerade än vi, skulle kunna anfalla jorden och helt enkelt utrota oss – kanske för att komma över våra råvaror eller för att hindra att vi med tiden skulle kunna utgöra en fara för dem (s. 221). Men finns det någon civilisation på något annat håll i universum? ”Fermiparadoxen”, eller ”Den Stora Tystnaden” är beteckningar på problemet om varför vi inte har observerat några utomjordingar – skulle det kunna tyda på att vi är ensamma i universum? Häggström ägnar ett kapitel åt att försöka uppskatta sannolikheten för att vi inte är ensamma.

Han menar också att våra försök att komma i kontakt med andra civilisationer i universum, genom att skicka ut meddelanden i rymden, via radio eller rymdsonder, är oansvariga och icke önskvärda. Vi kan inte utgå ifrån att främmande civilisationer ska vara fredliga, så det kan vara mycket farligt att göra dem uppmärksamma på att vi existerar (s. 223). Kanske bör vi också sluta med radarastronomi av samma skäl (s. 225).

Man lär sig alltså en hel del om riskabel teknikutveckling i denna bok. Men i fortsättningen ska jag här fokusera på några mer filosofiska problem som Häggström kommer in på.

MEDVETANDE

Häggström lutar åt en teori om jaget och medvetandet, som brukar kallas ”the computational theory of mind” (CTM). Han anser att CTM är minst lika plausibel som varje alternativ (s. 68). Nu finns det ju många olika varianter av CTM[1] och det framgår inte riktigt vilken av dessa som är Häggströms favorit, men det spelar kanske mindre roll i sammanhanget. Medvetandet är i alla fall enligt CTM en sorts mekanism som givet *input* (stimuli av olika slag) åstadkommer en viss *output* (beteende, eventuellt språkligt). Inte vilken input-output relation som helst, förstås, utan en som liknar eller motsvarar den som förbinder sinnesintryck och beteenden hos människor.

Man kan undra om den CTM-mekanism eller algoritm som från input framgångsrikt levererar output också har någon sorts ”inre liv”. Kan den t.ex. känna glädje, sorg och obehag? Talet om ”medvetande” (*mind*) antyder ju det, men det är långtifrån klart hur det skulle gå till. Häggström har emellertid ett argument för att datorer inte bara kan *bete* sig intelligent (simulera intelligens), utan också kan *känna* sådant som glädje och smärta. Beviset är följande. Om en av hans goda vänner, som han betraktar som människa med ett inre liv som liknar hans eget, plötsligt skulle visa sig ha en dator i stället för en hjärna bakom pannbenet, så skulle Häggström ”sannolikt” inte ändra uppfattning om att vännen har medvetande och upplevelser som han själv och andra människor (s. 68–9).

Jag tycker inte att Häggströms prognos om sin egen reaktion i detta tankeexperiment har så stort bevisvärde. Själv vet jag inte hur jag skulle reagera i en liknande situation, men jag

[1] Eller CTOM som den betecknas i boken.

misstänker att jag snarare skulle tycka att jag hade haft fel och att den jag betraktat som människa med medvetande egentligen var en konstgjord zombie. Men inte heller det har något större bevisvärde.

Ett välkänt argument mot CTM, som Häggström anser sig vederlägga, är filosofen John Searles tankeexperiment med ”det kinesiska rummet”. Detta rum innehåller Searle själv som följer nedskrivna instruktioner, som innebär att när vissa lappar med kinesiska tecken på lämnas in i rummet, så ska Searle lämna ut vissa andra lappar med kinesiska tecken. Om en kines skriver ned en fråga på en lapp och lämnar in den i rummet, så kommer det alltså, med hjälp av Searle och hans instruktioner, ut ett lämpligt svar på en annan lapp. Kinesen utanför kan på detta sätt konversera med rummet. Men Searle inne i rummet förstår inte kinesiska. Han kan inte översätta det som sägs på lapparna till sitt eget språk, engelska. Och själva rummet kan ju inte heller kinesiska.

