Recension

Liv 3.0. Att vara människa i den artificiella intelligensens tid
Max Tegmark
Översättning: Helena Sjöstrand Svenn & Gösta Svenn
Volante 2017, 462 s.
ISBN 978–91–88123–98–5

Max Tegmark är en framstående svensk fysiker, sedan länge verksam i USA. Han har tidigare publicerat en populärvetenskaplig bok med titeln *Vårt matematiska universum*, där han även kommer in på mer filosofiska frågor. Den nya boken är också bitvis filosofisk. Åtminstone amatörfilosofisk.

Uttrycket ”Liv 3.0” i titeln syftar på ett stadium i livets utveckling. Tegmark skiljer mellan tre olika stadier av liv. Det första är enkelt biologiskt liv, i det andra tillkommer inlärningsförmåga, dvs. förmåga att modifiera sin mjukvara. Det tredje, Liv 3.0, innebär dessutom en förmåga att modifiera sin hårdvara och att sålunda vara oberoende av evolutionen.

Gränserna mellan de tre stadierna är inte skarpa. Men Liv 3.0 är alltså artificiell generell intelligens (AGI) på minst mänsklig, och kanske betydligt högre, nivå. Denna nivå har man ju ännu inte uppnått, men Tegmark tycks anse att det eventuellt kan ske rätt snart och att det kräver noggranna förberedelser om vi inte ska riskera att råka riktigt illa ut. Denna bedömning är han som bekant inte ensam om. Många har varnat för en maskinell intelligensexplosion.

Tegmark skriver: ”Eftersom vi människor har lyckats härska över jordens övriga livsformer genom att vara smartare än de, är det också rimligt att vi på samma sätt skulle kunna överlistas och behärskas av en superintelligens” (s. 180). Med ett sådant analogiresonemang är det väl också lika orimligt att vi skulle kunna kontrollera superintelligensen, som att fiskarna eller hästarna skulle kunna kontrollera oss människor.

Boken börjar med en berättelse om en grupp AI-forskare, som försöker utveckla en artificiell generell intelligens. De lyckas också med detta och i hemlighet skaffar de sig sedan med hjälp av denna AGI ekonomiska resurser, politiskt inflytande och stort stöd i den allmänna opinionen, vilket gör att de till

sist kan skaffa sig världsherravälde. Deras politiska program är tilltalande, men man vet inte vad det hela kan leda till längre fram.

Tegmark är ju fysiker och mycket i boken handlar också om de rent fysikaliska betingelserna för digital intelligens och hur den i framtiden kan tänkas handskas med universum. Här bjuds man på en sorts häpnadsväckande science fiction, delvis med hänvisningar till faktiskt existerande forskning.

Man får reda på märkliga saker. Till exempel: ”Ett problem med att använda avdunstning av ett svart hål som energikälla är att såvida det svarta hålet inte är mycket mindre än en atom är det en olidligt långsam process som tar längre tid än universums nuvarande ålder och utstrålar mindre energi än ett stearinljus” (s. 284). Ett annat exempel: ”även om vi fortsätter fördubbla våra datorers förmåga vartannat år kommer det att krävas mer än två århundraden innan vi når den slutliga gränsen” (s. 90).

I kapitel 2 finns också en lärorik beskrivning av hur materia kan ordnas så att den genom att följa fysikens lagar kan minnas, utföra beräkningar och lära sig något nytt. Allt som behövs är sådant som består av kvarkar och elektroner; avancerad teknologi kan arrangera om dem till vad som helst (s. 275).

Tegmark går också igenom olika sätt på vilket livet kan ta slut, antingen genom naturliga händelser (global pandemi, asteroidnedslag, supervulkaner) eller på grund av våra beslut (ovänlig artificiell intelligens, kärnvapenkrig, klimatförstöring). På sikt går ju även jorden under, som när solen blir för het eller vår vintergata kolliderar med Andromedagalaxen. Men Tegmark beskriver också metoder för mänskligheten att kolonisera resten av universum. Om man skickar iväg små mekaniska ”frösonder” som inte behöver ta med mat och vatten, så kan de bygga upp mottagarstationer på avlägsna planeter till vilka människor sedan kan teletransporteras med ljusets hastighet (s. 302f).

I och med att viktiga samhällsfunktioner alltmer hanteras digitalt är det viktigt att systemen inte kan hackas. Det är fortfarande ett stort problem. ”I den pågående kapprustningen mellan den offensiva respektive den defensiva sidan när det gäller datasäkerhet är det än så länge inte mycket som tyder på att den defensiva sidan håller på att vinna” (s. 139).

Bland mycket annat pekar Tegmark också på att rättsväsendet antagligen kan automatiseras, så att det kan råda likhet inför lagen. Maskininlärningstekniken blir succesivt allt ”bättre på att analysera hjärndata från magnetkameror och andra hjärnsensorer för att avgöra vad en person tänker på och, i synnerhet, om denne tala sanning eller ljuger”, vilket skulle kunna ”möjliggöra snabbare rättegångar och rättvisare domar” (s. 143).

