Vetenskap på gott och ont

Olle Häggström

1. Inledning

Min avsikt med denna text är att förklara och försvara den syn jag har på vetenskapens roll i samhället och på forskaretiken, och som genomsyrar de etiska resonemangen i min senaste bok *Here Be Dragons: Science, Technology and the Future of Humanity* (Häggström, 2016). För att tydliggöra min ståndpunkt skall jag ta spjärn mot två mer allmänt förekommande sätt att se på saken, vilka jag valt att kalla det *akademisk-romantiska* respektive det *ekonomistisk-vulgära* synsättet. Dessa skall jag skissera i Avsnitt 2. Därefter, i Avsnitt 3, förklarar jag vad jag anser fattas i dessa, nämligen insikten om att vetenskapliga framsteg inte bara kan göra världen bättre utan också sämre, något som gör att vi behöver agera med långt större framsynthet än vad som idag är legio. För att ge en smula konkretion åt diskussionen tar jag i Avsnitt 4 upp vad detta kan innebära för ett specifikt forskningsområde: artificiell intelligens. I det avslutande Avsnitt 5 återgår jag till mer allmänna resonemang om vad som bör göras.

2. Två otillräckliga synsätt

Till representant i detta avsnitt för det första av de båda synsätt – det **akademisk-romantiska** – jag här skall kritisera utser jag den häromåret bortgångne ungersk-svenske cancerforskaren och författaren Georg Klein. Jag ger honom denna otacksamma roll trots (eller kanske tack vare) att han under decennier haft ett stort inflytande på mitt tänkande, genom de essäböcker från 80-talet och framåt där han i karaktäristiskt eftertänksam stil med rikligt bruk av självbiografiska hågkomster resonerar kring frågor om bland annat vetenskapssyn, moral, kreativitet, människans utsatthet och livets mening.

I sin bok *Korpens blick* från 1998 återger Klein en brevväxling med Göran Rosenberg om forskarens ansvar för de eventuella konsekvenserna av forskningsresultaten. Rosenberg hävdar att forskaren har ett tungt sådant ansvar, med Klein håller emot, och menar att forskarens kreativitet fungerar fullt ut enbart när denne inte har något annat än det rena och okorrumperade sökandet efter sanningen för ögonen. Forskaren bör därför isolera sig från nyttoaspekter och överväganden om samhällskonsekvenser – överväganden som riskerar inte bara att få forskaren att tappa fokus, utan också stör idén om det objektiva sanningssökandet, och därmed i förlängningen kan slå mot vetenskapens själva trovärdighet. Denna tankegång har för övrigt en parallell i journalistiken och det som kommit att kallas *konsekvensneutralitet*: för att journalistiken skall vara trovärdig krävs att journalister och redaktörer inte låter beslut om vad som skall respektive inte skall publiceras styras av strategiska hänsyn till vilka politiska och andra konsekvenser en publicering kan få.[1] I sin iver att befria forskaren från ansvar för konsekvenserna av dennes forskningsresultat gör Klein en jämförelse med andra yrkesgrupper, och söker sig längre bort än till journalistiken:

> Kräver man av en slaktare att han skall fundera över köttätandets etiska berättigande medan han utför sitt yrkesarbete? Och hur ofta grubblar flygvärdinnor över bullerskador som orsakas av flygtrafiken?

Vidare ger Klein exempel där forskare knutit stora förhoppningar till hur deras framsteg skall bidra till en bättre värld, men där omgivande krafter bortom forskarens kontroll tagit dem i bruk på annat vis. Konsekvenserna blir för forskaren oförutsägbara och oöverblickbara. Underförstått: det är ingen idé

[1] Fichtelius (2016). Som läsaren snart kommer att inse har jag inte mycket till övers för en alltför långt driven konsekvensneutralitet.

för forskaren att ens tänka på sådant. Och med Kleins egna ord: ”Vem skall kontrollera lavinerna? Inte är det forskaren inte.”

Den akademisk-romantiska synen på forskning, som alltså förespråkar ett nyfikenhetsdrivet men i övrigt opartiskt sökande efter sanningen (vilken denna än må vara), okorrumperat av strategiska överväganden om konsekvenser, är den som dominerar i de akademiska kretsar jag främst rört mig i – inte minst bland matematiker. Den kan kontrasteras mot den **ekonomistisk-vulgära** syn, som är vanligare bland politiker och i någon mån bland universitetsledningar och på vissa av de mer tillämpat inriktade institutionerna på en teknisk högskola som det Chalmers vid vilket jag tjänar mitt levebröd.