Searle anser att detta visar att CTM-teorin inte kan vara riktig. Här finns ju en framgångsrik kombination av input och output. Man kan säga att de nedskrivna instruktionerna i rummet, plus Searle i rollen som processor, uppfyller det s.k. Turingtestet: kinesen utanför rummet kan inte avgöra om hen konverserar med en människa eller en maskin (om vi bortser från processens hastighet). Och även om Searle skulle ha lärt sig instruktionerna utantill, så kan han fortfarande inte kinesiska. Han kan bara följa de engelska instruktionerna, men helt utan att förstå vad de kinesiska tecknen på lapparna betyder.

Men Häggström anser att Searle har fel. Han menar att Searle i sitt agerande uppvisar en sorts personlighetsklyvning (s. 71). Den ena halvan av Searle ”kan” kinesiska i den meningen att han kan leverera en rad lämpliga svar; den andra halvan ”kan inte” kinesiska, ty han förstår inte vad tecknen betyder.

Jag tycker inte att detta är en rimlig analys. Searle ”kan” ju här kinesiska bara i samma mening som jag ”kan” spela schack på stormästarnivå – nämligen om jag har en stormästare bredvid mig som talar om för mig vilka drag jag ska göra. Vi skulle knappast säga att jag då har en personlighetsklyvning och att ”ett av mina jag” kan spela stormästarschack. Sanningen är ju att hela jag är en usel schackspelare, som bara följer mästarens instruktioner. Så Searles kritik av CTM är, såvitt jag kan se, rätt slående.

En idé som ibland förknippas med CTM är att medvetandet förhåller sig till hjärnan som ett program (mjukvaran) till datorn (hårdvaran). Men det är inte särskilt rimligt att säga att det är ”programmet”, dvs. instruktionerna, som kan kinesiska (respektive spela schack). Snarare är det den som har konstruerat programmet, som kan kinesiska. Liksom det är stormästaren i schack som kan spela schack, inte de instruktioner han ger mig.

UPPLADDADE PERSONER

Hur som helst, för den som godtar CTM kan det te sig principiellt möjligt att ”ladda upp” en persons medvetande i digital form på en dator. Som vi har sett anser Häggström att en dator kan vara medveten. Om han har rätt skulle en person i princip kunna kopieras och transporteras – mejlas? – till en mängd olika platser. Och Häggström tycks (med viss tvekan) mena att det då är ”man själv” som finns på alla dessa platser (s. 74–5). Och så länge en uppladdad kopia av en person finns kvar har personen inte dött.

Eftersom boken handlar om risker kan man undra vilka risker som är förknippade med uppladdade personer. Häggström nämner här att företag i framtiden kan komma att inte längre

anställa lågpresterande personer, utan i stället ladda upp stora mängder av sin nyttigaste medarbetare och anställa dem i stället (s. 78). Kanske kan det göras till ett mycket billigt pris. Antalet människor kan också antas öka våldsamt med hjälp av uppladdning. Och uppladdade personer behöver väl inte äta och dricka – och de tar kanske inte heller så stor plats.

Ett problem med detta är att det kan verka omöjligt att en och samma person finns på flera olika platser samtidigt. Anta att A och B är uppladdningar av en och samma person, som befinner sig på olika platser vid en viss tidpunkt. Om A är identisk med B, så måste A och B ha precis samma egenskaper (enligt Leibniz princip). Men om A befinner sig på Chalmers i Göteborg och B samtidigt befinner sig i New York, och inte i Göteborg, så kan de alltså inte vara identiska.[2]

INTELLIGENSEXPLOSION

Att vi kan bygga maskiner eller tekniska system som (av misstag eller avsiktligt) faktiskt råkar ta livet av oss är ju inte alls osannolikt (vätebomber, smittsamma virus, automatiska antimissilförsvar, osv). Men Häggström tror dessutom, som flera andra som intresserar sig för existentiella risker, att vi kommer att kunna bygga maskiner som är lika intelligenta eller *intelligentare* än människor – och att detta mycket väl kan komma att ske under detta sekel (s. 101). Det låter kanske inte så farligt, men Häggström med flera misstänker också att vi då inte längre kommer att kunna kontrollera dessa superintelligenta maskiner (s. 102). Och att maskinerna kanske – eller troligen – kommer att förslava eller utrota oss.

Enligt Häggström anser visserligen de flesta experter att en sådan intelligensexplosion är helt orealistisk (s. 123). Men han själv och många andra anser att risken bör tas på allvar.