”Om superintelligensen blir verklighet kommer frestelsen att förvandlas till cyborger eller uppladdningar att vara stark” (s. 206). Cyborger är teknologiska förstärkningar av våra biologiska kroppar, uppladdningar (eller emuleringar) är uppladdningar av våra hjärnor (mjukvaran) till maskiner. Emulering är bara den mest extrema formen av cyborg. Men sådana förstärkningar är enligt Tegmark mindre troliga än att stark AGI utvecklas direkt från AI (s. 224).

Ett kapitel handlar om fördelar och nackdelar med tolv tänkbara scenarier för framtiden. En möjlighet är att någon superintelligens inte uppstår, utan att ägandet avskaffas och det råder fredlig samexistens och jämlikhet mellan människor, cyborger och uppladdningar. Om superintelligensen däremot uppstår kan den exempelvis tillåta en nyliberal ordning, vara välvillig diktator, tillåta vad som helst utom att någon skapar en ny superintelligens, maximera människornas lycka, vara människornas slav, utrota människorna, behålla dem i en sorts djurparker, styra en övervakningsstat eller återgå till ett förteknologiskt samhälle (s. 219).

En superintelligens kan utplåna oss för att den är rädd för att vi ska förstöra planeten, t.ex. genom atomvapenkrig, eller för att den ogillar vårt sätt att behandla naturen. Eller för att den är rädd för att bekämpas av människor (s. 248). Eller av svårförutsebara skäl som kan jämföras med människornas relation till elefanter: vi har ”utrotat åtta av elva elefantarter och dödat de allra flesta individerna av de återstående tre” (s. 249). Tegmark noterar att det kan gå med människorna som med hästarna: de behövs inte längre i arbetslivet, men de kan hållas vid liv med ett välfärdssystem för nöjes skull (s. 168).

Såvitt jag förstår är det framför allt tre frågor man inledningsvis borde ställa sig när det gäller superintelligens. För det första: vad är intelligens? För det andra: kan en superintelligens alls uppstå? För det tredje: vad bestämmer dess beteende?

Den första frågan besvarar Tegmark. Han definierar ”intelligens” som ”förmåga att uppnå komplexa mål” (s. 66). Det är en lite oklar och rätt underlig definition. Enligt den har tydligen även spisar och motorer intelligens – de har nämligen enligt Tegmark förmåga att uppnå mål: spisar värmer upp maten och motorer omvandlar elektricitet till rörelse (s. 345).

Är en miniräknare intelligent i Tegmarks mening? Den tycks ju ha förmåga att utföra aritmetiska operationer. Den kan t.ex. multiplicera 491 med 7368. Men det betyder ju bara att den kan *användas* för att utföra sådana uppgifter. I den meningen är alla verktyg intelligenta. En hammare är t.ex. intelligent eftersom den kan användas för att slå i spik och bygga hus. På samma sätt med spisar och

motorer. Och schackspelande dataprogram. Och på samma sätt med den AGI som AI-forskarna konstruerar i bokens inledning: den är ett *verktyg* som de använder för att uppnå *sina* mål. Men verktyg brukar vi inte betrakta som intelligenta. De är snarare mer eller mindre *användbara*.

Verktyg tycks inte ha några egna mål, men vi kan använda dem för att uppnå vissa av våra mål. Tegmark och andra tänkare som varnar för risker med superintelligenser lägger däremot mycket stor vikt vid att dessa har *egna* mål, som kan skilja sig från våra. Det är just detta antagande som ligger till grund för riskvarningarna.

Invändningen att maskiner inte kan ha egna mål bemöter Tegmark med att en värmesökande missil har ett mål (s. 55). Men är det missilen som har målet eller är det operatören? Man får väl säga att operatören har ett mål, som han eller hon inprogrammerar i missilen. Varefter missilen utför sitt uppdrag. Ungefär som spisen som värmer upp maten.

I själva verket är talet om ”mål” tvetydigt. Ett verktyg kunde ju sägas ha ett eget mål i den meningen att det kan användas för att utföra en viss sorts arbete. En hammare har då målet att slå i spikar, en miniräknare har målet att ge korrekta svar på aritmetiska frågor, en självkörande bil har målet att transportera passageraren till önskade adresser utan att krocka. Men ett sådant *generellt* mål måste skiljas från de *specifika* mål som verktyget kan användas för att uppnå i en viss situation, t.ex. att slå i en viss bestämd spik, att multiplicera 491 med 7368, eller att transportera mig till Stockholms stadshus en viss dag. Dessa specifika mål är användarens mål, de generella ”målen” är bara verktygets funktion.

Distinktionen kunde kanske uttryckas så här. En maskins generella mål, dess funktion, är det den *kan* (användas för att) göra. Dess specifika mål är vad den *ska* (eller avses) göra i en bestämd situation.

