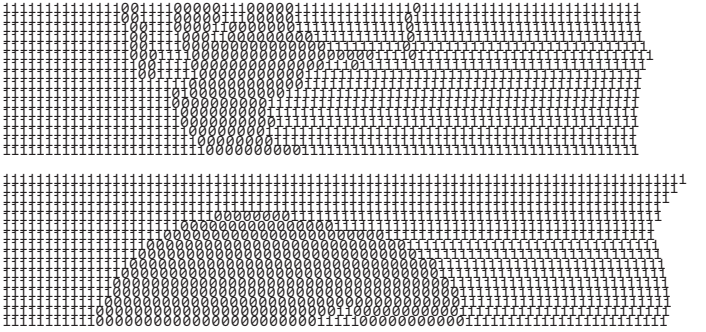
TÄNKANDE MASKINER

Olle Häggström



Innehåll

Förord till andra upplagan

DET ÄR FÖRBLUFFANDE hur mycket vatten som hinner rinna under broarna på bara två år. AI-utvecklingen har efter publiceringen våren 2021 av denna boks första upplaga skjutit iväg som en raket, och ChatGPT är blott toppen på ett isberg av revolutionerande teknikframsteg. Varje försök att idag behandla ämnet AI och tillhörande framtidsfrågor utan att beakta vad som hänt under 2021–2023 skulle framstå som inte bara ofullständigt utan rejält förlegat.

Härav denna nya upplaga, som jag avslutar med ett rejält bonuskapitel ägnat utvecklingen under sagda tvåårsperiod, och de konsekvenser denna utveckling haft på AI-debatten och på hur vägen framåt för att säkerställa en bra framtid behöver se ut. Trots nödvändigheten i denna komplettering känner jag att den text jag färdigställde för drygt två år sedan fortfarande ger en fungerande redogörelse för AI-teknikens filosofiska och ingenjörsmässiga grunder, och för dess historik fram till år 2021. Jag har därför valt att inte gå in och peta i de kapitel som utgjorde första upplagan.

Denna andra upplaga kan läsas på (minst) två sätt. Den läsare som redan tycker sig ha koll på ämnets grunder kan gå direkt på bonuskapitlet, för att endast vid behov fräscha upp minnet genom att gå tillbaka till de relevanta passager i de äldre kapitlen som jag hänvisar till i det nya kapitlet. Men det går givetvis också bra att läsa boken från början till slut. Den som anammar

detta mer konventionella sätt att läsa boken kan eventuellt ha glädje av att hålla i bakhuvudet de huvudsakliga förändringar i min syn på AI som de två senaste årens utveckling framtvingat, och som jag diskuterar i detalj i bonuskapitlet. I korthet är dessa förändringar följande:

Begreppet AGI (artificiell generell intelligens) definieras i slutet av Kapitel 1 och används därefter flitigt i efterföljande kapitel, men jag är idag betydligt mer skeptisk än tidigare till att begreppet är idealiskt för diskussion av det framtida eventuella AI-genombrott som kan väntas bli avgörande för mänsklighetens framtid. I bonuskapitlet håller jag mig därför mestadels till alternativa begrepp.

Mina tidtabeller för när detta AI-genombrott är att vänta har tidigarelagts rejält i ljuset av de senaste årens utveckling. Stora osäkerheter i tidtabellerna kvarstår, men jag tror nu till skillnad mot för ett par år sedan att vi behöver ta på allvar att den avgörande brytpunkten kan komma redan under innevarande decennium (2020-talet).

Delvis till följd av dessa accelererade tidtabeller är jag nu mer orolig än tidigare över risken att genombrottet kommer innan vi tillräckligt lyckats bemästra de säkerhetsfrågor som faller inom begreppet AI Alignment, och jag förespråkar därför idag att åtgärder vidtas för att bromsa utvecklingen – något jag var helt avvisande till i första upplagan.

* * *

Otaliga är de personer vilkas tankar jag inspirerats av i arbetet med denna nya upplaga. Mitt största tack går som alltid till min livskamrat Marita Olsson, vars ständiga kärlek och stöd skapar den grundtrygghet i tillvaron som möjliggör för mig att, utan

att totalt gå ned mig i grubblerier, kunna arbeta med de svåra och bitvis skrämmande frågor som står i centrum för boken. Därtill vill jag rikta ett stort tack till Marina Axelson-Fisk, Björn Bengtsson, Emma Jonson, Stefan Schubert och Thomas Weibull för värdefulla kommentarer till bonuskapitlet. Ingen skugga må emellertid falla över dem för de eventuella fel och anstötligheter som kvarstår i texten och som jag i vanlig ordning ensam bär ansvaret för.