Här undrar man förstås vad ”intelligens” betyder i sammanhanget. Schackprogrammet Deep Blue som besegrade världsmästaren Kasparov klarar ju av en intellektuell uppgift som nästan ingen människa klarar,[3] men vi skulle kanske ändå inte kalla det intelligent. En miniräknare är ju inte heller intelligent, trots att den löser aritmetiska problem långt bättre än de flesta människor. Varken Deep Blue eller miniräknaren klarar ju heller Turingtestet.

Häggström argumenterar, plausibelt enligt min mening, för att framgång på Turingtestet varken är en nödvändig eller tillräcklig betingelse för intelligens (s. 103). Han tycks i stället anse att allmän intelligens visar sig i förmåga att lösa många olika sorters problem snabbt och effektivt (s. 104). Det låter rimligt. Men Steven Pinker har ett definitionsförslag, citerat av Häggström, som tillför ytterligare en komponent: ”Intelligens är förmågan att använda *nya medel* (”novel means”) att uppnå ett mål” (s. 116, min kursiv). Då ligger det nära till hands att fråga sig om datorprogram har denna förmåga att hitta nya medel. Är inte deras förmågor helt fixerade så fort de är färdigskrivna? De kan visserligen ha förmågan att förbättra eller effektivisera sina förmågor – de kan lära sig av framgångar och misstag – men detta sker hela tiden inom ramen för programmets ursprungliga instruktioner. Kan t.ex. Deep Blue hitta på ”något nytt”? Den kan väl välja ett drag i en given situation som tidigare aldrig använts. Men skulle den kunna komma på något verkligt nytt, som att förgifta Kasparovs kaffe eller att ta livet av alla världens alla schackspelare? Med andra ord: en nyhet som skadar människor?

---

[2] I min recension av Max Tegmarks bok *Vårt matematiska universum* har jag skisserat ett sätt att hantera ett liknande problem (se *Filosofisk tidskrifts* hemsida). Men detta förslag har inte övertygat alla läsare.

[3] Och numera är schackprogrammen såvitt jag förstår ännu duktigare.

Nu är ju Deep Blues kompetens väldigt speciell. Den superintelligens som Häggström föreställer sig ska inte vara specialiserad på något särskilt område; den ska vara en ”artificiell generell intelligens” (AGI) på minst mänsklig nivå (s. 106).

Häggström lägger vidare stor vikt vid en distinktion mellan finala och instrumentella mål. Finala mål introduceras utifrån i en AI (artificiell intelligens) eller AGI, av den som skriver programmet, och de ändras sedan inte. En AGI kan däremot själv välja instrumentella mål, som är medel att uppnå dess finala mål (s. 114–6).

Ett exempel (hämtat från Nick Bostrom) är en AGI vars finala mål är att tillverka så många gem som möjligt. (Det måste vara en rätt dum programmerare som har gett den detta finala mål.) Datorn/programmet väljer då som instrumentellt mål att förbättra sig själv, så att den bättre ska kunna förverkliga det finala målet. Och så vidare, gång på gång, så att den till sist förvandlar hela jorden, och kanske hela solsystemet eller hela universum, till gem. Mänskligheten stryker förstås med på kuppen (s. 116).

Har vanliga intelligenta människor över huvud taget några finala mål? Det verkar väldigt tvivelaktigt och i alla händelser tycks de i så fall – till skillnad från en AGI – kunna ändra sina finala mål lite hur som helst. Och även om de inte har någon riktig kontroll över sina finala mål (om de har några), så är dessa knappast konstanta över tid, som hos en AGI.

Apropå kontroll. Häggström tycker att det är ”fruktansvärt naivt” att tro att människor kan ha kontroll över en superintelligens. Många har tyckt att superintelligensen kan stoppas helt enkelt genom att man drar ur sladden till elkontakten. Men Häggström menar att superintelligensen kan förhindra detta genom att hota människan med hemska straff (s. 114–5). Det verkar inte övertygande. I och med att sladden är utdragen är ju hoten verkningslösa. (En annan sak är förstås att processen kan gå så fort att människan inte hinner upptäcka faran. Men då behövs ju inte heller några hot.) Hur som helst kan ju en fånge vara väldigt mycket intelligentare än sin fångvaktare, samtidigt som fångvaktaren kontrollerar fången.

