TIDSKRIFT FÖR POLITISK FILOSOFI
NR 2 2022 | ÅRGÅNG 26

Bokförlaget THALES

→ ARTIFICIELL INTELLIGENS, LIKRIKTNINGS-PROBLEMET OCH POLITISK FILOSOFI – EN ÖVERSIKT

Markus Furendal

1 Inledning

FÅ KAN HA missat att forskningen kring Artificiell Intelligens (AI) har gjort stora framsteg under det senaste decenniet. Media förmedlar ett stadigt flöde av rapporter om nya genombrott, och AI-system kan numera producera text, manipulera bilder och analysera oöverskådliga mängder material. AI får allt oftare sköta sysslor som vi traditionellt har förmodat kräver mänsklig intelligens, och såväl privata företag som offentliga institutioner har börjat automatisera vissa sorters beslutsfattande. Förklaringen till framstegen beskrivs ofta som att AI-forskning har bytt strategi, från att försöka bygga en komplett och färdig intelligent agent, till att istället med hjälp av så kallad maskininlärning skapa vad som kan liknas vid ett barn med förmågan att lära sig själv. Startar vi inlärningsprocesser på rätt sätt kan en rad förmågor tränas upp tack vare kraftigare datorer och tillgången till den enorma mängd data som finns på Internet. Därav den AI-hajp vi just nu befinner oss i.

Dessa tekniska framsteg väcker en rad filosofiska frågor kring vad intelligens och medvetande är, men också frågor av moralisk och politisk karaktär kring hur AI-system bör användas, och hur de påverkar våra samhällen och våra relationer. I kölvattnet av den tekniska utvecklingen samt några inflytelserika böcker som publicerades under första halvan av 2010-talet (t.ex. Bostrom 2014; Brynjolfsson & McAfee 2014) har det nu i praktiken uppstått ett nytt fält inom tillämpad etik, ofta kallat AI-etik (›AI ethics‹), som fokuserar på de många frågor och problem som aktualiseras av AI-tekniken. Denna artikel ämnar ge en överblick över några av de mest diskuterade frågorna inom samtida AI-etik genom att fokusera på fyra nyligen utgivna böcker. Böckerna som diskuteras är journalisten Brian Christians *The Alignment Problem* (2020), datavetaren och AI-nestorn Stuart Russells *Human Compatible* (2019), ekonomen Carl Benedikt Freys *The Technology Trap* (2019), samt

ISSN 2002-3383

Tänkande maskiner av Olle Häggström (2021), professor i matematisk statistik. Artikelns övergripande syfte är inte att presentera eller utvärdera en specifik AI-etisk fråga, utan snarare att ge en introduktion till fältet för politiskt-filosofiskt sinnade personer som är intresserade av AI, och samtidigt signalera några vita fläckar på kartan som är angelägna att fylla i.

2 Likriktningsproblemet

DEN INTERNATIONELLA AKADEMISKA och populärvetenskapliga diskussionen kring moraliska och politisk-filosofiska aspekter av AI-tekniken är nämligen överraskande nog präglad av en stor brist på bidrag från specialister inom just politisk filosofi.[1] Istället är det i första hand AI-experter som har börjat intressera sig för och publicera kring dessa frågor. Det finns så klart något gott i det, eftersom filosofiska analyser och slutsatser kommer att behöva operationaliseras och programmeras in i maskinerna om vi ska ha någon nytta av dem, och det krävs tekniskt kunnande för att förstå AI-teknikens förmågor och begränsningar. Samtidigt medför ju detta att problemformuleringar och lösningsförslag kommer ha en viss slagsida åt de tekniska aspekterna. Det är kanske mest uppenbart i faktumet att mycket av diskussionen kring AI-etik har kommit att handla om det som på engelska kallas *the alignment problem* (vilket jag här kallar *likriktningsproblemet*): ungefär, problemet med hur vi ska få en AI att frambringa ett utfall som ligger i linje med vad vi vill. Hur detta problem ska lösas har kommit att bli den drivande frågan för mycket av AI-etik, och det gemensamma temat för böckerna som diskuteras i denna artikel är att de diskuterar problemet eller föreslår lösningar.