Edsleskog, september 2023
Olle Häggström

Förord till första upplagan

EN UTOMJORDISK BETRAKTARE av vår planet skulle knappast kunna undgå att lägga märke till den dominerande ställning över andra djur och växter som *Homo sapiens* skaffat sig. Den manifesterar sig bland annat i storstädernas imponerande skylines och i de ofantliga arealer vi lagt under oss för jord- och skogsbruk, men också i en eskalerande atmosfärisk koldioxidhalt och i den tillbakagång av biodiversitet som brukar betecknas som planetens sjätte massutdöendeepok. Allt detta och mer därtill är människans verk. Vår dominans har dock väldigt lite att göra med vår muskelstyrka eller fysiska uthållighet: istället är det vår *intelligens* som ger oss ett avgörande övertag. Intelligensen är ett universalredskap, med hjälp av vilket vi lyckats med så skilda saker som att bygga broar, monument och världsomspännande mobiltelefonnät; skriva poesi; klyva atomkärnor; bekämpa pandemier; och till och med landsätta några av oss på månen. Den enorma förmåga att omstöpa världen som intelligens och tankekraft ger oss stämmer till eftertanke när vi nu är i full färd med projektet att skapa tänkande maskiner.

Rapporteringen om artificiell intelligens (AI) har de senaste åren kommit att uppta en successivt ökande del av mediaflödet, men jag vågar påstå att detta utrymme alltjämt är långt ifrån att stå i proportion till det genomgripande inflytande som fortsatt AI-utveckling kan väntas få på samhället och våra liv. Mycket av rapporteringen har därtill ett drag av naiv framstegsoptimism,

där specifika AI-tillämpningars potential att förbättra världen framhålls utan att teknikens risker och avigsidor vägs in i diskussionen. Min avsikt med denna bok är att korrigera den balansen, och ge den mer allsidiga bild av teknikens möjligheter och risker som behövs för att vi – mänskligheten – ska kunna navigera undan de svåraste blindskären och gå mot den lysande framtid som tekniken rätt hanterad kan bidra till att skapa.

De samhällsutmaningar som AI-utvecklingen aktualiserar är av många slag. Redan existerande AI-teknik reser svåra frågor rörande allt från risken för automatiserad diskriminering, och effekter av sociala mediers AI-algoritmer, till hur vi på olika vis kan sätta AI i tjänst för att uppnå FN:s hållbarhetsmål. Jag tar i boken upp ett brett spektrum av sådana frågor, jämte ett knippe något mer långsiktiga frågor som den om vad AI-driven automatisering kan komma att innebära för arbetsmarknad och ekonomisk ojämlikhet, och vad vi eventuellt kan och bör göra för att bana väg för ett samhälle där vi arbetar mindre eller inte alls.

Det allra största utrymmet i boken får dock den på lång sikt kanske mest betydelsefulla AI-frågan av alla, nämligen den om vad som kan väntas ske den dag vi lyckas skapa en maskin som i termer av allmänintelligens matchar eller överträffar mänsklig förmåga. Frågan avfärdas ibland som esoterisk och alltför fantasifull för att tas på allvar, men jag menar att vi inte har råd att ignorera den. Den är numera föremål för ett växande forskningsområde, och ju mer jag studerat detta område desto mer övertygad har jag blivit om att ett AI-genombrott av det slag som avses, vilket skulle försätta oss i ett läge där vi inte längre kan ta vår roll som planetens härskare för given, kommer att bli antingen det bästa eller det mest katastrofala som hänt i mänsklighetens historia. Vilket av dessa båda diametralt motsatta utfall vi landar i kommer i hög grad att bero på hur noggrant

vi förberett oss och hur väl vi lyckats styra AI-utvecklingen i rätt riktning.

Ungefär halvvägs in i mitt arbete med boken slog coronapandemin till. Den samhällskris som följde har jag iakttagit med stort intresse (och givetvis oro), och jag tror att det kommer att bli möjligt att dra viktiga lärdomar inför framtiden ur den, varav en del kan ha bäring på eventuella framtida AI-relaterade kriser. Det känns dock lite tidigt att dra några bestämda slutsatser av den ännu pågående krisen, och jag har därför valt att i denna bok nästan helt ignorera den.

