Ekonomiska följder av en intelligensexplosion

Olle Häggström

I samband med långsiktiga samhällsekonomiska kalkyler, exempelvis i klimatsammanhang, brukar ekonomer ansätta en positiv diskonteringsränta. En sådan motiveras vanligen, åtminstone delvis, med hänvisning till framtida ekonomisk tillväxt: eftersom framtida generationer kan väntas leva i större rikedom än vi, så har ett ekonomiskt tillskott av given storlek större betydelse för oss än för dem.[1] Den framtida ekonomiska tillväxten väntas i sin tur bli följden av fortsatt teknisk utveckling. Det sistnämnda kan alltså ses som ett mainstreamantagande inom nationalekonomin. Något ovanligare, däremot, är att precisera vad för slags teknik som i första hand väntas gå framåt på ett sätt som lyfter ekonomin. En tillförlitlig sådan precisering kan naturligtvis vara svår att göra, men är likväl väsentlig för att förstå hur ekonomin och samhället kan väntas komma att gestalta sig. Om vi som exempel tar den aktuella utveckling på IT- och telekomområdet som gör att vi via våra surfplattor och mobiltelefoner ständigt är blott några knapptryckningar bort från Internets enorma informationsarkiv, så har den givetvis haft genomgripande följder för världsekonomin även på andra vis än blott genom att få den att växa. Detsamma kan vi vänta oss av en rad framtida teknologier.

Ekonomiska analyser av specifika framtidsteknologier och deras potentiella följdverkningar är därför ett angeläget ämne, och därför välkomnar jag den amerikanske ekonomen James Millers nya bok *Singularity Rising: Surviving and Thriving in a Smarter, Richer and More Dangerous World*. Han fokuserar på de möjligheter som fortsatt datorutveckling erbjuder, särskilt inom det område som kallas artificiell intelligens (AI) och som syftar till att framställa ”verkligt intelligenta” (vad nu det betyder) robotar och datorprogram. Titelns *Singularity* syftar på den hypotetiska utveckling i vilken ett genombrott på AI-området sätter igång en kedjereaktion med en snabb följd av mer och mer intelligenta maskiner – en intelligensexplosion, eller en Singularitet som det också kommit att kallas.

Bokens undertitel är lite olyckligt vald, då den felaktigt för tankarna till det slags hjälp-till-självhjälpslitteratur som lovar läsaren ett bättre liv om bara hon eller han följer en uppsättning käcka råd från författaren. Mot slutet av boken bjuder Miller visserligen på lite grand av den varan, men på det stora hela är *Singularity Rising* en helt annan slags bok, långt mer akademisk och analytisk än vad den lätt imbecillt klingande undertiteln antyder.

Tanken bakom talet om Singulariteten är följande. Låt oss anta att vi skapat en verklig AI – ett datorprogram med högre allmänintelligens än en människa. I kraft av sin högre intelligens borde denna AI specifikt vara bättre än vi på att skapa AI, och därmed vara i stånd att skapa en ännu smartare AI, som i sin tur... och så vidare i en gränslös spiral mot allt högre intelligensnivåer.

För att matematikens singularitetsbegrepp strängt taget skall vara tillämpligt måste vi ha en kvantitet som rusar iväg mot oändligheten *på ändlig tid*. Riktigt så bokstavligt bör det inte tolkas i detta sammanhang, om inte annat så för att fysikens lagar till slut torde sätta käppar i hjulet för fortsatt acceleration. Men det finns en trevlig kalkyl som ändå låter oss föreställa oss hur en bokstavlig singularitet skulle kunna uppstå. Kalkylen bygger på Moores lag – det empiriskt grundade samband som säger att datorers hårdvaruprestanda växer exponentiellt, med en extremt kort fördubblingstid om 18 till 24 månader. Låt oss anta (en smula orealistiskt) att Moores lag kan extrapoleras obegränsat framåt, och att det dominerande dröjsmålet fram till nästa fördubbling

1 Detta formaliseras ofta med Ramseys formel, enligt vilken diskonteringsräntan r ges av $r=\eta g+\delta$, där η är riskaversionskoefficienten, g är förväntad tillväxttakt, och δ är den rena tidspreferensdiskontering som indikerar hur mycket mer vi bryr oss om vår egen välfärd jämfört med kommande generationers.

orsakas av vår trögtänkthet i att utforma nya smarta hårdvarudesigner. Låt oss vidare tänka oss att ett AI-genombrott har skett, så att världens skickligaste hårdvarudesigners nu är datorprogram och inte människor. Om vi försiktigt räknar med en fördubblingstid om 24 månader i Moores lag, så kommer denna AI efter 24 månader att kunna flytta över till en dubbelt så snabb hårdvara som den ursprungliga. För en ny fördubbling krävs ytterligare 24 månader av AI:ns *subjektiva* tid, men eftersom den går på dubbelt så snabb hårdvara så svarar dessa subjektiva 24 månader mot 12 månaders *objektiv* tid. Nästa fördubbling sker efter 6 månaders objektiv tid, och så vidare. Eftersom den geometriska serien 24+12+6+3+1½+... (med ett oändligt antal termer) summerar sig till 48, så kommer AI:n efter fyra år av objektiv tid att ha åstadkommit *ett oändligt antal fördubblingar* av sina hårdvaruprestanda.