Men kanske spelar det här roll *hur mycket* intelligentare än människor en AGI kan bli. Finns det ingen gräns? En dators beräkningskapacitet kan kanske ökas nästan obegränsat, men intelligens består knappast bara av snabbhet och minneskapacitet. Bland annat kreativitet hör också dit. Detta hänger ihop med Pinkers idé om att hitta ”nya medel” att uppnå mål. Kanske kan det också innefatta en förmåga att hitta nya finala mål? Och man kan verkligen undra om en maskin bli mer kreativ än, låt oss säga, Kurt Gödel eller Albert Einstein. I så fall, hur mycket mer? Och hur mäter man över huvud taget sådant?

En superintelligent robot består väl, enkelt uttryckt, av en elektronisk dator inklusive minne, eventuellt någon sorts sensorer och beteendeorgan, plus ett program. Vad den gör bestäms av programmet. Robotens ”finala mål” är att följa programmets instruktioner. Den kan, såvitt jag förstår, inte ta några egna initiativ – utöver vad programmet påbjuder (eller uttryckligen tillåter). När den väljer ett ”instrumentellt mål”, så beror det på att programmet bestämt det (eller möjligen att programmet bestämmer att en slumpmekanism ska välja ett av vissa bestämda alternativ). Någon särskild ”intelligens”, som innefattar förmågan att använda ”nya medel” tycks inte vara inblandad. Roboten lyder bara order.

Men alldeles oavsett om roboten är ”intelligent”, så kan man ju fråga sig om den kan löpa amok – genom att modifiera sig själv eller skapa nya robotar som i sin tur blir farliga för människor. Jag vet inte hur man ska kunna utesluta det. Men att maskiner kan vara farliga – eventuellt genom att användas på ett vårdslöst sätt – är ju ingen nyhet.

DISKONTERING

Ett intressant avsnitt i boken handlar om diskontering av framtida värden. Om vi vill värdera framtida katastrofer, så måste vi kunna jämföra en viss nivå av välfärd hos oss idag med samma grad av välfärd hos framtida människor (s. 230).

Vi kan tänka oss att vi behöver göra en viss uppoffring idag för att minska risken för en framtida katastrof. Om vi exempelvis nu sänker vår levnadsstandard från bra till halvbra, så kan vi kanske öka levnadsstandarden hos framtida generationer från hemsk (om katastrofen inträffar) till låt oss säga halvbra (om katastrofen inte inträffar). Steget från bra till halvbra kan antas vara mindre än steget från hemsk till halvbra. Men de framtida människorna är kanske dessutom väldigt många fler än vi som skulle drabbas av kostnaden för att förhindra katastrofen. Så kostnaden för oss skulle då väga mycket lätt i förhållande till deras samlade vinst. Ur moralisk synpunkt kan det då verka självklart att vi bör ta på oss kostnaden. Men om nuvärdet av den framtida vinsten diskonteras – vilket många tycks anse vara rimligt – så kanske den inte längre väger tillräckligt tungt för att uppväga vår kostnad.

Frågan är alltså om en sådan diskontering kan vara moraliskt acceptabel. Vad säger Häggström om det? Hans inställning är en smula vacklande. Å ena sidan är han motståndare till diskontering (av detta slag); å andra sidan noterar han att diskontering är oundviklig, om inte nuvarande välfärd ska minimeras till fördel för kommande generationers välfärd (s. 235). Det är lätt att förstå att han vacklar här. Problemet kan te sig olösligt. (Det är på sätt och vis analogt till problemet med hur mycket av sina resurser välbeställda människor bör överföra till dem som nu har det sämre ställt. Bör man donera allt man har, utom ett minimum som krävs för att överleva, om det krävs för att alla ska komma upp på så likvärdig nivå som möjligt?)

Diskonteringsproblemet hänger också ihop med frågan om hur viktigt det är att mänskligheten alls överlever på lång sikt. Om man tillåter diskontering av framtida välfärd, så kan det tyckas relativt oviktigt (ur vår nuvarande synvinkel) att mänskligheten överlever. Detta då under förutsättning att välfärd är det enda som räknas – vilket kanske inte är självklart.

