**Recension**

*Superintelligens: Vägar, faror, strategier*
Nick Bostrom
Översättning: Jim Jakobsson
Fri Tanke, 2017, 516 s.
ISBN 978–91–87513–08–4

Nick Bostrom är en framgångsrik svenskfödd filosof vid universitetet i Oxford, där han leder en avdelning som heter ”The Future of Humanity Institute”. En annan internationellt framgångsrik svensk, fysikern Max Tegmark vid MIT, anser enligt förlagsreklamen för denna bok att Bostrom är ”en av världens vassaste tänkare”.

Bostrom hävdar här att superintelligenta maskiner kan komma att utklassa – och i värsta fall utrota – biologiska människor. ”Maskiner har ett antal grundläggande fördelar som kommer att göra dem enormt överlägsna. Biologiska människor, även om de är förstärkta, kommer att bli utklassade” (s. 88).

Denna profetia låter väldigt definitiv, och rätt skrämmande, men i förordet säger Bostrom också att boken ”sannolikt är allvarligt felaktig och vilseledande” (s. 11). Så man vet inte riktigt vad man ska tro. I alla händelser menar han att en maskinell superintelligens ”skulle själv kunna vara en extremt mäktig agent, som med framgång kunde hävda sig mot det projekt som skapat det likaväl som mot resten av världen” (s. 152).

Utvecklingen av maskinintelligens kan dessutom bli explosionsartad. Denna idé om en ”intelligensexplosion” har framförts av många andra, tidigast kanske 1965 av matematikern I. J. Good, som också citeras av Bostrom. När – och om – man uppnår superintelligens kommer den vidare utvecklingen att gå mycket fort, ty då ”äger utveckling och forskning rum i de tidsskalor som kännetecknar maskinell superintelligens – kanske tusentals eller miljontals gånger snabbare än den forskning som bedrivs i en biologisk mänsklig tidsskala” (s. 276).

Enligt Bostrom kan en superintelligens innebära stora fördelar för oss människor. ”Risker som härrör från naturen – exempelvis asteroidnedslag, supervulkaner och naturliga pandemier – skulle närmast elimineras, eftersom superintelligensen kunde vidta motåtgärder mot de flesta sådana risker eller

åtminstone flytta ner dem till den icke-existentiella kategorin (till exempel genom kolonisering av rymden)”. Den skulle också ”minska risken för oavsiktlig förstörelse, inklusive risken för olyckor som är relaterade till ny teknologi” (s. 356).

Inte nog med det. ”Om revolutionen i maskinintelligens avlöper väl, skulle den superintelligens som blir resultatet nästan säkert kunna utveckla metoder för att efter behag förlänga livet för de då ännu levande människorna, och inte bara hålla dem vid liv utan återge dem hälsa och ungdomlig vitalitet och öka deras förmåga långt bortom vad vi idag ser som det mänskliga spektrumet” (s. 379).

## KONTROLLPROBLEMET

Men även om en superintelligens kan vara oss till stor hjälp – om nu Bostrom har rätt – så kan den som sagt också bli livsfarlig. Även om den håller sig till de ”slutmål” vi har bestämt för den, så kan den komma på helt oförutsedda medel att uppnå dem och dessa medel kan drabba oss. Till exempel så att superintelligensen röjer oss ur vägen för att vi hindrar dess verksamhet – t.ex. genom att hota att ”dra ur kontakten” – eller använder oss som råvara för att fabricera något den anser sig behöva (t.ex. s. 195).

Därmed ställs vi inför det som Bostrom kallar ”kontrollproblemet” (s. 202f). Hur ska vi hindra maskinintelligensen att skada oss? Isaac Asimovs tre lagar för robotar räcker inte som skydd (s. 220).

Men superintelligensen har knappast någon fri vilja, även om Bostroms formuleringar ibland kan tyda på motsatsen (som när han säger att den kan vara ”en extremt mäktig agent”). Såvitt jag förstår gör den bara det den är programmerad att göra (så länge det inte uppstår något mekaniskt fel). Den kan förstås vara programmerad att lära sig saker, så att den till sist vet och kan mycket mer än programmeraren. Och den kan förstås tänka mycket snabbare. Men den gör ändå bara det programmeraren direkt eller indirekt har beordrat den att göra. Kontrollproblemet löser man genom att se till att superintelligensens program förhindrar att den gör något dumt. Gör den något dumt, så är det alltså programmerarens fel. Ytterst är det inte superintelligensen som är livsfarlig, utan programmeraren. Alltså är det egentligen *programmeraren* – eller snarare alla AI-forskare – som måste kontrolleras! Vilket nog är omöjligt. Så även om man skulle hitta ett program som hindrar en superintelligens att åstadkomma skada, så är kontrollproblemet därmed inte löst.

Man kunde kanske tänka sig att överlåta kontrollproblemet till den väldigt smarta superintelligensen. Men Bostrom skulle antagligen invända att

superintelligensen då kan lösa problemet genom att utplåna mänskligheten. Om vi inte längre finns löper vi ju ingen risk att skadas av superintelligensen!