Den ekonomistisk-vulgära synen förkastar det akademisk-romantiska idealet om kunskapssökande för dess egen skull, och hävdar istället att vetenskapen är till för att generera innovationer och patent åt industrin, för att på så vis elda på den ekonomiska tillväxten och landets konkurrenskraft på världsmarknaden. Detta kan låta som en grov karikatyr, men ligger mycket nära den nuvarande svenska regeringens syn på forskning, och det är lätt att hitta oförblommerade uttryck för den ekonomistisk-vulgära synen. Nästan parodiskt blir det i en intervju i tidskriften *Universitetsläraren* 2006 med dåvarande rektorn Thomas Nordström vid Högskolan i Kristianstad. Nordström frågar sig retoriskt hur det kan vara möjligt ”att Scania, Volvo, Ericsson, Ikea och Sandvik, för att nämna en handfull företag, håller världsklass när våra universitet och högskolor inte klarar av det”, och svarar själv att det universiteten behöver är ett mer renodlat nyttoperspektiv och hårdare styrning. Ty hur skulle det, som Nordström uttrycker det i ännu en retorisk fråga, ”se ut om man på Scania tillät någon tillverka mopeder, någon annan gräsklippare, en tredje brödrostar”?[2]

Om vi accepterar den ekonomistisk-vulgära synen på vetenskap så följer att vi dömer ut forskning om exempelvis dinosaurier, Big Bang, genusuppfattningar i Selma Lagerlöfs författarskap, och allt annat som inte snabbt låter sig omsättas i nya industriprodukter. För mig – liksom gissningsvis för de flesta försvarare av den akademisk-romantiska synen – räcker detta med råge för att påvisa ohållbarheten i den ekonomistisk-vulgära synen. Jag har emellertid en annan kritik som slår lika hårt mot den akademisk-romantiska som den ekonomistisk-vulgära synen, och som utgör startpunkten för nästa avsnitt.

3. Vetenskapliga framsteg kan göra världen bättre eller sämre

Det akademisk-romantiska synsättet och det ekonomistisk-vulgära förenas i den implicita ståndpunkten att det sämsta som kan hända med ett forskningsresultat är att det visar sig vara irrelevant – att det får noll citeringar, för att tala i de bibliometriska termer som på senare år dessvärre kommit att få allt mer genomträngande inflytande i universitetsvärlden. Den ståndpunkten är emellertid helt på tok, då långt värre saker kan hända. Ett forskningsresultat kan ge ringar på vattnet som gör världen sämre. En forskargrupp som lyckas med konststycket att sekvensera genomet för det virus som låg bakom spanska sjukan (den fruktansvärda influensavariant som under 1918-1920 tog död på mer än 50 miljoner människor) och sedan offentliggör det fullständiga genomet har gjort världen osäkrare och sämre.[3] På en skala som mäter ett forskningsresultats genomslag bör alltså inte den totala irrelevansen (noll genomslag) ses som skalans bottenpunkt, utan snarare som en mittpunkt, på en skala som sträcker sig från total katastrof till saliggörande genombrott.[4]

---

[2] Eliasson (2006). Se även Häggström (2006b) för min (sarkastiska) reaktion på Nordströms förslag.

[3] Detta har hänt på riktigt; se t.ex. von Bubnoff (2005), van Aken (2007) och Häggström (2016).

[4] Detta bör dock enligt min mening inte förstås som att det skulle vara en *merit* att ligga på mittpunkten, en linje som den framstående matematikern G.H. Hardy emellertid drev i sin berömda essä *A Mathematician's Apology* från 1940. Hardy hävdade att den matematik han ägnade sig åt saknande tillämpningar (något som senare visade sig felaktigt, vilket dock inte är poängen här), och därför är renare och finare än den kemi som ligger till grund för t.ex. nervgas. Många matematiker av tunnare kaliber än Hardys har i dennes efterföljd

För att ge det rätta perspektivet här, låt mig ägna ett kort stycke åt det allmänna läget för mänskligheten så här i början av det 21:a århundradet. Tillståndet i världen är i fråga om en lång rad välfärdsmått bättre än någonsin (Pinker, 2018; Rosling, Rosling och Rosling Rönnlund 2018). Samtidigt står emellertid mänskligheten under återstoden av århundradet inför enorma utmaningar inte bara när det gäller miljö- och naturresursfrågor och undvikandet av en kärnvapenapokalyps, utan också i fråga om en rad framväxande teknologier som, om de inte hanteras rätt, kan föra med sig risker så stora att mänsklighetens undergång är ett högst tänkbart scenario. Som fysikern Max Tegmark nyligen uttryckte saken:

> För första gången i vår planets 4.5 miljarder år långa historia står vi inför ett vägskäl, där det troligtvis inom vår livstid kommer att avgöras huruvida vi går under eller tar oss samman. [Harris, Goldstein och Tegmark, 2018, 1:16:40 in i bandinspelningen]

De försök som gjorts att åtminstone halvt systematiskt gå igenom de olika undergångsrisker som kan tänkas föreligga tenderar att landa i två slutsatser: för det första att sannolikheten att mänskligheten går under någon gång under innevarande århundrade är långt ifrån försumbar, och för det andra att merparten av denna sannolikhet kommer från risker som härrör från oss själva och våra teknologier, snarare än från naturliga orsaker som exempelvis asteroidnedslag (Bostrom och Cirkovic, 2008; Pamlin och Armstrong, 2015; Häggström, 2016). Som jag ser på saken står vi inför ett i hög grad okänt landskap av potentiella vetenskapliga och tekniska framsteg, varav många kan skänka mänskligheten rikedom och välgång, medan andra utgör dödliga hot mot hela vår existens.