Trots att begreppet är så spritt är det emellertid svårt att förstå exakt vad likriktningsproblemet består i. I sin mest inkluderande tolkning verkar det kunna fyllas med alla tänkbara AI-etiska problem som uppstår när AI-system byggda för en specifik uppgift misslyckas med den. AI-algoritmer beskrivs exempelvis ibland som fördomsfulla (›biased‹) om de, på grund av bristfälliga träningsdata, reproducerar strukturer i samhället som vi bedömer som pro-

blematiska. Ett välkänt fall är det AI-baserade rekryteringsverktyg som företaget Amazon år 2015 testade för att finna nya medarbetare: eftersom programmet tränats för att söka efter kandidater som var lika dem som redan jobbade på Amazon sorterade verktyget systematiskt bort kvinnliga kandidater (Christian 2020: 40). Sådan AI-teknik beskrivs som *smal* (›narrow›) – den är expert på en uppgift men klarar inget annat. Vi kan således kalla problem som det Amazon erfor för *smala likriktningsproblem*.

En alternativ tolkning är emellertid att det mest relevanta likriktningsproblemet handlar om ett möjligt framtidsscenario där vi lyckas skapa system med Artificiell Generell Intelligens (AGI) som överträffar mänskliga kognitiva förmågor och klarar alla möjliga uppgifter. Enligt exempelvis Olle Häggström (2021) är det allmänt antaget bland AI-experter att detta skulle innebära ett enormt genombrott med potentiellt underbara eller katastrofala följder, beroende på om vi lyckas se till att denna hypotetiska AGI har våra bästa intressen som mål. Oron är att en sådan agent lätt skulle kunna bli övermänskligt kompetent inom alla områden, och därför vara omöjlig att kontrollera. Redan 1960 konstaterade AI-forskaren Norbert Wiener att vi i detta fall bör ha försäkrat oss om att »[…] det mål vi stoppat in i maskinen verkligen är det mål vi eftersträvar, och inte bara en färgrik imitation av det» (citerad av Christian 2020: 295, min övers.). Eftersom en AGI enfaldigt drivs av det övergripande mål vi givit den och de instrumentella mål som är nödvändiga för att uppnå det övergripande målet riskerar alla andra mänskliga värden, och till och med mänsklighetens fortlevnad, att offras. För att illustrera denna insikt har Nick Bostrom beskrivit ett tankeexperiment där en AGI får uppgiften att producera gem. Om den är tillräckligt kraftfull, och vi glömmer att specificera en övre gräns för produktionen, kan vi få en situation där AGI:n till slut omvandlar alla möjliga resurser – inklusive människor – till gem. Varje invändning mot detta scenario möts i regel av en fantasifull förklaring. Vi skulle till exempel inte kunna stänga av maskinen, eftersom den tack vare övermänsklig intelligens kan förutse detta och göra back up-kopior av sig själv, övertala människor att

hjälpa den, eller på andra sätt neutralisera hotet. Tramsigheten i detta tankeexperiment är, som Häggström (2021: 105) påpekar själva poängen, eftersom det visar att intelligens bara handlar om måluppfyllelse, och inte vilka målen är. Om en övermänskligt intelligent AGI skulle fungera på detta sätt kan alltså vilket godtyckligt mål som helst leda till mänsklighetens undergång. Sådana system kan så klart också få egna mål att sträva mot, vilket i värsta fall kan medföra att vi som mänsklighet i framtiden förhåller oss till AGI på samma sätt som gorillor förhåller sig till människor idag: vi får existera, men detta är helt avhängigt maskinernas goda vilja (Russell 2019: kap. 5). Eftersom AGI:er är just generella kommer jag kalla denna typ av problem för *generella likriktningsproblem*.[2]