## GENERELL ARTIFICIELL INTELLIGENS

För att en artificiell superintelligens ska uppstå krävs två saker. För det första att man skapar en generell artificiell intelligens (AI) på minst mänsklig nivå. För det andra att en sådan artificiell intelligens sätter igång en vidare utveckling till allt högre grader av intelligens.

Kan dessa två betingelser uppfyllas?

Det är fortfarande en öppen fråga om man kan skapa generell intelligens på hög mänsklig nivå. Bostrom anser att det är möjligt och att det finns olika vägar till superintelligens, men han säger: ”Sann superintelligens (i kontrast till marginella ökningar i nuvarande intelligensnivåer) kan troligen uppnås först på AI-vägen” (s. 86).

Vi bör alltså fokusera på *artificiell* intelligens. (Observera att en maskin som lyckas imitera en människa i ett Turingtest kan vara lika ointelligent som en ointelligent människa.) Men det är oklart vad som menas med ”mänsklig nivå”. AI-experter talar om ”human-level machine intelligence, HLMI” (s. 39), men man kan ju, som Bostrom själv senare påpekar, skilja mellan t.ex. snabbhet och kvalitet (s. 89–94). Digital intelligens är ju oerhört mycket snabbare än biologisk (s. 99). Ska den ligga på ”mänsklig nivå” måste den alltså vara kvalitativt betydligt sämre, för att kompensera för överlägsenheten i snabbhet (jfr s. 118). Kort sagt: snabb men dum.

Det framgår inte särskilt tydligt i Bostroms bok vad som avses med ”intelligens” (och ordet finns inte heller med i det omfattande registret). På ett ställe definieras det som ”något i stil med förmåga till förutsägelse, planering och resonemang om mål och medel generellt” (s. 171).

Kanske kan man säga att intelligens består i förmåga att *lösa problem*. Människor är intelligentare än maskar och skalbaggar, vilket kunde betyda att vi kan lösa en väldig massa problem som de varken kan eller vill lösa. Och en superintelligens kan förhålla sig till oss som vi till maskar och skalbaggar (s. 149). Bostrom talar dunkelt om ”möjliga men oförverkligade kognitiva talanger” (s. 96), som gör det möjligt att lösa problem som ”inte kan lösas bit för bit och som kanske kräver kvalitativt nya typer av förståelse eller nya representationella ramverk som är för djupa eller för komplicerade för att dödliga varelser av nuvarande snitt ska kunna upptäcka eller använda dem på ett effektivt sätt” (s. 98).

Det är klart att vi kan skapa maskiner som löser vissa problem oerhört mycket snabbare än vi. Redan en miniräknare gör det, för att inte tala om en avancerad schackdator. Men de har inte *generell* intelligens. Det skulle kräva att de kan lösa ungefär samma problem som vi på samma områden som vi.

Men kan de också, på samma sätt som vi, vara *nyfikna*? Att de *kan* lösa problem är en sak, men *vill* de lösa problem? Och i så fall *vilka* problem?

En miniräknare kan lösa aritmetiska problem, men den är inte nyfiken. Den frågar sig ingenting. Den levererar bara lösningar på problem som vi har ställt upp. Detsamma gäller schackprogram. Gäller det även en *generell* intelligens?

Miniräknare och schackprogram kräver (utöver energitillförsel) en viss *input* – eller mer speciellt: en problemformulering eller en fråga – för att leverera en *output,* ett svar. Detsamma gäller såvitt jag förstår en artificiell generell intelligens. Liksom de program som kan klara ett Turingtest. (Det gäller nog också människor, även om många av oss tror att det är ”vi själva” som så att säga ”inifrån” producerar input till den output vi sedan uppvisar i vårt handlande.)

Bostrom menar att en generell intelligens, eller åtminstone en superintelligens, har ett *slutmål*. Som exempel nämner han bl.a. sådant som att ”producera så många gem som möjligt” (s. 195) eller att ”att göra projektets finansiär nöjd” (s. 189) eller ”att göra oss lyckliga” (s. 192). Det verkar väldigt konstigt att en generell superintelligens skulle ha ett sådant slutmål. Såvitt jag förstår kan ”slutmålet” knappast vara något annat än att lösa de problem som matas in som input i systemet. Miniräknare och schackprogram har heller inga slutmål utöver att lösa de problem som de har programmerats att lösa.

Bostrom tänker sig för övrigt också att superintelligensen är en Bayesiansk agent (s. 196), med en viss nyttofunktion och en apriori sannolikhetsfördelning över alla möjliga världar (s. 346). Kanske menar han att detta (eller enbart nyttofunktionen) är dess ”slutmål”? Och att detsamma gäller varje generell intelligens?