Det är därför av största vikt att vi gör vårt bästa att orientera rätt i detta fruktbara men samtidigt väldigt farliga landskap. Såväl det akademisk-romantiska som det ekonomistisk-vulgära synsättet på vetenskap och forskaretik är liktydiga med att blint rusa rakt fram i detta minfält. De behöver därför avvisas.

Förvisso ligger det något i argumenten för journalistikens konsekvensneutralitet och den motsvarande akademisk-romantiska principen inom vetenskapen. Jag kan gott se hur sanningssökandets effektivitet och trovärdighet mår bäst av att inte störas av frågeställningar om vilka sanningar som bör respektive inte bör dras fram i ljuset. Och visst är det viktigt med sanningssökandets effektivitet och trovärdighet, men det är…

…*inte fullt så viktigt att det automatiskt övertrumfar allt annat*! Som exempel på ett annat värde som jag tycker förtjänar att beaktas, och vägas mot det akademisk-romantiska idealet, kan nämnas mänsklighetens överlevnad och möjlighet till en blomstrande framtid som om vi sköter våra kort rätt kan komma att sträcka sig över årtusenden och årmiljoner.

Enligt min (måhända något primitiva) moralsyn är varje människa skyldig att beakta konsekvenserna av sina handlingar – detta gäller även forskare precis som för journalister, slaktare och flygvärdinnor. För just forskare skulle jag vilja inskärpa följande moralregel:

Det kan aldrig vara acceptabelt att bedriva forskning vars risk att störta mänskligheten i fördärv och utplåning inte uppvägs av dess potential att skapa mänsklig blomstring och välfärd. Ej heller går det an att inleda ett forskningsprojekt utan att noggrant och uppriktigt ha övervägt denna aspekt.

När jag hävdar detta bland andra forskare får jag ett blandat mottagande. Vissa tycker att min föreslagna moralregel är självklar, medan andra håller emot, och en och annan börjar förhandla. ”En

försökt sig på samma argumentation, dock utan att tänka på att om den alls kan ses som ett försvar för offentligt finansierad matematikforskning så är dess logiska slutpunkt att matematiker är så odugliga till nyttigt arbete att man får vara nöjd om de kan sysselsättas med något helt harmlöst. Jag tror inte på den slutsatsen, och låt mig inskärpa: matematisk forskning utan uppenbar praktisk tillämpning kan mycket väl vara värd att satsa på, men då är det inte *tack vare* utan *trots* bristen på tillämpning (Häggström, 2006a).

sådan förhållningsregel", kan det exempelvis heta, "kan vi kanske ha inom tillämpad forskning, men att kräva det för grundforskning vore väl att ta i?".[5] Här är jag dock benhård, ty jag kan inte se några goda skäl för att just grundforskare (hur man nu definierar en sådan) skulle befrias från det allmänmänskliga kravet att reflektera i förväg över sina handlingar och försöka undvika sådana som gör mer skada än nytta. Jag inser givetvis att den konsekvensanalys jag här förordar är ett synnerligen svårt företag, och att vi sällan eller aldrig kan räkna med några säkra svar, men jag kan inte godta att den saken tas som ursäkt för att rycka på axlarna och inte ens försöka.

Jag vill inte bli övertolkad här. Får något år sedan, när jag talade med en reporter om dessa ting, och hen före publicering skickade mig en text för påseende, så visade det sig att hen tillskrivit mig påståendet att "all farlig forskning borde förbjudas". Det var då för väl att jag fick se och korrigera det innan det gick i tryck! Frågan om vad som eventuellt borde förbjudas och vad som kan hanteras på annat vis är besvärlig, men framför allt är jag inte beredd att döma ut all forskning som för med sig risker i fall då dessa risker uppvägs av potentialen att göra gott.