Under arbetets gång har jag haft god hjälp av en rad personer som generöst ställt upp med att läsa och ge synpunkter på olika utkast. Av dessa vill jag rikta ett särskilt stort tack till Björn Bengtsson och Stefan Schubert för deras fenomenala engagemang i boken; utan deras utförliga och genomtänka kommentarer hade delar av den varit betydligt torftigare och sämre. Tack också till Hannes Bengtsson, Susanna Göranzon, Emma Jonson, Carl Lindberg, Patrik Lindenfors, Vilhelm Verendel, Edvin Wedin, Thomas Weibull och Anna Wisakanto, som alla bistått med värdefulla kommentarer och råd. Detsamma gäller förlagsredaktörerna Magnus Linton och Martina Stenström och förläggaren Christer Sturmark.

Mitt allra varmaste tack till går till Marita Olsson, min käresta och livskamrat sedan snart 28 år tillbaka. Sedan coronan slog till och vi övergick till att sköta våra respektive arbeten på distans har vi levt tillsammans tätare än någonsin – så gott som 24/7. Hennes ständiga kärlek och stöd, obrutet även när mina bryderier om AI-futurologi och bokredigering någon gång övergått i ältande, har hjälpt mig framåt på så många vis, inklusive färdigställandet av denna bok.

Edsleskog, januari 2021
Olle Häggström

KAPITEL 1

Människor och AI

TIDIGT PÅ MORGONEN den 11 februari 2013 rådde vinterlugnet i den lilla staden La Crosse i Wisconsin, USA, tills det plötsligt bröts av två gevärsskott mot ett hus. Ett ögonvittne rapporterade att skotten kom från en förbipasserande bil, vilken polisen snabbt kunde lokalisera, varpå en biljakt vidtog som inte tog slut förrän den jagade bilen kraschade i en snödriva. Polisen kunde utan ytterligare dramatik arrestera föraren, en man i 30-årsåldern vid namn Eric Loomis. Ingen människa kom till skada.[1]

Loomis förnekade inblandning i skjutningen, men medgav en rad andra brott, inklusive bilstöld, olaga vapeninnehav och allmänfarligt beteende under själva biljakten. För dessa brott dömdes han senare samma år till sex års fängelse. Händelsen hade säkert inte nått längre än till lokalpress om inte domen hade överklagats. Loomis och hans advokat Michael Rosenberg bestred inte de faktiska omständigheterna den där februarimorgonen, utan invände mot straffets längd, som ju utöver dessa omständigheter påverkas även av andra faktorer, inklusive hur återfallsbenägen gärningsmannen bedöms vara. Rosenberg pekade i överklagan på att domaren i detta fall hade använt sig av ett datorprogram kallat COMPAS (Correctional Offender

1 Garber (2016).

Management Profiling for Alternative Sanctions), som räknar fram en återfallsbenägenhet och klassar individen som antingen högrisk eller lågrisk.[2] Loomis, som hade varit i klammeri med rättvisan även tidigare, klassades av COMPAS som högrisk – vilket bidrog till det långa fängelsestraffet.

Rosenberg framhöll att varken han eller hans klient – och faktiskt inte ens domaren – hade insyn i datorprogrammets detaljer. Dessa var en affärshemlighet hos företaget Northpointe som utvecklat COMPAS.[3] Undergräver inte detta hemlighetsmakeri den anklagades möjlighet att försvara sig? Har rättssäkerheten satts ur spel? Wisconsins högsta domstol fastslog till slut den ursprungliga domen,[4] men de principiella frågorna kvarstår.

Northpointe och deras datorprogram, vilket sedan tidigt 00-tal använts för återfallsriskbedömningar i domstolar på många håll i USA, skulle snart hamna i ytterligare blåsväder. I maj 2016 publicerades i den New York-baserade nättidningen *ProPublica* en undersökning av en grupp grävande reportrar med Julia Angwin i spetsen, som hävdade att COMPAS fungerar som ett slags automatiserad rasdiskriminering, och lättare bedömer svarta än vita som högriskpersoner i fråga om återfallsbenägenhet.[5] Publiceringen ledde till livlig debatt som ännu pågår, om maskiners roll i detta slags bedömningar, och om vad som egentligen bör menas med diskriminering. Den sista frågan är knivigare än den först kan verka, och det finns

2 Denna förenklade uppdelning räcker för den följande diskussionen, fastän klassificeringen egentligen är något mer komplicerad är så: individer tilldelas en risknivå på en skala 1 till 10, där 1 till 4 räknas som »låg risk«, 5 till 7 som »medelhög risk«, och 8 till 10 som »hög risk«; se Larson m. fl. (2016).