Men det gäller i alla fall inte människor. Även om människor skulle ha en nyttofunktion och en sannolikhetsfördelning vid varje given tidpunkt – vilket verkligen kan betvivlas – så skulle de i alla fall inte vara konstanta över tid. Vår hjärna förändras hela tiden och vi utsätts för olika stimuli (input) vid olika tillfällen.

För övrigt tror jag att om en artificiell generell intelligens har en nyttofunktion och en sannolikhetsfördelning över alla möjliga världar, så försvinner distinktionen mellan slutmål och instrumentellalla mål (dvs. medel),

som Bostrom lägger så stor vikt vid och som i sin tur ska motivera intelligensexplosion och kontrollproblem. Ty *all* ”motivation” specificeras då av dessa funktioner och allt är lika mycket ett ”slutmål”.

## INTELLIGENSEXPLOSION

Vad är det som kan sätta igång en digital intelligensexplosion? Enligt Bostrom är det att en AI får förmåga att förbättra sig själv, speciellt att den får en ”områdesspecifik talang för kodning och AI-forskning”. Den kan då komma in i en process av ”rekursiv självförbättring” (s. 54). Denna process kan bli väldigt snabb och leda fram till en ”stark superintelligens”, dvs. ”en intelligensnivå långt över den samtida mänsklighetens samlade intellektuella resurser” (s. 105).

Kort sagt, superintelligensen kan programmera mycket bättre än människor och den kan därför lösa många svåra problem som ligger långt utanför mänsklig räckvidd. Inte bara enskilda geniala människors räckvidd, utan hela mänsklighetens räckvidd.

Lösningarna på de problem som mänskligheten just nu har kunnat lösa finns, kan man kanske säga, på nätet. Bostrom säger också: ”Googles sökmotor skulle kunna beskrivas som det största AI-system som hittills konstruerats” (s. 35). Men är Googles sökmotor över huvud taget ”intelligent”? Den kan inte lösa problem som människor inte kan lösa. Den kan inte heller programmera. Den är kanske i någon mening en ”superintelligens”, men den kan knappast ”tänka själv” och den kan inte råka in i en process av ”rekursiv självförbättring” och den vill nog inte utrota oss. Bostrom själv menar att Googles sökmotor ”ligger långt under den mänskliga baslinjen för varje rimligt mått på generell intellektuell förmåga” (s. 105).

Så *varför* börjar en superintelligens på ”den mänskliga baslinjen” förbättra sig själv? Bostrom har antagligen ett svar på detta, i kapitel 4, ”En intelligensexplosions kinetik” (s. 104–127), men jag har tyvärr inte lyckats förstå vad detta svar går ut på. Program skapade av människor kan väl förväntas bli bättre och bättre, som hittills, men varför skulle ett program som ligger på den mänskliga baslinjen, om något sådant skulle bli möjligt, fortsätta *på egen hand* att skapa nya program? Det kan lära sig mer och mer, men det är inte detsamma som att skapa nya program, dvs. program som löser nya problem.

## VÄRDEFRÅGOR

Bostroms bok innehåller massor av detaljer och spekulationer. För den specialintresserade finns det mycket att hämta. Det finns information om allt

möjligt som har samband med människor, hjärnforskning, maskiner, databehandling och den allmänna teknologiska utvecklingen sedan antiken. Noter och litteraturförteckning omfattar ungefär hundra tätskrivna sidor.

Men boken tangerar också rena värdefrågor. En stark superintelligens borde kanske även ha värderingar. Men vilka? ”Det är för närvarande inte känt hur mänskliga värden kan överföras till en digital dator, även givet maskinintelligens på mänsklig nivå” (s. 320). Men dessutom är det kanske inte just *våra* värden som borde överföras. Bostrom anser tydligen att en superintelligens kunde prestera en mycket *bättre* värdelära än vad vi primitiva varelser hittills har lyckats åstadkomma (jfr s. 323f). Han tycks vara moralisk realist! Han anser att ”vi kan ha fel om moralen” och att filosofernas oenighet dessutom visar att ”de flesta filosofer måste ha fel” (s. 324). Superintelligensen är mer tillförlitlig.

Men i så fall borde vi kanske akta oss väldigt noga för att försöka överföra våra ”mänskliga värden” till en maktfullkomlig digital dator. Å andra sidan kan man undra om vi verkligen bör låta superintelligensen sköta värderingarna. Den kommer kanske fram till att livet är meningslöst och att både den själv och det som eventuellt återstår av mänskligheten därför bör utrotas för gott. Kanske inser den att mänskligheten på det hela taget är av ondo och skadlig för sig själv och för andra levande varelser. Den värderingen har nog inte Bostrom själv, men han anser ju att han kan ha fel och att han bör lita på superintelligensen. Kanske är det vår undermåliga intelligens som gör att så många av oss finner livet uthärdligt och ibland till och med trevligt. Och att vi har fått för oss att det är värdefullt att mänskligheten överlever inom överskådlig framtid. I så fall kan vi strunta i kontrollproblemet.

LARS BERGSTRÖM