Det mest kända fall ur historien där forskare ställts inför detta slags valsituation är Manhattanprojektet, i vilket en extraordinärt talangfull grupp av fysiker under 1940-talet utvecklade atombomben. Detta genombrotts fasansfulla konsekvenser – först med bomberna över Hiroshima och Nagasaki, och sedan med hur miljarder människor under de årtionden som därefter förflutit suttit som gisslan under ett antal politiska och militära ledares dödliga grepp – kunde inte förutses i detalj av de forskare som deltog, men att de var på väg att ta fram ett ohyggligt vapen med genomgripande konsekvenser för mänskligheten var de klara över.[6] Konsekvenserna av att *inte* genomföra projektet föreföll, i det kunskapsläge som låg för handen, även de ohyggliga: man trodde att Nazityskland drev ett kärnvapenprojekt som hade kunnat ge dem världsherravälde om de hunnit först. Huruvida det var rätt eller fel i det läget att driva eller medverka i Manhattanprojektet är en knivig fråga som jag inte behöver ta ställning till här, utan nöjer mig med att hävda att det hade varit fel att gå in i projektet utan att ens reflektera över frågan. Flera av de medverkande fysikerna har efteråt frikostigt delat med sig av sina etiska överväganden; se exempelvis Dyson (1979), Feynman (1985) och Bethe (1991), samt Ottaviani and Myrick (2011). Psykologiskt särskilt intressant finner jag Richard Feynmans observation att han både inför projektet och efteråt brottades med svåra etiska dubier, men att han medan det vetenskapliga arbetet pågick blev så absorberad av detta att den etiska problematiken kom helt i skymundan – så till den grad att han inte ens lade märke till att när Tyskland i maj 1945 kapitulerade så försvann hans ursprungliga bevekelsegrund för att delta i projektet.

Vilka är då vår tids Manhattanprojekt? Vilka är de forskningsområden som vi bör hålla ett extra öga på när det gäller framtida konsekvenser för mänskligheten och därmed förknippade etiska överväganden? Som framgått är jag obenägen att dra gränser som avskärmar delar av forkningslandskapet från imperativet att göra en kritisk konsekvensanalys, men det går ändå att hitta områden som framstår som (ännu) angelägnare att utsätta för sådan jämfört med, låt oss säga, komparativa studier av litterära verk från modernismens genombrott. Hit hör en rad snabbt framväxande teknologiområden, som bio- och nanoteknik samt artificiell intelligens (AI). Det sistnämnda är ämnet för nästa avsnitt; se annars Häggström (2016) för ett bredare svep över teknologier som på gott eller ont kan väntas får genomgripande konsekvenser.

---

[5] Jag nödgas avstå från att nämna namn i detta fall, men kan avslöja att förslaget kom från en i forskningssverige högt uppsatt person.

[6] Därtill fanns problemet med huruvida den första kärnvapenprovsprängningen i öknen i New Mexico den 16 juli 1945 (blott några veckor före Hiroshima) skulle antända hela atmosfären och därmed göra slut på både mänskligheten och allt annat jordiskt liv. Det fanns teoretiska kalkyler som pekade mot att så inte skulle ske, men de medverkande fysikerna var vid tiden för provsprängningen långt ifrån eniga om att frågan var tillräckligt utredd. Se Ellsberg (2017).

4. Exemplet AI

Nya allt kraftfullare appar till våra mobiltelefoner, och en automatiseringsvåg som kan komma att revolutionera bransch efter bransch – ingen läsare kan rimligtvis ha undgått nyhetsrapporteringen kring den på senare år allt snabbare AI-utvecklingen, och de (enligt min mening berättigade) förhoppningar som ställs till hur innovationer inom AI och robotik kan komma att står för en stor andel av den förväntade ekonomiska tillväxten kommande decennium eller två. I längre perspektiv än så är möjligheterna praktiskt taget obegränsade, annat än av fysikens lagar. Men jämte de enorma möjligheterna finns också stora risker, och låt mig i detta avsnitt (delvis baserat på Häggström, 2018c) lyfta fram några av dessa.

Tidsperspektiven för de olika riskerna varierar. Till dem som står omedelbart för dörren hör vissa problem kring AI-baserad bildbehandling. Sådan är av stort värde inom filmindustrin, men att den också har en baksida stod, om inte förr, klart i slutet av 2017 då en uppsättning pornografiska videoklipp publicerades på Internet, vilka felaktigt men mycket realistiskt gav intryck av att visa några av världens mest kända skådespelerskor. De var gjorda med så kallad *face swap* – en AI teknik med hjälp av vilken en persons ansikte kan bytas mot en annans – och inte långt senare släpptes en app som låter vem som helst göra samma sak (se Jerräng, 2018). Huruvida följden blir en våg av hämndporr och andra skadliga tillämpningar återstår att se, men kan knappast uteslutas. Den optimistiskt lagde kan i och för sig tänka sig att problemet löser sjg självt då ju teknikens tillgänglighet gör att den utsatte kan hävda att det rör sig om förfalskningar, men om det är riktigt, hur går det då med videobevis i rättssystemet? Och vad händer med vår förmåga att skilja mellan *fake news* och korrekt nyhetsrapportering? Problemen hopar sig inom detta och närbelägna områden, inklusive vad förbättrad teknik för ansiktsigenkänning kan göra med vår personliga integritet.

Ett annat problem som i princip redan är över oss rör utvecklingen av militära drönare och besläktad AI-teknologi för så kallat autonoma vapen – eller, med en mindre artig term, mördarrobotar. Sommaren 2015 anslöt jag mig till tusentals andra forskare i undertecknandet av ett öppet brev med rubriken *Autonomous Weapons: An Open Letter from AI and Robotics Researchers* som pekar på riskerna med denna utveckling och uppmanar till moratorium för utveckling av autonoma vapen. Situationens allvar framgår av följande passage i brevet.