I den mån de fyra böcker som diskuteras här utgör ett rättvist tvärsnitt av befintlig och inflytelserik AI-etik (fältet utvecklas, precis som tekniken, oerhört fort), vill jag driva följande tes: i en bemärkelse kan så klart alla de moraliska och politiska frågor som väcks av AI-teknik förstås som likriktningsproblem av varierande bredd, eftersom tekniken frambringar andra utfall än dem vi önskar se. Det verkar emellertid rimligt att vi lätt kan vinna analytisk skärpa om vi är noggrannare med vilken roll AI faktiskt spelar, och vilken normativ relevans den egentligen har. Att förstå AI-etik på detta sätt kan annars komma att dölja de i grunden politiska aspekterna av hur AI-teknik förstås, utvecklas, och tillämpas. Jag återvänder till denna poäng nedan.

3 Fyra böcker om likriktningsproblemet

BRIAN CHRISTIANS *The Alignment Problem* (2020) torde vara den just nu bästa och mest pedagogiska introduktion till de tekniska framsteg som har lett fram till var AI-utvecklingen står idag. Det är emellertid ingen filosofisk bok. Vi erbjuds egentligen aldrig någon tydlig definition av vad likriktningsproblemet är och Christian diskuterar smala och breda likriktningsproblem om vartannat. Boken är snarare en idéhistorisk beskrivning baserad på intervjuer med ledande forskare och den ger en inblick i hur vår samtida förståelse av maskininlärningsteknik och AI-etik har vuxit fram, samt

vilka möjliga lösningar som just nu prövas. Dess främsta förtjänst är hur väl Christian förklarar de tekniska begränsningar som finns hos existerande maskininlärningssystem, och därmed sticker hål på många myter. Christians mer generella poäng är att AI-system egentligen inte är tränade på världen som sådan, utan på en förenklad modell av världen. Ett bildigenkänningsprogram har till exempel endast en begränsad uppsättning kategorier att sortera in bilder i, och det förutsätts att kategorierna är ömsesidigt uteslutande och uttömmande, vilket självklart inte är fallet. Ett system som endast har kategorierna ›häst‹ och ›åsna‹ men inte ›mulåsna‹ kommer stå handfallet inför en bild på detta djur, men *måste* sortera det som antingen häst eller åsna (Christian 2020: 315). Detta teoretiska antagande får praktiska konsekvenser när AI lämnar modellerna och tillämpas i den verkliga världen: den dödliga kollisionen mellan en självkörande Uberbil och en fotgängare i Arizona år 2018 skedde delvis för att AI-systemet inte förmådde klassificera fotgängaren, eftersom det inte hade tränats på möjligheten att människor ibland går över gatan på andra ställen än övergångsställen (Christian 2020: 326–327). Det är en dyrköpt insikt, för utanför labbet finns ingen möjlighet att starta om scenariot. Christian (2020: 325) menar därför att det främsta problemet inte är att vi riskerar att förlora kontrollen till AI-system, utan snarare till de formella, numeriska och begränsade modeller av världen och våra värderingar som AI-system ofrånkomligen utgår från.

Stuart Russell är en inflytelserik datavetare vid UC-Berkeley som tillsammans med Peter Norvig har skrivit standardtextboken för AI- och maskininlärningsstudenter. Han medverkar ofta i den allmänna debatten kring AI och i *Human Compatible* (Russell 2019) diskuterar han smala och generella likriktningsproblem och möjliga tekniska lösningar. Likt Christian ägnar Russell en stor del av boken åt att förklara vilka möjligheter och risker AI medför, och han ger också en lättfattlig introduktion till de inre mekanismerna i AI (för den intresserade finns dessutom ett antal mer tekniska appendix). Mest intressant är dock diskussionen kring det Russell kallar *Inversed Reinforcement Learning*, vilket han hoppas kom-