MÄNSKLIGHETENS UNDERGÅNG

Häggström resonerar också kring Derek Parfits tes att det är *mycket* viktigare att 1 % av mänskligheten överlever än att ingen överlever (s. 237). Skillnaden mellan att alla överlever och att bara 1 % gör det är nästan försumbar i jämförelse. Tanken är förstås att om ingen överlever, så är det definitivt slut för mänskligheten, men om 1 % överlever, så kan det ge upphov till väldigt många framtida generationer, som annars aldrig skulle existera.

Frågan är då om vi på grund av ett sådant resonemang borde satsa nästan alla resurser vi kan uppbåda på att minska sannolikheten för mänsklighetens undergång. Häggström tycks anse det, men han värjer sig också mot denna slutsats (s. 241). Det kan tyckas innebära alltför stora uppoffringar för de människor som nu existerar.

Nu är ju själva huvudsyftet med Häggströms bok att peka på möjliga existentiella risker som skulle kunna leda till mänsklighetens undergång, så att man kan satsa forskningspengar på att hitta metoder att eliminera eller minimera sådana risker. Så man får väl anta att han själv anser det vara värdefullt att mänskligheten överlever – så länge som möjligt. Han

förordar också forskning om kolonisation av andra planeter för att öka sannolikheten för mänsklighetens överlevnad (s. 248).

Man kan då fråga sig: hur vet Häggström att det är önskvärt att mänskligheten överlever? Ja, han skulle väl säga att han inte *vet* det. Han tror inte på objektiva sanningar i moral eller andra värdesammanhang (s. 228). Hans argument för detta är visserligen inte särskilt starkt; han tycker inte att vi har någon empirisk evidens för en objektiv moral – som inte kan förklaras bättre utan antagandet av en objektiv moral. Men denna brist på empirisk evidens är ju precis vad han borde vänta sig, eftersom han också är anhängare av Humes lag rörande distinktionen mellan moral och empiriska fakta (s. 226). Att moraliska omdömen inte är empiriska, visar ju inte att de inte kan vara (objektivt) sanna. (Hur skulle det i så fall gå med matematikern Häggströms matematiska påståenden? Har han någon empirisk evidens för att det finns objektiva matematiska sanningar?)

Det finns kanske någon bra evolutionspsykologisk förklaring till att de flesta människor tycker det värdefullt att mänskligheten överlever, om de nu tycker det. Men det visar ju inte att det är sant. (Humes lag, igen.) Med andra ord, det visar inte att det *är* värdefullt att mänskligheten överlever. Man borde ha något *argument* för detta, men det har inte Häggström. Eller också menar han kanske att det inte behövs, eftersom de flesta människor ändå *tycker* att det är värdefullt att mänskligheten överlever – och att det är detta som är relevant för vad vi bör forska om och vad vi i övrigt bör göra. Vi borde kanske rösta om saken?

Kanske är det för mycket begärt att Häggström ska avgöra om det vore bra eller dåligt att mänskligheten överlever. Men jag gissar att han skulle hålla med om att det beror på hur eventuella överlevande generationer skulle komma att ha det. Om de skulle leva under vedervärdiga förhållanden och plågas av krig, sjukdom och fattigdom, så hade det nog varit bättre om de inte alls existerade. Ska vi utgå från att detta inte är vad som skulle hända, om mänskligheten överlever de existentiella riskerna? Jag tror inte Häggström har någon ”empirisk evidens” för ett sådant antagande. Statistiska kalkyler kan knappast hjälpa oss här.

För den som tror att mänskligheten kan komma att ersättas av robotar, och dessutom att dessa robotar har medvetande, så blir väl frågan om de i framtiden har det bättre eller sämre än vad människor skulle ha haft det. Kanske skulle de bli mycket lyckligare. De vore kanske så förnuftiga att de kunde undvika krig, terrorism och andra olyckor som människor kan ställa till med. Och med sin höga intelligens kan de kanske också ”förbättra” sig själva, så att deras förmåga att uppleva lycka blir oerhört mycket större än människors. Att människorna har gått under skulle då kanske inte vara mer katastrofalt än att neandertalarna och andra ”förmänniskor” har gått under för länge sedan. Ur hedonistisk synpunkt vore det snarare bingo!

LARS BERGSTRÖM