> Om någon av de stora militärmakterna väljer att satsa på utveckling av AI-vapen så uppstår oundvikligen en global kapprustning vars teknologiska slutpunkt är uppenbar: autonoma vapen kommer att bli framtidens Kalashnikovs. Till skillnad mot kärnvapen kräver inte dessa några kostsamma eller svårtillgängliga råvaror, och de kommer därför att bli rikligt förekommande och lätta för alla någorlunda stora länder att massproducera. Det blir bara en tidsfråga innan de dyker upp på den svarta marknaden och i händerna på terrorister, diktatorer som vill öka kontrollen över den egna befolkningen, krigsherrar med avsikt att genomföra etnisk rensning, etc. Autonoma vapen är idealiska för lönnmord, folkförtryck, destabilisering av nationer, och det selektiva dödandet av någon viss etnisk grupp. Vi tror av dessa skäl att en militär AI-kapprustning inte vore till gagn för mänskligheten. (Russell m.fl., 2015)

Den sista meningen är givetvis en kraftig underdrift. Jag vill i allmänhet gärna vara nyanserad och undviker helst att tvärsäkert och på otillräcklig grund döma ut viss forskning som etiskt oacceptabel, men i det här fallet vågar jag sätter ned foten: den som bidrar till AI-kapprustningen mot autonoma vapen bidrar också till att göra världen sämre.

En annan konsekvens av AI-utvecklingen som bör beaktas är dess inverkan på arbetsmarknaden. Kan den väntas leda till så kallad teknologisk arbetslöshet, i betydelsen arbetslöshet orsakad av

rationalisering till följd av tekniska framsteg? Att människors arbetsuppgifter övergår till maskiner är givetvis ingen nyhet. Vi kan ta det svenska jordbruket som ett typiskt exempel: i takt med att detta effektiviserades gick jordbrukssektorns andel av arbetskraften från cirka 75% vid mitten av 1800-talet till cirka 3% idag, men de 72% som därmed blivit över har mestadels inte hamnat i arbetslöshet utan istället gått till andra arbetsmarknadssektorer (Schermer, 2017). Huruvida detta fenomen – att arbetskraften hittar anställningar i nya sektorer i ungefär samma takt som de rationaliseras bort från gamla – kan väntas bestå är knappast självklart med tanke på ett antal nya omständigheter, som att det idag inte längre bara är fysiska och manuella arbetsuppgifter som automatiseras bort, utan även intellektuella.[7] Att vi i evighet skall kunna hitta nya områden där den mänskliga förmågan utkonkurrerar maskinernas är inte någon självklar sanning.

Att teknologisk arbetslöshet skulle vara något i grunden dåligt tycks förutsätta att lönearbete är nödvändigt för ett gott liv, vilket givetvis kan ifrågasättas. Visdiktaren Kjell Höglund gläds i sin sång *Maskinerna är våra vänner* åt att de kan ta över våra arbeten så att vi får mer tid över för konst, kultur, kärlek och förströelse. Ett samhälle med 100% arbetslöshet är i viss mening jämlikt, men hur ett sådant kan organiseras är knappast uppenbart. Och när vi väl har preciserat vart vi vill så behöver vi också ha en fungerande plan för vägen dit, vilken rimligtvis går via alla de mellanliggande arbetslöshetsnivåerna: 20%, 50%, 90%... Hur kan vi passera dessa övergångsstadier utan att de ekonomiska klyftorna vidgas ofantligt, med vidhängande risk för social instabilitet? Det här är svåra frågor som vi inte har tydliga svar på idag.

Dagens arbetslöshetssiffror kan troligtvis inte tolkas som början på en eskalerande teknologisk arbetslöshet (Alexander, 2018), men läget kan komma att förändras snabbt. Utvecklingen av autonoma fordon kan på bara ett par decenniers sikt väntas resultera i en närmast fullständig utradering av en hel arbetsmarknadssektor, och en liknande utveckling kan komma snabbt även i en rad andra sektorer (Frey och Osborne, 2013; Brynjolfsson och McAfee, 2014).

Jag har sparat AI-forskningens ultimata vision till sist – skapandet av en **artificiell generell intelligens** (AGI), det vill säga en maskin vars intelligens matchar eller överstiger människans över hela spektret av relevanta kognitiva förmågor inklusive kreativitet och förmåga att tänka utanför boxen. När – om överhuvudtaget – ett sådant genombrott är att vänta är föremål för stor oenighet bland expertisen, som fördelar sina uppskattningar relativt jämnt över hela det innevarande århundradet, och en del ännu längre bort (Müller och Bostrom, 2016; Häggström, 2016, 2018b; Tegmark, 2017). I ett sådant läge kör vi klokt i att bereda oss på alla möjligheter. Oenigheten är stor också vad gäller följderna av ett sådant genombrott, men en ganska vanlig uppfattning bland AI-futurologer är att ett AGI-genombrott snabbt kan trigga igång en accelererande självförbättringsspiral som mycket snabbt leder till en AGI med **superintelligens**, i betydelsen en intelligens som *vida* överstiger människans över hela spektret av relevanta kognitiva förmågor. Denna snabba (men än så länge hypotetiska) dynamik har omväxlande benämnts **Singularitet** och **intelligensexplosion** (Yudkowsky, 2013; Bostrom, 2014).