mer kunna lösa likriktningsproblemet. I korthet tar Russell fasta på Wieners insikt från 1960, som visar att standardmodellen för AI – där vi ger AI:n ett mål och ber den maximera det – ofta kommer att leda fel, eftersom det är så svårt att formulera målet på rätt sätt (Russell (2019: 136–140) nämner till exempel Kung Midas som en sedelärande historia med denna sensmoral.) Istället bör AI-utveckling följa tre principer: i) maskinens enda mål är att maximera förverkligandet av mänskliga preferenser, ii) maskinen är från början osäker på vilka dessa preferenser är, och iii) den yttersta källan till information om mänskliga preferenser är mänskligt beteende (Russell 2019: 173). Till skillnad från den gemproducerande AGI:n ovan skulle ett system designat i enlighet med dessa principer gladeligen låta sig stängas av eller bli omprogrammerad, eftersom den andra och tredje principen bygger in ett slags ödmjukhet och vilja att erkänna sina brister. Russell (2019: kap. 8) ger ett antal enkla spelteoretiska exempel där mänskliga och artificiella agenter lyckas samarbeta tack vare dessa principer. Det är tyvärr oklart hur vi ska ta steget från enkla modeller där AI:n ska inferera vad *en* person vill, till en modell med *många* personer där preferenser behöver vägas mot varandra. Dessutom uppstår ju uppenbara frågor kring vad mänskliga preferenser är, och om maskiner (eller vi) överhuvudtaget kan identifiera dem utifrån observerat beteende.[3] Detta kräver moralteori, svarar Russell, men för sedan ingen allmän filosofisk diskussion, utan slår snabbt fast att svaret står att finna i ett slags preferensutilitarism inspirerad av Harsanyi (Russell 2019: 217–221). Ett kapitel ägnas åt att bemöta ett antal standardinvändningar mot denna syn, men ytterst lite argumentation ges för varför vi i denna fråga bör vara konsekventialister till att börja med. Ett svårbegripligt argument verkar till exempel vara att eftersom vi bygger maskiner för att uppnå vissa konsekvenser bör vi föredra att bygga maskiner som frambringar dessa konsekvenser (Russell 2019: 217). Att be maskinerna observera vårt beteende för att identifiera våra preferenser verkar ju också riskera att ge för mycket vikt åt *status quo*, men Russell antar helt enkelt att »[v]i har skapat [världen] på ett visst sätt eftersom vi – grovt räknat – föredrar den

på det sättet» (2019: 181, min övers.). Uppenbara motexempel baserade på hur svårigheter att koordinera kollektivt handlande leder till omständigheter som ingen önskar, som till exempel klimatförändringarna, lyser med sin frånvaro.

Sammantaget kan Russells bok sägas vara ett värdefullt och viktigt bidrag till debatten, men också ett exempel på hur AI-etiken färgas av den disciplinära bakgrund man har. Att behandla alla dessa frågor som vore de utmaningar för en ingenjör bygger in en alltför grund förståelse för politiska och samhälleliga fenomen i AI-etiken.

Carl Benedikt Frey är en svensk-tysk ekonom verksam vid Oxford som år 2013 medförfattade den uppmärksammade studie som uppskattade att 47 % av alla amerikanska jobb skulle kunna komma att automatiseras (sedermera publicerad som Frey & Osborne 2017). I en mening är detta ett slags likriktningsproblem (även om inte Frey beskriver det så), eftersom vi implementerar AI-teknik med ett visst syfte (att öka produktiviteten) men frambringar ett icke-önskat utfall (vissa får det sämre till följd av förlorade arbetstillfällen). Å andra sidan talar vi ju därmed om AI på ett helt annat sätt än i de ovan diskuterade exemplen, eftersom de normativt relevanta frågorna här är oberoende av vilka specifika mål AI-systemen har: felriktningen beror ju på att vissa människors mål (vinstmaximering) skiljer sig från andra människors mål (att arbeta och få en inkomst). Freys bok *The Technology Trap* (2019) följer upp den första studien och syftar till att ge en mycket djupare ekonomisk-historisk förståelse för hur teknologi påverkar ekonomier och samhällen. Specifikt försvarar Frey tesen att människor historiskt sett har motsatt sig ny teknik när den har *ersatt* mänsklig arbetskraft, men accepterat den när den har *förstärkt* densamma. Att ludditerna slog sönder maskiner berodde till exempel på att de levde under en period då den arbetskraftsersättande teknologin som möjliggjorde den industriella revolutionen hade effektiviserat produktionen, men de samhällsekonomiska vinsterna av produktivitetsökningen ännu inte hade kommit särskilt många till del (Frey 2019: kap. 5). Denna period brukar kallas Engels paus, efter Friedrich Engels studier av

