Innehåll

Sparvarna och ugglan – en oavslutad fabel

Det var mitt uppe i bobyggartiden, men efter många långa dagar av hårt arbete kunde sparvarna äntligen sätta sig och vila i det varma skymningsljuset, under stillsamt småkvitter.

”Vi är alla så små och svaga. Tänk så enkelt livet skulle vara om vi hade en klok uggla som kunde hjälpa oss att bygga våra bon!”

”Ja!” inföll en annan sparv. ”Och vi kunde låta den ta hand om våra gamla och våra ungar.”

”Den kunde ge oss goda råd och hålla utkik efter den där katten som stryker omkring”, tillade en tredje.

Då tog ålderfågeln Pastus till orda: ”Låt oss sända ut spejare i alla väderstreck för att försöka hitta en övergiven uggleunge, eller kanske ett ägg. En kråkunge kunde också duga, eller en liten vessla. Det här kan bli det bästa som någonsin hänt oss, i alla fall sedan de Obegränsade Frönas Paviljong slog upp portarna på gården här intill.”

Hela flocken blev alldeles till sig, och överallt började sparvar kvittra av alla krafter.

Den enda som inte var övertygad om det kloka i detta företag var Vrångfink, en enögd sparv med vresigt humör. Han sade: ”Detta blir med all säkerhet vår undergång. Borde vi inte sätta oss in i konsten att tämja ugglor först, innan vi tar in en sådan varelse i vår krets?”

Pastus svarade: ”Det låter väldigt svårt att tämja en uggla. Det lär bli svårt nog redan att hitta ett uggleägg. Så låt oss börja med det. När vi sedan har lyckats få fram en uggla, kan vi fundera på hur vi ska hantera den där andra utmaningen.”

”Så kan man inte göra!” pep Vrångfink, men hans protest förklingade ohörd, då flocken redan hade flugit upp för att sätta Pastus plan i verket.

Bara två eller tre sparvar stannade kvar. Tillsammans började de resonera om hur en uggla skulle kunna tämjas. De insåg snart att Pastus hade rätt: det var en utomordentligt svår uppgift, särskilt när man inte hade en uggla att öva på. Men de funderade vidare, efter förmåga, hela tiden oroliga för att flocken skulle återvända med ett ägg innan de hittat en lösning på kontrollproblemet.

Det är inte känt hur berättelsen slutar, men författaren vill tillägna sin bok Vrångfink och alla hans meningsfränder.

Förord

Inuti din skalle finns något som läser den här meningen. Detta något, den mänskliga hjärnan, har vissa förmågor som hjärnorna hos andra djur saknar. Det är dessa särskilda förmågor som vi har att tacka för vår dominerande ställning på planeten. Andra djur har starkare muskler och vassare klor, men vi har smartare hjärnor. Vår blygsamma fördel i fråga om generell intelligens har lett oss att utveckla språk, teknik och en komplex social organisation. Fördelen har ökat med tiden, genom att varje generation har byggt vidare på föregångarnas resultat.

Om vi någon dag bygger maskinella hjärnor som överträffar mänskliga hjärnor i generell intelligens, skulle denna nya superintelligens kunna bli mycket mäktig. Och liksom gorillornas öde numera beror mer på oss människor än på gorillorna själva, skulle ödet för vår egen art hänga på vad den maskinella superintelligensen tar sig för.

En fördel har vi trots allt: det är vi som ska bygga den. I princip kunde vi bygga en superintelligens som värnar om mänskliga värden. Och vi skulle förvisso ha starka skäl att göra det. Men kontrollproblemet – problemet att styra vad superintelligensen gör – ser i praktiken ut att bli mycket besvärligt. Det verkar också som om vi bara får en chans. När en ovänlig superintelligens väl existerar, skulle den hindra oss från att byta ut den eller ändra dess inställningar. Vårt öde vore beseglat.

I den här boken försöker jag förstå den utmaning som utsikten till superintelligens utgör, och hur vi bäst kan hantera den. Detta kan

mycket väl vara den viktigaste och mest överväldigande utmaning som mänskligheten någonsin ställts inför. Och vare sig vi lyckas eller inte, är det sannolikt den sista utmaning vi någonsin kommer att ställas inför.

Det hör inte till argumentet i denna bok att vi står på tröskeln till ett stort genombrott inom artificiell intelligens, eller att vi med någon större precision kan förutsäga när en sådan utveckling skulle kunna äga rum. Det verkar ganska troligt att det kommer att ske någon gång under detta århundrade, men vi vet inte säkert. De inledande kapitlen diskuterar möjliga vägar och säger något om den möjliga tidpunkten. Men merparten av boken handlar om vad som händer därefter. Vi studerar kinetiken i en intelligensexplosion, superintelligensens former och förmågor och de strategiska alternativen för en superintelligent agent som uppnår ett avgörande övertag. Därefter skiftar vi fokus till kontrollproblemet, och frågar vad vi skulle kunna göra för att forma utgångsvillkoren så att resultatet blir möjligt att överleva och gynnsamt för oss. Mot slutet av boken tar vi ett steg tillbaka och begrundar den mer övergripande bild som växer fram ur våra undersökningar, och ger några förslag på vad som bör göras nu för att öka våra möjligheter att undvika en existentiell katastrof senare.

Detta har inte varit en lätt bok att skriva. Jag hoppas att den väg som nu har röjts ska göra det möjligt för andra forskare att nå fram till den nya frontlinjen snabbare och enklare, så att de med friska krafter kan bidra till att ytterligare vidga vår förståelse. (Och om den väg som har banats på sina håll är en smula knagglig och slingrig, hoppas jag att recensenter som bedömer resultatet inte underskattar hur besvärlig terrängen var i utgångsläget!)