Redan datavetenskapens fader Alan Turing (1951) diskuterade på fullt allvar möjligheten av en framtida superintelligent maskin. Det skulle dock dröja ända till 2005 och IT-visionären Ray Kurzweils bok *The Singularity is Near* innan tanken vann någorlunda bred popularitet (förutom i science fiction-litteraturen). Kurzweil bekriver genombrottet för superintelligens som det avgörande steg som skall hjälpa oss människor att befria oss från våra skröpliga kroppar och ge oss allt vi kan önska, inklusive erövrandet av världsrymden. Därefter har diskussionen till stor del kommit att skifta karaktär – bort från Kurzweils evangeliska tonfall och mot en mer balanserad syn och ökad medvetenhet om att ett

[7] Ett ofta citerat exempel är journalistik (Marr, 2017).

AI-genombrott kan ha betydande risker, inklusive risk för total utplåning av mänskligheten (Yudkowsky, 2008; Bostrom, 2014).

Turing (1951) insåg att när väl en superintelligent AGI är på plats och vi människor alltså inte längre är de intelligentaste varelserna på vår planet så kommer vi antagligen inte längre att kunna behålla kontrollen, varför vårt öde kommer att vila i maskinernas händer.[8] Allt hänger då på vilka drivkrafter och mål maskinerna har. Det huvudsakliga sätt som föreslagits att hantera detta är att på något vis säkerställa att den första superintelligenta AGI:n har värderingar som prioriterar mänsklig välfärd och även i övrigt är i samklang med mänskliga värderingar (vad som nu menas med sådana); detta projekt döptes av Yudkowsky (2008) till **Friendly AI** men kallas numera **AI Alignment**. Det anses av flera skäl vara mycket svårt. Ett skäl är det som alla programmerare känner till: när vi programmerar tenderar vi att göra fel, och när en diskrepans föreligger mellan vad vi programmerat och vad vi egentligen menade så är det det förstnämnda som blir styrande. Ett annat skäl ligger i den grundläggande instabilitet som tycks föreligga, och som gör att till synes små avvikelser från de mål vi avser kan ha katastrofala konsekvenser (Bostrom, 2014). Det gäller också att vara noggrann med valet av mål: ett förmånligt klingande mål som ”maximera mängden välbefinnande minus lidande i världen” skulle måhända få i någon mening gynnsamma konsekvenser för universum, men skulle troligen också leda till mänsklighetens utplåning, då ju våra kroppar och hjärnor rimligtvis är mycket långt ifrån optimala när det gäller att maximera mängden hedoniskt välbefinnande per kilogram materia.

Och som om inte dessa svårigheter vore nog, så kan situationen förvärras av ett slags kapprustningsscenario. Om två eller flera intressenter (företag eller nationer) tävlar om att bli först med att skapa en superintelligens, så kommer den som väljer att försöka lösa *både* superintelligens- och AI Alignment-problemen att stå inför en svårare uppgift än den som nöjer sig med att *enbart* skapa en superintelligens, och kan därför hamna på efterkälken. Detta slags ekonomistisk-vulgära logik riskerar därmed att leda till att AI Alignment nedprioriteras (Miller, 2012).

Jag har här knappt ens skrapat på ytan till det framväxande och viktiga forskningsområdet kring konsekvenser och hantering av ett eventuellt genombrott inom AGI och superintelligens. En vanlig första reaktion för den som inte är bekant med området är att på stående fot improvisera ihop ett motargument mot att något farligt skulle kunna inträffa, och raskt låsa fast sig i uppfattningen att detta motargument är avgörande. Jag vill uppmana den läsare som till äventyrs känner en impuls i den riktningen att behärska sig, och istället med den rätta kombinationen av intellektuell öppenhet och kritiskt tänkande ta del av den spännande diskussion som pågår. Några av de vanligaste alltför frestande motargumenten diskuterar jag i Häggström (2018b), och för den som vill fördjupa sig ytterligare rekommenderar jag böckerna av Bostrom (2014) och Tegmark (2017).