det industriella England. De enorma produktivitetsökningar som följde av mekaniseringen och automatiseringen av arbetsuppgifter under 1900-talet orsakade dock inte tillnärmelsevis samma motstånd, vilket Frey (2019: kap. 6–8) förklarar med att löntagarnas relativa förhandlingsstyrka, samt politiska beslut, gjorde att välståndet spreds mer jämlikt. Med denna historiska bakgrund vänder sig Frey sedan mot samtiden, och beskriver hur reallönerna har stagnerat, produktivitetsökningen bromsat in och andelen av BNP som tillfaller kapital snarare än arbete ökat under de senaste decennierna. Dessa tendenser kan komma att förstärkas ytterligare när arbetare ersätts av maskiner, det vill säga kapital. Men automatiseringen är redan igång och vi har varit här förut, konstaterar Frey (2019: 348), och eftersom det är ett fördelningsproblem krävs politiska lösningar. Frey nämner ett antal olika förslag, som till en viss besvikelse emellertid inte är specifika för just AI-teknik utan redan genomtröskade svar på ökande ekonomisk ojämlikhet i allmänhet. Alla syftar till att underlätta för arbetare att anpassa sig till automatisering, som till exempel ett slags jobbskatteavdrag, ökade möjligheter att bygga tätt i ekonomiskt starka områden, och förbättrad infrastruktur för pendling (Frey 2019: kap. 13).

Freys bok är ett välkommet tillskott till debatten kring de externaliteter som implementeringen av AI-teknik kan komma att medföra. Alltför ofta framförs onyanserade scenarier, som att *alla* jobb kan komma att försvinna, eller att produktivitetsökningen kommer skapa så mycket ny efterfrågan att *ingen* nettoförlust av jobb kommer ske. Frey visar istället att automatisering är ett komplicerat fenomen vars effekter inte är desamma oavsett geografi eller sektor (nästan alla amerikanska tillverkningsrobotar är till exempel koncentrerade till de industriella områdena i östra USA. Se Frey 2019: kap. 10), och projektet verkar drivas av övertygelsen att även om alla får det bättre i det långa loppet så kan flera generationer få det sämre innan dess (Frey 2019: 126). Samtidigt får vi aldrig riktigt någon förklaring till *vad* det är som gör att teknologi antingen ersätter eller förstärker mänsklig arbetskraft, och ibland verkar AI-utvecklingen nästan betraktas som en deterministisk kraft som kan

hanteras, men inte styras. Det vore inte rättvist att kritisera boken för att sakna politiska förslag på hur sådan styrning skulle kunna se ut, eller mer rättviseteoretiska resonemang kring varför den behövs. Men bristen på detta visar samtidigt på ett område där politisk filosofi kan bidra.