3 Smith (2016).

4 Liu m. fl. (2019).

5 Angwin m. fl. (2016), Larson m. fl. (2016).

en rad tänkbara förslag på vad det borde innebära att behandla två eller flera grupper – i detta fall svarta och vita – rättvist och utan diskriminering.

Det kanske allra mest uppenbara förslaget går ut på att COMPAS och liknande programvaror inte ska få lov att använda sig av data om gärningspersonens rastillhörighet.[6] Detta är i själva verket redan fastslaget i amerikansk lagstiftning, och uppfylls av COMPAS,[7] men det räcker dessvärre inte. Skälet till det är att det finns många olika egenskaper som korrelerar med ras, vilket öppnar för datorprogrammet att (oavsett om det var programutvecklarnas avsikt eller inte) indirekt ta hänsyn till ras. Det var motsvarande möjlighet att hitta ledtrådar om kön med hjälp av korrelerande egenskaper som orsakade det uppmärksammade fiasko där det amerikanska teknikföretaget Amazon till slut gav upp sitt försök med programvara för automatiserat urval av jobbsökande att kalla till intervju. Hur de än skruvade på algoritmens parametrar så blev de inte kvitt dess tendens att favorisera manliga sökande framför kvinnliga.[8]

Nästa förslag till rättvisekriterium skulle kunna vara att kräva att den andel gärningspersoner som klassas som högrisk ska vara densamma i varje raskategori, så att om exempelvis 30 % av vita gärningspersoner klassas som hög risk, så ska även

---

6 Begreppet ras är givetvis problematiskt. I USA har man löst det juridiskt genom att låta individen själv definiera sin rastillhörighet.

7 Hao och Stray (2019).

8 Dastin (2018) betonar att programmets könsdiskriminering berodde på att det tränats med data från företagets tidigare anställningspraktik och därigenom reproducerade orättvisor därifrån. Tendensen att företrädesvis anställa män syntes i data som en tendens att anställa personer med egenskaper som korrelerar med manligt kön, och algoritmen tolkade dessa egenskaper som önskvärda, vilket fick den att förorda manliga kandidater.

30 % av svarta gärningspersoner få samma klassificering. Detta kan i förstone tyckas rättvist, men är bara rimligt om vi kan ta för givet att andelen personer med den eftersökta egenskapen är densamma för olika kategorier, vilket ofta inte är fallet. De data från brottsregistret i Boward County i Florida som reportrarna på *ProPublica* använde sig av, och som förutom COMPAS-bedömningar innehåller data om vilka som inom en given tidsperiod faktiskt återföll, visar på en sådan diskrepans. Av vita gärningspersoner var det 39 % som återföll, medan motsvarande siffra för svarta gärningspersoner var 51 % – en betydande skillnad.[9] Om vi gör (det något orealistiska) tankeexperimentet att vi lyckas utveckla ett program som perfekt förutser alla gärningspersoners eventuella återfall i brottslighet, så skulle detta program klassificera 39 % av de vita och 51 % av de svarta som högrisk, vilket enligt detta kriterium skulle ses som diskriminerande. Men att ha ett rättvisekriterium som dömer ut en metod som behandlar varje enskild individ exakt rätt verkar inte rimligt, och det är inte svårt att koka ihop andra exempel som visar metodens orimlighet.[10] Att samma andel ska

9 Vad denna skillnad kommer sig är en intressant fråga, liksom den i USA länge omdebatterade frågan om varför brottsstatistiken ser så olika ut i olika raskategorier. Troligtvis är orsakerna mer än en, och några huvudkandidater är socioekonomiska faktorer som segregation, utbildningsnivå och fattigdom, jämte förekomsten av diskriminerande särbehandling inom polis och rättsväsende; se exempelvis Gabbidion och Greene (2005) och Sampson m. fl. (2005). Vad gäller den aktuella datamängden från Boward County har Uppsalamatematikern David Sumpter funnit att skillnaden i återfallsfrekvens mellan grupperna svarta och vita helt kan förklaras med att de svarta gärningspersonerna i medeltal är yngre än de vita, och att yngre brottslingar har större återfallsbenägenhet än äldre (Sumpter, 2018).