Det finns flera goda skäl att inte bokstavligt acceptera ett sådant scenario: det är troligt att fundamentala fysikaliska lagar till slut sätter stopp för Moores lag, och det kan hända att logistiska problem (AI:n behöver ju inte blott tänka ut den nya hårdvaran, utan även tillse att den *byggs*) sätter käppar i hjulen för den accelererande utvecklingen. Men uträkningen hjälper oss ändå att förstå att saker kan komma att hända väldigt snabbt då AI-teknologin nått sitt stora genombrott.

Observera också att scenariot förutsätter att AI-forskningen till slut når fram till målet att matcha mänsklig intelligens. Detta har ofta ifrågasatts. Som forskningsområde har AI tidvis lidit stora PR-förluster sedan överoptimistiska förutsägelser kommit på skam. Ett känt exempel är hur den på sin tid ledande AI-forskaren Herbert Simon 1965 meddelade att ”inom 20 år kommer maskiner att kunna göra allt arbete som människor kan”.

Icke desto mindre är det på tok för tidigt att förklara AI-visionen död. Tvärtom finns flera förslag om hur teknologin kan föras framåt, vilka ger anledning att tro att ett genombrott så småningom kommer. Ett sådant, vilket förespråkas av Singularitetens främste popularisator Ray Kurzweil (2005, 2012), är att helt enkelt plagiera den mänskliga hjärnan. Tekniken för att scanna hjärnan går snabbt framåt, och kan till slut väntas vara så förfinad att vi helt enkelt kommer att kunna emulera hjärnan på en tillräckligt kraftfull dator, och ”verklig intelligens” därmed härbärgeras på en sådan.

Därmed är inte sagt att det är omöjligt att AI-forskningen stöter på patrull långt innan någon intelligensexplosion kan komma till stånd. Miller ägnar ett särskilt kapitel åt sådana scenarier. Min egen bedömning, efter att ha tagit del av vad de vassaste tänkarna på området har att säga om saken, är att det är en vidöppen fråga huruvida vi har att se fram emot någon superintelligens. Miller skriver mer utförligt om saken, men för den analytiskt lagde rekommenderar jag i första hand filosofen David Chalmers (2010).

AI-teknologin är inte riskfri. Den amerikanske fysikern och IT-entreprenören Steve Omohundro har utvecklat en teori för så kallade autonoma agenter, som vi kan tänka oss som datorprogram eller robotar som är tillräckligt avancerade för att deras skapare inte skall kunna förutse deras agerande i detalj. Även agenter för vilka till synes helt harmlösa mål uppställts kan komma att utveckla oönskade så kallade *instrumentella* mål (Omohundro, 2012). Ett datorprogram med målet att besegra så många starka spelare som möjligt i schack kan få för sig att (för detta syfte) förhindra att det stängs av, att tillskansa sig så mycket datorkraft (t.ex. genom att via Internet kapa andras datorer) och ekonomiska resurser som möjligt, och så vidare.

Att förutsäga vad som händer den dag Singulariteten är ett faktum är behäftat med enorma svårigheter. Det rimligaste att tänka sig är att vi människor då inte längre sitter i utvecklingens förarsäte. Miller citerar en av del ledande tänkarna på området, Eliezer Yudkowsky, som anger som ett slags default-scenario att ”AI-varelsen varken hatar dig eller älskar dig, men du är gjord av atomer som AI:n kan ha annan användning för”. Detta skrämmande scenario bör vi såklart försöka undvika, och Yudkowsky driver därför ett projekt han kallar *Friendly AI*, som handlar om att skapa

en superintelligens med drivkrafter som garanterar att den agerar på ett för oss människor och våra egna mål gynnsamt vis. Detta verkar dock extremt vanskligt och extremt svårt.

Miller medger i sin bok att världens beskaffenhet efter Singulariteten är närmast hopplös att förutsäga. Det är lätt att koka ihop dystopiska scenarier, där mänskligheten antingen drabbas av snabb utrotning eller reduceras till irrelevans på grund av våra jämförelsevis försumbara kognitiva och andra förmågor. Men också utopier har målats upp, där en super-AI som vill oss väl låter alla våra drömmar och mer därtill komma till uppfyllelse i fantastiska virtuella världar.