Olle Häggström är professor i matematisk statistik vid Chalmers och via sin blogg och annan publik filosofi är han en viktig spridare av kunskap om AI-etik och relaterade frågor i svensk offentlighet. Hans nyligen utgivna *Tänkande maskiner* (2021) är det första större verket om AI-etiska frågor på svenska. Häggström summerar och ger en översikt över det senaste decenniets internationella diskussion kring riskerna med AI, och kanske framförallt AGI. Häggström skiljer nämligen mellan ›jordnära‹ och ›högtflygande‹ frågor på ett sätt som ungefärligen överlappar med de smala och breda likriktningsproblemen jag diskuterat ovan: jordnära frågor inbegriper till exempel hur självkörande bilar bör utformas och hur rätten till våra data ska skyddas från techjättar, medan högtflygande frågor inbegriper bland annat huruvida en AI kan uppnå medvetande och hur vi ska kunna förhindra att en superintelligent AGI utplånar mänskligheten (Häggström 2021: 20–26). Detta breda anslag är välkommet för en bok som syftar till att introducera frågorna för en intresserad allmänhet, och distinktionen är användbar. Samtidigt diskuteras de jordnära frågorna relativt kortfattat (se framförallt Häggström 2021: kap. 3) och desto mer tid ägnas åt de högtflygande frågorna, och särskilt riskerna med en potentiell AGI. Det är inte förvånande, eftersom Häggström hör till en skara tänkare som menar att denna eventualitet är ett av de största hoten mänskligheten någonsin har stått inför. Baserat på resonemang från bland annat Nick Bostrom och Effektiva Altruister som Toby Ord är idén, i korthet, att det vore katastrofalt om mänskligheten dog ut till följd av exempelvis en superintelligent AGI, eftersom det inte bara utsläcker alla existerande människors liv, utan också berövar universum alla framtida potentiella människor och de förmodat värdefulla liv de hade kunnat leva. En överslagskalkyl visar att så många som 10^{32} människor skulle kunna hinna leva om mänsklig-

heten fick fortsätta att existera i en miljard år. Det innebär att om vi i konsekventialistisk anda funderar över hur vi kan frambringa det bästa möjliga utfallet kommer alternativet ›minska existentiell risk‹ alltid väga oerhört tungt. Om vi kan välja att rädda en miljon nu levande människor eller minska risken att mänskligheten dör ut med en miljondels procent, kommer 100 gånger fler räddas om vi väljer det senare alternativet (Häggström 2021: 205). Alltså spelar det i praktiken ingen roll, menar Häggström, om vi tycker att gemapokalypsen är en tramsig fantasi, eller om vi tror att genombrottet som skapar AGI ligger 1, 100 eller 1000 år in i framtiden. Med tanke på vad som står på spel kan vi inte rättfärdiga att ignorera risken.

Det är uppenbart att denna insikt driver Häggströms bok, men insikten är också central i mycket av den internationella filosofiska och tekniska forskningen kring det breda likriktningsproblemet som han syntetiserar i boken. Häggström beskriver till exempel de bästa uppskattningarna av när en AGI kan tänkas skapas, och vad det kan tänkas medföra (inklusive »gemapokalypsen») (2021: kap. 4, 5 & 8). Han förklarar varför det vore fel och kanske till och med oansvarigt att avfärda risken (Häggström 2021: kap. 6, 9 & 10), och diskuterar ett antal tekniska och politiska strategier för att minska den (Häggström 2021: kap. 7 & 11). Även om han inte vill stoppa den tekniska utvecklingen, eftersom det verkar både omöjligt och dumt med tanke på vilka goda konsekvenser AGI kan medföra om vi löser problemet, så finns det vissa typer av applikationer som bör förbjudas (Häggström (2019: 288) nämner till exempel autonoma vapen som drönarsvärmar.) Likt andra AI-futurologer föreslår Häggström också att vi bör styra om forskningsmedel från försöken att skapa så kraftfulla AI-system som möjligt till forskning »[…] vars fokus är att göra AI *säker* och ge den *goda konsekvenser för samhället*» (2021: 278, Häggströms kurs.).

Häggströms bok lyckas väl med sitt syfte att introducera AI-etiska frågor för svenska läsare, och kanske framförallt vikten av att bry sig om den existentiella risk AGI medför. Men, samtidigt som argumenten för denna slutsats verkar framkalla en stark hängivenhet hos vissa misstänker jag att många kan komma att avfärda hela resone-

manget som alltför abstrakt eller konstlat. Spekulationer kring hur många ofödda människor som aldrig får chans att existera kan kanske, ungefär som *den motbjudande slutsatsen* (Parfit 1984), uppfattas som exempel på hur filosofer börjar med något trivialt och sedan vrider på det till dess att vansinniga implikationer verkar följa. Likväl är det seriösa filosofiska argument som leder dem till slutsatsen att vi bör bry oss om AGI:s hot mot vår existens. Om vi motstår impulsen att förkasta hela resonemanget – och bortser från frågan hur framgångsrik strategin är för att övertyga skeptiker – måste vi konstatera att det skulle krävas seriös filosofisk argumentation för att ens försöka avfärda slutsatsen, och det är inte något jag ska göra i denna text.