5. Slutsatser

I denna uppsats har jag argumenterat för en större framsynthet i valet av vilka forskningsgenombrott vi skall sträva efter, och hävdat att bristande sådan kan få katastrofala konsekvenser. Vem har då ansvaret för denna framsynthet och att den omsätts i praktisk handling? I Avsnitt 2 och 3 betonade jag den enskilde forskarens ansvar, men jag menar också att det vore dumdristigt av samhället i övrigt att sätta sin tillit enbart till detta. Bland såväl akademisk-romantiskt som ekonomistisk-vulgärt

[8] Vissa försök att undvika denna slutsats har gjorts. Det oftast framförda är olika varianter på ”Vi kan ju alltid dra ur sladden”, men de tenderar att vara bottenlöst naiva (Häggström, 2014). Aningen mer lovande är den så kallade AI-in-a-box-tanken om att hålla maskinen isolerad från omvärlden annat än genom en snäv och noga kontrollerad kommunikationskanal, men tills vidare pekar det mesta på att den sortens lösning som mest kan fungera under en tillfällig och ganska kort övergångsperiod (Armstrong, Sandberg och Bostrom, 2012; Häggström, 2018a).

inriktade forskare är omedvetenheten om problematiken stor, liksom kreativiteten i att hitta på skäl för att slippa ta moraliskt ansvar.

I augusti 2015, inte långt efter publiceringen av det öppna brev om autonoma vapen jag citerade i Avsnitt 4, åhörde jag ett föredrag av datalogen Patrick Doherty, som berättade om sin fascinerande forskning kring AI-teknologi för drönare. Den avsedda tillämpningen var civil (lantmäteri), men eftersom brevet var så aktuellt och hade fått en del uppmärksamhet kände han sig nödgad att kommentera det, och meddelade att han ”inte har undertecknat brevet, då det ju inte finns några onda eller goda teknologier, utan bara onda och goda användningar” (citerat ur minnet och översatt från engelska). Att med den svepande tankegången undandra sig allt ansvar för huruvida ens forskning skulle kunna leda exempelvis till en situation där terrorister har tillgång till en förödande kraftfull teknologi, svår för samhället att försvara sig emot, duger enligt min mening inte alls. Jag vill knappt ens kalla det en tankegång; snarare är det att betrakta som ett simpelt slagord avsett för att slippa tänka obekväma tankar.

Ett mer oförblommerat exempel – skrämmande men på ett vis imponerande ärligt – står AI-forskaren Geoffrey Hinton för. I en intervju i *The New Yorker* (Khatchadourian, 2015) uttrycker Hinton stark pessimism inför vilka samhällskonsekvenser den teknologi kan väntas få som hans forskning syftar till att frambringa, och menar att det troliga är att den kommer att användas av politiska makthavare för att förtrycka befolkningen. På frågan om varför han då bedriver sin forskning svarar Hinton att han skulle kunna ge ”de vanliga argumenten”, men att (och nu hittar jag ingen svensk översättning med tillräcklig tonträff) ”the prospect of discovery is so *sweet*”.

Doherty och Hinton är inte några ovanliga avvikare i forskarsamhället. Ovanlig är knappast heller den Feynmanska psykologi jag återgav i Avsnitt 3. Forskare är inte mer än människor, och är därför helt enkelt inte att lita på när det gäller det konsekvensetiska tänkande jag anser vara av nöden. Så även om jag givetvis inte på minsta vis vill befria forskaren från ansvar, så behövs ett ansvarstagande även från andra parter: universitet och högskolor, teknikföretag, forskningsfinansieringsstiftelser, investerare, politiker, media och vanliga samhällsmedborgare. Detta ansvar tas inte idag, annat än högst sporadiskt.

Jag har inget färdigt svar på huruvida någon ny instans borde inrättas med särskilt ansvar för att besluta om vilken forskning som med hänsyn tagen till effekterna på mänsklighetens framtid skall få genomföras, och vilken som inte skall få genomföras; givetvis inser jag det problematiska med en sådan toppstyrning. Men även om vi bordlägger frågan om den sortens institutionella förändringar, så finns mycket att vinna på om alla de parter jag räknade upp i förra stycket, vilka var och en på ett eller annat vis har inflytande över forskningen, tog frågan om dess långsiktiga samhällskonsekvenser och eventuella risker på allvar. Givetvis är det inte lätt för alla dessa att göra välinformerade konsekvensanalyser och riskbedömningar, men om något offentligt organ inrättades med ansvar för att göra välbalanserade sammanställningar av kunskapsläget om samhällskonsekvenser av nya och framtida teknologier – exempelvis med FN:s klimatpanel IPCC (Intergovernmental Panel on Climate Change) som förebild – så skulle det kunna vara till stor hjälp. Det tror jag i alla fall. Vad vi hur som helst inte bör göra är att fortsätta ignorera problemet.

Bibliografi

van Aken, J. (2007) Ethics of reconstructing Spanish Flu: Is it wise to resurrect a deadly virus? *Heredity* **98**, 1-2.

Alexander, S. (2018) Technological unemployment: much more than you wanted to know, *Slate Star Codex*, 19 februari.

Armstrong, S., Sandberg, A. och Bostrom, N. (2012) Thinking inside the box: controlling and using an oracle AI, *Minds and Machines* **22**, 299-324.