Detta har inte varit en lätt bok att skriva: jag har försökt göra det till en lätt bok att läsa – men jag tror inte att jag har lyckats helt. Den målgrupp som föresvävade mig under arbetet var en tidigare version av mig själv, och jag försökte skriva en bok som jag själv skulle ha tyckt om att läsa. Detta kan visa sig vara en ganska smal läsekrets. Likväl tror jag att innehållet bör vara tillgängligt för många människor, om de gör sig mödan att tänka efter och motstår frestelsen att genast missförstå

varje ny idé genom att assimilera den till mest närliggande kliché i deras eget kulturella förråd. Tekniskt obevandrade läsare bör inte låta sig avskräckas av enstaka inslag av matematik eller specialterminologi, för huvudpoängen kan alltid utvinnas ur de omgivande förklaringarna. (För läsare som tvärtom vill ha *fler* tekniska detaljer finns åtskilligt att hämta i fotnoterna.[1])

Många poänger som görs i den här boken är sannolikt felaktiga.[2] Det är också troligt att det finns avgörande överväganden som jag inte har tagit hänsyn till, så att vissa av mina slutsatser – eller alla – är ogiltiga. Jag har konsekvent bemödat mig om att indikera olika nyanser och grader av osäkerhet, och därigenom belamrat texten med en uppsjö av ”möjligen”, ”skulle kunna”, ”kanske”, ”kunde mycket väl”, ”tycks”, ”sannolikt”, ”mycket troligt”, ”nästan säkert”. Varje bestämning och reservation har placerats ut noggrant och medvetet. Men dessa lokala markeringar av epistemisk anspråkslöshet räcker inte; de måste här kompletteras med ett övergripande erkännande av min ovisshet och felbarhet. Detta handlar inte om falsk blygsamhet: medan jag tror att min bok sannolikt är allvarligt felaktig och vilseledande, menar jag att de alternativa synsätt som har presenterats i litteraturen är betydligt sämre – däribland den mest utbredda uppfattningen (”nollhypotesen”), som säger att det tills vidare är riskfritt och rationellt att bortse från möjligheten av superintelligens.

KAPITEL 1

Tidigare utveckling och nuvarande kapaciteter

Vi börjar med att blicka tillbaka. Historien, sedd i sin allra största skala, verkar uppvisa en följd av tydligt skilda tillväxtlägen, vart och ett mycket snabbare än det föregående. Mot bakgrund av detta mönster har det förmodats att ännu ett (ännu snabbare) tillväxtläge kunde vara möjligt. Men vi lägger ingen större vikt vid denna observation – detta är inte en bok om ”teknologisk acceleration”, ”exponentiell tillväxt” eller de olika idéer som ibland förs samman under rubriken ”singulariteten”. I nästa steg överblickar vi den artificiella intelligensens historia. Därefter går vi igenom vilka kapaciteter som idag uppnåtts på området. Slutligen kastar vi ett öga på några nyare expertenkäter och begrundar vår okunnighet när det gäller tidtabellen för framtida framsteg.

Tillväxtlägen och ”big history”

För bara några miljoner år sedan klättrade våra förfäder fortfarande omkring i de afrikanska trädtopparna. Ur såväl geologiskt som evolutionärt tidsperspektiv var utvecklingen av *Homo sapiens* från vår senaste gemensamma förfader med de stora aporna en mycket snabb process.

Vi utvecklade upprätt hållning, motsättbara tummar och – framför allt – vissa relativt små förändringar i hjärnstorlek och neurologisk organisation som ledde till ett stort språng i kognitiv förmåga. Till följd av detta kan människor tänka abstrakt, kommunicera komplexa tankar och genom kulturen ackumulera information från generation till generation långt bättre än någon annan art på planeten.

Dessa förmågor tillät människor att utveckla allt effektivare produktionsteknologier, vilka gjorde det möjligt för våra förfäder att migrera långt bort från regnskogen och savannen. Särskilt efter jordbrukets införande och spridning ökade befolkningstätheten, tillsammans med den mänskliga befolkningens totala storlek. Fler människor innebar fler idéer; större täthet innebar att idéer kunde spridas lättare och att en del individer kunde ägna sig åt att utveckla specialiserade färdigheter. Denna utveckling ökade *tillväxttakten* i fråga om ekonomisk produktivitet och teknologisk kapacitet. Senare utveckling, relaterad till den industriella revolutionen, medförde ett andra, jämförbart omslag i tillväxttakt.

Sådana förändringar i tillväxttakt har viktiga konsekvenser. För några hundra tusen år sedan, i den tidiga mänskliga (eller hominida) förhistorien, var tillväxten så långsam att det tog omkring en miljon år för människans produktionskapacitet att öka tillräckligt för att livnära ytterligare en miljon individer på existensminimum. Omkring 5000 f.Kr., efter den agrikulturella revolutionen, hade tillväxttakten ökat till en punkt där motsvarande tillväxt tog bara två århundraden. Idag, efter den industriella revolutionen, växer världsekonomin med i genomsnitt lika mycket var nittionde minut.[1]

Redan den nuvarande tillväxttakten kommer att producera imponerande resultat om den upprätthålls under någorlunda lång tid. Om världsekonomin fortsätter att växa i samma takt som den har gjort de senaste femtio åren, kommer världen år 2050 att vara omkring 4,8 gånger rikare än idag, och år 2100 omkring 34 gånger rikare.[2]

Men tanken på en fortsatt stadig exponentiell tillväxt bleknar i jämförelse med vad som skulle hända om världen genomgick ett nytt