Däremot kan vi notera att när dessa frågor anses allt viktigare knuffas andra frågor nedåt på dagordningen. Häggström noterade ju själv att vår tid och våra forskningsmedel är begränsade. Samma skäl kan rimligen anföras för varför medel också bör omfördelas för att teoretisera kring de mer jordnära frågor som jag här har kallat smala likriktningsproblem. För den rimligaste tolkningen av oron kring en övermänsklig AGI är ju inte att det med all säkerhet kommer att ske, utan att det finns en åtminstone minimal risk. Vågskålen med de katastrofala följder som en AGI kan orsaka kan väl inte *helt och hållet* väga tyngre än problemen vi observerar kring redan idag tillämpade smala AI-system? Häggström hävdar aldrig att det är så, och så vitt jag känner till är det inte en position som försvaras i den internationella litteraturen. Trots det händer det lätt att de mer högtflygande frågorna får stort utrymme i samtida AI-etik, och en större balans skulle vara hälsosam.

4 Avslutning – vad bör nästa våg av AI-etik fokusera på?

LÅT OSS AVSLUTNINGSVIS återvända till tesen jag introducerade ovan, som sa ungefär att det kan finnas en poäng med att ha en bredare och mer nyanserad analysapparat kring AI-etiska frågor än att förstå det som att allt handlar om att få AI-system att frambringa de utfall vi önskar se. Oron är att om likriktningsproblemet blir vår AI-etiska hammare riskerar alla problem att förvandlas till spikar.

Denna reflexion aktualiseras inte bara i ljuset av dessa böcker utan också den lilla men snabbt växande internationella facklitteraturen kring AI-etik som fortfarande verkar värdesätta teknisk kompetens högre än samhällsvetenskaplig eller filosofisk kompetens. Att fältet i ett tidigt skede dominerades av AI-experter kan kanske förklara varför just problemformuleringen fokuserad kring likriktning fortfarande hänger kvar.

Rent konkret finns det goda skäl att tro att problemformuleringen döljer centrala skillnader mellan det AGI-relaterade generella likriktningsproblemet och så kallade smala likriktningsproblem, och att de senare ofta kan beskrivas med hjälp av existerande analysverktyg. Som jag har antytt ovan så skapar smala AI-applikationer politiska och moraliska problem när de antingen fungerar för dåligt (självkörande bilar), eller när de reproducerar redan existerande orättvisor (fördomsfull AI). Men normativa frågor uppstår också när AI fungerar för bra, till exempel när tekniken möjliggör nya former av massövervakning. I de fallen är det dock inte bristen på harmoni mellan användarens avsikt och systemets effekt som är problemet, utan snarare tvärtom: med hjälp av en effektiv AI kan vissa aktörers dåliga avsikter få större genomslag i världen. Trots att smal AI kan frambringa nya fenomen verkar det alltså inte som att det spelar roll ur ett normativt perspektiv att allt detta drivs av AI. I någon mening är tekniken bara ett nytt slags verktyg; det som egentligen spelar roll är de förmågor och den makt de skänker åt dem som använder dem (som i fallet med övervakning), eller den förmenta felfrihet och objektivitet de antas ha (som i fallet med diskriminering). Relevant politisk-filosofisk analys måste vara känslig för denna observation, för den innebär att vi bör rikta blicken mot människorna och institutionerna som utvecklar och tillämpar smal AI snarare än systemen själva. En AGI:s intressen, däremot, kan kanske vara mer eller mindre likriktade med våra, och här spelar det ju roll att vi talar om en agent som kan vara vår vän eller fiende, eller likgiltig inför oss. Men den stora mängden AI-etiska frågor som väcks av smala AI-system kan förstås utan att vi tar till likriktningsmetaforen. Tesen jag introducerat här säger alltså att detta

riskerar att glömmas bort om vi pratar om AI som verktyg i samma andetag som vi pratar om AI som ett subjekt.