En superintelligens kan ha ett generellt mål, t.ex. att besvara de frågor vi matar in i den. Men den har rimligen inga specifika mål, inga särskilda problem som den ska eller ”vill” lösa. Det som bestämmer dess beteende i bestämda situationer är alltså de önskemål användaren matar in i den, inte dess egen förmåga eller funktion.

En superintelligens ska för övrigt inte bara vara intelligent, den ska ha vad Tegmark kallar ”universell intelligens”, vilket innebär att den ligger över ”den kritiska intelligenströskel som krävs för AI-design” (s. 72). Den ska med andra ord kunna programmera. Och konstruera hårdvara. Det kan inte en spis eller en hammare eller en värmesökande missil. Det kan inte heller ett avancerat

schackprogram eller ett program som kan förbättra sig självt genom maskininlärning.

Såvitt jag kan se har Tegmark inget argument för att AI kan uppnå universell intelligens eller för att en AGI verkligen kan uppstå. Och det finns i alla fall ”absolut ingen garanti för att vi kommer att lyckas skapa AGI under vår livstid – eller någonsin. Men det finns inte heller något vattentätt argument för att vi inte kommer att göra det” (s. 175).

Hur som helst anser Tegmark som sagt att en superintelligens har egna mål och han accepterar Nick Bostroms tes att ett systems slutmål är oberoende av dess intelligens, vilket han tolkar till att innebära att ”slutmålen för livet i vårt kosmos inte är förutbestämda, utan att vi har friheten och makten att utforma dem” (s. 368). Hur detta kan följa är för mig ett fullständigt mysterium.

Man kan också fråga sig om vi *bör* utforma superintelligensens ”slutmål” (om vi nu skulle kunna göra det). Tegmark lutar åt det han kallar ”arvsprincipen”, som säger att dagens människor har rätt att bestämma vad morgondagens aktörer (inklusive superintelligenser) ska göra, även om de senare har helt andra önskemål än vi. Men han inser också att detta är rätt problematiskt. Analogt skulle i så fall antikens människor ha rätt att bestämma vad vi nu ska göra (s. 365).

Här kommer han alltså in på mer traditionella filosofiska problem. Han skriver:

> För att programmera en vänlig AI måste vi fastställa meningen med livet. Vad är ”mening”? Vad är ”liv”? Vilket är det yttersta etiska imperativet? Med andra ord, hur ska vi göra för att skapa universums framtid? Om vi lämnar ifrån oss kontrollen till en superintelligens innan vi noga har besvarat de här frågorna, kommer dess beslut näppeligen att inbegripa oss. Det är dags att väcka liv i de klassiska filosofi- och etikdiskussionerna, för samtalet brådskar! (s. 272)

Vad gäller ”mening” anser han att utan (subjektiva) upplevelser finns det ingen mening eller något som är etiskt relevant (s. 362). Det håller jag med om.

Det leder honom till följande målsättning: ”Så det allra första målet på vår önskelista för framtiden borde vara att behålla (och förhoppningsvis vidga) biologiskt och/eller artificiellt medvetande i kosmos” (s. 417).

Det leder i sin tur till frågan om maskiner kan ha upplevelser. Kan de ha medvetande?

Medvetandet är enligt Tegmark ”ett fysiskt fenomen som känns icke-fysiskt därför att det […] har egenskaper som är oberoende av dess specifika fysiska substrat” (s. 404). Hans argument för att medvetandet verkligen är substratoberoende tycks vara att det *känns* ”icke-fysiskt” (s. 403). Det låter ju inte alls som ett bra argument. Men att medvetandet är substratoberoende tycks vara helt avgörande, när det gäller att bedöma om maskiner eller datorprogram – i likhet med biologiska organismer – kan ha medvetande.

Att medvetandet är substratoberoende är dessutom enligt Tegmark ”en logisk följd” av idén att medvetandet är ”sättet som information känns när den bearbetas på vissa sätt” (s. 404). Men här är ju den avgörande frågan vilka dessa ”vissa sätt” är. Även om information kan bearbetas på liknande sätt i en biologisk hjärna och i en icke-biologisk dator, så kanske den inte alls känns på samma sätt i de bägge fallen. Den kanske inte känns alls i datorn.

Slutligen, om vi för resonemangets skull antar att maskiner kan ha medvetande, så vore det kanske inte någon större förlust om superintelligenser utrotar mänskligheten. Medvetna aktörer finns då ändå kvar i universum – vilket möjliggör mening och värde – och eftersom de är mycket intelligentare än vi skulle de kanske inte heller ställa till så mycket elände som vi alltid tycks göra.

Att mänskligheten gör slut på sig själv med hjälp av medvetslösa maskiner är å andra sidan en påtaglig risk – en risk som vi redan har levt med under en längre tid.

LARS BERGSTRÖM