Mer kan måhända sägas om vägen dit. Här bjuder Miller på en rad läsvärda analyser, i vilka han på ekonomers typiska vis fokuserar på de olika aktörernas incitament i givna situationer, och vad dessa resulterar i. Ett scenario han målar upp är hur vägen ligger öppen för skapandet av en Singularitet, och att forskare har att välja mellan att gå så snabbt fram som möjligt, med stora risker, eller att satsa på ett *Friendly AI*-koncept à la Yudkowsky, vilket tar betydligt längre tid. Om vi vidare tänker oss att såväl amerikansk som kinesisk militär sysslar med detta, och att båda fruktar för den händelse att den andre hinner före, så kan en fångarnas dilemma-liknande situation uppkomma där båda parter känner sig pressade att öka tempot och därmed också riskera en Singularitet av mer katastrofalt slag. Liknande olyckliga situationer kan uppstå på en fri marknad.

Miller diskuterar också hur intelligensen, under återstoden av den pre-Singularitära eran, kan öka mer gradvis genom att vi gör ingrepp på oss själva, på farmakologisk väg, på genetisk, eller genom elektroniska implantat av olika slag i våra huvuden. Det har sedan länge pågått en IQ-ökning om cirka 3 IQ-poäng per decennium i befolkningarna i stora delar av världen (den så kallade Flynn-effekten), och med de nya teknologierna är det möjligt att denna utveckling kommer att accelerera. Vad gäller farmakologiska metoder har vi ju faktiskt redan lämnat startblocken. Miller skriver utförligt om det omfattande bruket av amfetaminbaserade prestationshöjande läkemedel bland studenter på hans eget college och annorstädes, och framställningen kryddas därtill av detaljer om hans eget bruk.

Även här tenderar analyserna att landa i kapprustningssituationer. Det kan mycket väl vara så att den enskilde känner tveksamhet inför tankar om att på artificiell väg modifiera den mänskliga naturen, men ändå väljer att anamma den nya tekniken då ju alternativet att avstå innebär att man åsamkar sig själv eller sina barn ett svagare konkurrensläge exempelvis på arbetsmarknaden. Miller tror också att regeringar kommer att stödja och subventionera nya intelligenshöjande tekniker för sina befolkningar, och drar en parallell till det omfattande statliga stöd till utbildningssektorn vi hittills sett i de flesta av världens länder.

Om man jämför med analyserna av en mer renodlad Singularitet, så är dessa delar av boken klart mer obehagliga att ta del av, i kraft av att beskriva en utveckling som ligger närmare i tiden. Med individens incitament i fokus blir Miller ofta otydlig i distinktionen mellan deskriptivt och normativt, och hans ställningstaganden framstår ibland som ogenerat rashygieniska.

Till de avsnitt som kan sägas vara särskilt magstarka, om än likväl av visst intellektuellt värde, hör hans analys av så kallade sexbotar – en utveckling som troligen inte ligger särskilt många år in i framtiden, där på ytan människolika robotar designas som ett slags uppgraderade ”uppblåsbara barbaror”. Hos vissa individer kommer denna teknik att minska intresset för att skaffa sig en mänsklig partner, menar Miller, och noterar att ”eftersom män jämfört med kvinnor i genomsnitt är mer ytliga i sina sexuella preferenser, så kommer sexbot-tekniken åtminstone initialt att avlägsna fler män än kvinnor från äktenskapsmarknaden”. Denna effekt menar han kan komma att få kinesiska makthavare att stödja sexbot-utvecklingen, som kan ses som en lösning på den demografiska obalans mellan män och kvinnor som inte minst deras ettbarnspolitik resulterat i. Å andra sidan varnar han för risken att många män kommer att bli mindre ambitiösa vad gäller

utbildning och karriär, då ett av deras incitament – partnerattraktion – för att lyckas på dessa områden försvagas.

Singularity Rising bjuder på långt fler spännande uppslag och ekonomiska analyser av radikala framtida teknikutvecklingsscenarier än jag kan redogöra för här. Som antytts finns mycket att bli provocerad av i boken. Trots vissa svackor betraktar jag den som helhet som läsvärd. Sitt största värde tror jag att den kan få om den väcker intresse bland ekonomer för radikal AI-teknik och dess möjliga konsekvenser. Hittills har intresset bland ekonomer för detta ohyggligt viktiga område varit nästan noll,[2] varför det torde finnas ypperliga möjligheter för den som känner sig manad att kavla upp skjortärmarna och ta itu med saken.

Referenser

Chalmers, D. (2010) The Singularity: a philosophical analysis, *Journal of Consciousness Studies* **17**, 7-65. http://consc.net/papers/singularity.pdf

Kurzweil, R. (2005) *The Singularity is Near: When Humans Transcend Biology*, Viking, New York.

Kurzweil, R. (2012) *How to Create a Mind: The Secret of Human Thought Revealed*, Viking, New York.

Omohundro, S. (2012) Rational artificial intelligence for the greater good, i den kommande volymen *The Singularity Hypothesis: A Scientific and Philosophical Assessment* (red. A. Eden, J. Søraker, J. Moor och E. Steinhart), Springer.
http://selfawaresystems.files.wordpress.com/2012/03/rational_ai_greater_good.pdf

2 Ett strålande undantag är Robin Hanson vid George Mason University, som på sin blogg Overcoming Bias (http://www.overcomingbias.com/) är frikostig med originella idéer och framtidsanalyser.