Detta är främst en begreppslig och intradisciplinär poäng, som jag hoppas kan hjälpa till att sortera upp de många olika normativt relevanta frågor som väcks av AI-teknologins framsteg. Eftersom det fortfarande är ett litet fält bör givetvis alla som är intresserade ansluta sig, oavsett om de lockas av de mer jordnära eller de mer högtflygande frågorna. Efter den andra våg av AI-etisk litteratur som diskuterats här kan vi nu förhoppningsvis börja bygga mer sammanhängande analyser och teorier med hjälp av den politisk-filosofiska traditionens tidigare insikter och angreppssätt.

→|

Markus Furendal är postdoktor i statsvetenskap vid Statsvetenskapliga institutionen på Stockholms universitet.

Noter

1 Undantag finns givetvis och nya bidrag publiceras hela tiden. Nyligen har det till exempel givits ut två handböcker om AI-etik (Dubber, Pasquale och Das 2020; Liao 2020) samt ett specialnummer av *Canadian Journal of Philosophy. Philosophical Studies* förbereder i skrivande stund ett specialnummer om normativ teori och AI, och en antologi om 'AI governance' ges ut under 2022 av Oxford University Press (Bullock et al. 2022). Se också översikten i Erman & Furendal (2022).

2 En del AI-utvecklare skiljer på 'outer alignment', vilket fångar in såväl smala som generella likriktningsproblem som jag beskrivit dem ovan, och 'inner alignment', vilket betecknar ett mer tekniskt problem kring hur vi kan veta att en AI-modell når fram till ett visst utfall på det sätt vi vill. Ibland skiljer man också mellan AI:s implikationer på kort och lång sikt, där smala likriktningsproblem kan sorteras som kortsiktiga och generella likriktningsproblem är långsiktiga.

3 Christian (2020) diskuterar en rad ytterligare problem med Russells modell i sitt kap. 8.

Referenser

BULLOCK, JUSTIN., YU-CHE CHEN, JOHANNES HIMMELREICH, VALERIE M. HUDSON, ANTON KORINEK, MATTHEW YOUNG & BAOBAO ZHANG (red.) (2022) *Oxford Handbook of AI Governance*, Oxford: Oxford University Press.

BOSTROM, NICK (2014) *Superintelligence: Paths, Dangers, Strategies*, Oxford: Oxford University Press.

BRYNJOLFSSON, ERIK & ANDREW MCAFEE (2014) *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, New York, NY: W. W. Norton & Company.

CHRISTIAN, BRIAN (2020) *The Alignment Problem: How Can Machines Learn Human Values?*, London: Atlantic Books.

DUBBER, MARKUS D., FRANK PASQUALE & SUNIT DAS (2020) *The Oxford Handbook of Ethics of AI*, New York, NY: Oxford University Press.

ERMAN, EVA & MARKUS FURENDAL (2022) »The Global Governance of Artificial Intelligence: Some Normative Concerns» *Moral Philosophy and Politics*, 9, ss. 267–291. Tillgänglig online på: https://doi.org/10.1515/mopp-2020-0046.

FREY, CARL BENEDIKT (2019) *The Technology Trap: Capital, Labor, and Power in the Age of Automation*, Princeton, NJ: Princeton University Press.

FREY, CARL BENEDIKT & MICHAEL A. OSBORNE (2017) »The Future of Employment: How Susceptible Are Jobs to Computerisation?», *Technological Forecasting and Social Change*, 114, ss. 254–280.

HÄGGSTRÖM, OLLE (2021) *Tänkande maskiner: den artificiella intelligensens genombrott*, Stockholm: Fri tanke.

LIAO, S. MATTHEW (red.) (2020) *Ethics of Artificial Intelligence*, New York, NY: Oxford University Press.

PARFIT, DEREK (1984) *Reasons and Persons*, Oxford: Clarendon Press.

RUSSELL, STUART (2019) *Human Compatible: Artificial Intelligence and the Problem of Control,* New York: Viking.