10 Ändå förekommer det att detta kriterium används, som exempelvis i de direktiv vi hade att rätta oss efter när jag under åren 2010 till 2015 deltog i

bedömas som högrisk oavsett raskategori är därför inte något bra rättvisekriterium.

Det finns dock andra och betydligt rimligare kriterier för att räkna en klassificeringsalgoritm som icke-diskriminerande. En sådan är att den ska vara *kalibrerad*, vilket i detta fall betyder att bland de individer som fått en viss klassificering, så ska andelen som faktiskt återfaller vara (ungefär) densamma för svarta som för vita.

Som ett alternativt kriterium kan man se till hur stor andel av individerna i olika kategorier som felklassificerats. Vi kan exempelvis jämföra andelen bland icke-återfallande vita som klassificeras som högrisk med andelen icke-återfallande svarta som råkat ut för samma felklassificering. Det är önskvärt att dessa andelar så kallade *felaktigt positiva* är (ungefär) samma för de båda raskategorierna, liksom detsamma bör gälla för andelarna *felaktigt negativa*, det vill säga för andelen återfallande i respektive kategori som felklassificeras som lågrisk.

Vid hastigt påseende kan det tyckas som om å ena sidan kalibrering, och å andra sidan att betrakta andelar felaktigt positiva och felaktigt negativa, vore ungefär samma sak, men det stämmer i själva verket inte alls, och skillnaden mellan kriterierna står i centrum för kontroversen kring COMPAS. Angwins och hennes *ProPublica*-kollegors påstående att COMPAS diskriminerar svarta gärningspersoner backas upp av Bowden County-data, som

---

Vetenskapsrådets arbete med att fördela medel till svenska universitetsforskare som ansökt om forskningsbidrag. Vi anmodades (a) att fördela bidrag enbart baserat på kvaliteten i den planerade forskningen och den sökandes meriter, och (b) att se till att samma andel manliga som kvinnliga sökande beviljades medel. Dessa båda anvisningar går bara att uppfylla samtidigt givet att det inte finns något statistiskt samband mellan ansökningarnas kvalitet och de sökandes kön, fast det knappast kan tas för givet att inte exempelvis kvinnliga sökande skulle skriva i snitt bättre ansökningar än de manliga.

visar att andelen felaktigt positiva är 45 % för svarta jämfört med bara 23 % för vita, jämte en lika slående diskrepans i fråga om felaktigt negativa (åter till svarta gärningspersoners nackdel). Till COMPAS försvar kan dock kalibreringskriteriet anföras. Av svarta som klassats som högrisk återföll 63 %, medan motsvarande siffra bland vita som klassats som högrisk var 59 %. Dessa siffror är tillräckligt nära varandra för att motivera att programmet räknas som någorlunda kalibrerat.[11] Tim Brennan, forskningschef på Northpointe och en av arkitekterna bakom COMPAS, pekar på detta till stöd för sin ståndpunkt att programmet alls inte gör sig skyldig till diskriminering, och tillägger i en intervju att kalibrering i själva verket är precis vad amerikansk rättspraxis kräver av program som detta.[12]

Så det finns alltså minst två olika till synes rimliga idéer om vad det bör betyda att ett program som COMPAS är rättvist och undviker rasdiskriminering: å ena sidan att det uppvisar samma andel felaktigt positiva respektive felaktigt negativa för olika raskategorier, å andra sidan att det är kalibrerat. Vilken bör hållas som rättesnöre? Om det kan man ha delade meningar, men den vakne läsaren kanske frågar sig om vi inte borde kombinera dem, och kräva att *båda* kriterierna är uppfyllda. Kan inte programmet utformas för att ge det bästa av två världar, på så vis att riskbedömningarna inte är rasdiskriminerande oavsett vilket av de båda kriterierna som än tillämpas?

Svaret är att det inte går! Så länge inte återfallsbenägenheten råkar vara densamma i de båda populationerna (svarta

11 Den som vill försvara Northpointe och deras programvara COMPAS mot anklagelsen om inbyggd diskriminering av svarta kan också notera att andelen återfall hos dem som klassificerats »högre risk« är något *större* hos svarta än hos vita, vilket kan förstås som att om någon grupp har anledning att klaga på diskriminering så är det i varje fall inte de svarta.

12 Sumpter (2018), s 69.