Bethe, H. (1991) *The Road from Los Alamos*, American Institute of Physics, New York.

Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, Oxford.

Bostrom, N. och Cirkovic, M. (2008) *Global Catastrophic Risks*, Oxford University Press, Oxford.

Brynjolfsson, E. och McAfee, A. (2014) *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, W.W. Norton, New York.

von Bubnoff, A. (2005) The 1918 flu virus is resurrected, *Nature* **437**, 794-795.

Dyson, F. (1979) *Disturbing the Universe*, Harper and Row, New York.

Eliasson, P.-O. (2006) Ny strategi behövs för nyttoperspektiv, *Universitetsläraren*, no. 9.

Ellsberg, D. (2017) *The Doomsday Machine: Confessions of a Nuclear War Planner*, Bloomsbury, New York.

Feynman, R.P. (1985) *Surely You're Joking, Mr. Feynman? Adventures of a Curious Character*, W.W. Norton, New York.

Fichtelius, E. (2016) Journalister ska inte ta hänsyn till konsekvenserna, *Svenska Dagbladet*, 16 januari.

Frey, C.B. och Osborne, M. (2013) The future of employment: how susceptible are jobs to computerisation?, preprint, http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf

Hardy, G.H. (1940) *A Mathematician's Apology*, Cambridge University Press, Cambridge, UK.

Harris, S., Goldstein, R. och Tegmark, M. (2018) What is and what matters, *Waking Up Podcast*, 19 mars.

Häggström, O. (2006a) Angående attityder inom vetenskapen, *Svenska Matematikersamfundets Medlemsutskick*, 1 februari.

Häggström, O. (2006b) Ett anspråkslöst förslag rörande svensk forskning, *Tentakel,* no.7.

Häggström, O. (2014) Om Singulariteten i DN, *Häggström hävdar*, 22 juni.

Häggström, O. (2016) *Here Be Dragons: Science, Technology and the Future of Humanity*, Oxford University Press, Oxford.

Häggström, O. (2018a) Strategies for an unfriendly oracle AI with reset button, i den kommande boken *Artificial Intelligence Safety and Security* (red. R. Yampolskiy), CRC Press, Boca Raton, FL.

Häggström, O. (2018b) Remarks on artificial intelligence and rational optimism, *Should we Fear Artificial Intelligence?*, The EU Parliament's STOA Committee, Bryssel, s 19-26.

Häggström, O. (2018c) AI-utvecklingen och dess yttersta risker, i den kommande rapporten *Livet med AI*, Stiftelsen för Strategisk Forskning, Stockholm.

Jerräng, M. (2018) Deepfakes är det läskigaste på nätet just nu – och ett tydligt exempel på riskerna med AI, *ComputerSweden*, 31 januari.

Khatchadourian, R. (2015) The doomsday invention, *The New Yorker*, 23 november.

Klein, G. (1998) *Korpens blick: Essäer om vetenskap och moral*, Bonniers, Stockholm.

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

Marr, B. (2017) Another example of how artificial intelligence will transform news and journalism, *Forbes*, 18 juli.

Miller, J. (2012) *Singularity Rising: Surviving and Thriving in a Smarter, Richer, and More Dangerous World*, Benbella, Dallas, TX.

Müller, V. och Bostrom, N. (2016) Future progress in artificial intelligence: A survey of expert opinion. *Fundamental Issues of Artificial Intelligence,* Springer, Berlin*,* s 553-571.

Ottaviani J. och Myrick, L. (2011) *Feynman*, First and Second, New York.

Pamlin, D. och Armstrong, S. (2015) *12 Risks That Threaten Human Civilization*, Global Challenges Foundation, Stockholm.

Pinker, S. (2018) *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress*, Viking, New York.

Rosling, H., Rosling, O. och Rosling Rönnlund, A. (2018) *Factfulness: Ten Reasons We're Wrong About the World – and Why Things Are Better Than You Think*, Flatiron Books, New York.

Russell, S. m.fl. (2015) *Autonomous Weapons: An Open Letter from AI and Robotics Researchers*, Future of Life Institute.

Schermer, I.G. (2017) Strukturförändringar i sysselsättningen, *EkonomiFakta*, https://www.ekonomifakta.se/Fakta/Arbetsmarknad/Sysselsattning/Strukturforandringar-i-sysselsattningen/

Tegmark, M. (2017) *Life 3.0: Being Human in the Age of Artificial Intelligence*, Brockman Inc, New York.

Turing, A. (1951) Intelligent machinery: a heretical theory, BBC, http://philmat.oxfordjournals.org/content/4/3/256

Yudkowsky, E. (2008) Artificial intelligence as a positive and negative factor in global risk, i Bostrom och Cirkovic (2008), s 308-345.

Yudkowsky, E. (2013) *Intelligence Explosion Microeconomics*, Machine Intelligence Research Institute, Berkeley, CA.