AI i EU

Etiska riktlinjer som styrmedel

av Stefan Larsson

Under 2010-talet har betydande framsteg gjorts inom området för artificiell intelligens (AI) och i synnerhet inom ramen för maskininlärning. Delvis för att främja denna utveckling antog EU en strategi för AI i april 2018. En expertgrupp med 52 medlemmar (AI HLEG) utsågs för att ge råd om såväl investeringar som etiska styrningsfrågor när det gäller AI i Europa. Expertgruppen publicerade i april 2019 etiska riktlinjer för ”tillförlitlig AI”, som – trots att man explicit påpekar att riktlinjerna inte hanterar rättsliga spörsmål – tydligt lyfter fram ansvarsfrågor, transparensfrågor samt dataskydd som helt centrala delar för utvecklingen av en tillförlitlig AI. Expertgruppens riktlinjer är inte ett isolerat fenomen utan kan tvärtom ses som del av en växande trend där en mängd etiska riktlinjer gällande AI tagits fram av både företag, forskningssammanslutningar och statliga företrädare. Många överlappar delvis redan befintlig lagstiftning, men det är ofta oklart hur lagstiftning och riktlinjer mer precist är tänkta att samspela. Särskilt är det många gånger otydligt hur etiska principer är tänkta att implementeras. Riktlinjerna fokuserar på normativa ställningstaganden, men är i allmänhet processuellt svaga.

På ett mer generellt plan är de etiska riktlinjer gällande AI som EU-kommissionens expertgrupp tagit fram ett tydligt tecken på en pågående styrningsutmaning för EU och dess medlemsstater. Samtidigt har ordföranden för EU-kommissionen, Ursula von der Leyen, under sin kandidatur uttryckt att hon under sina första 100 dagar i tjänst ska ”lägga fram lagstiftning om en samordnad EU-strategi för de mänskliga och etiska konsekvenserna av artificiell intelligens”. Exakt

vilka delar av de vidsträckta riktlinjerna hon avser att juridifiera är ännu oklart. En del av utmaningen ligger rimligen i att balansera reglering mot den tilltro som finns till teknisk innovation, och samhällelig utveckling överlag, som AI och dess metoder kan bidra till och som man därmed inte vill riskera att dämpa med reglering.

I förevarande kapitel refererar jag till AI och maskininlärning delvis som *självlärande teknologier*. Detta för att betona å ena sidan elementen av autonomi eller agens som utmärker dessa teknologier och, å andra sidan, den lärande, mönsterfinnande och prediktiva aspekt som mycket av den nutida maskininlärningen handlar om, vare sig det enbart är i form av en digital mjukvara eller i form av en integrerad del i en fysisk produkt. I dess vida betydelse är användningsområdena för AI och ytorna för interaktion med människor många, speciellt för de informationsintensiva och mer digitala miljöerna. I takt med ett ökat samhälleligt användande och beroende av AI och maskininlärning ökar också samhällets behov av att förstå dess eventuellt negativa konsekvenser, hur intressen och makt fördelas och vilka behov det finns av både rättslig och annan typ av styrning. Detta kapitel fokuserar på etiska riktlinjer som styrmedel, för att peka på sådana riktlinjers samspel med rättsliga styrmedel och diskutera vad det är med AI-utvecklingen som särart som lett till att just etiska frågor fått en så framskjutande plats. Det pågår en bredare diskussion om styrningen av AI – ofta uttryckt i termer av ”AI governance” – som detta kapitel bidrar till genom att anlägga ett europeiskt perspektiv. Framförallt behandlar kapitlet nämnda etiska riktlinjer för s.k. tillförlitlig AI som EU-kommissionens expertgrupp publicerat. Kapitlet är strukturerat kring följande tre frågor:

1. Hur kan man förstå en databeroende AI utifrån dess samspel med mänskliga värderingar och samhälleliga strukturer?
2. Varför tar reglering av AI så ofta formen av etiska riktlinjer?
3. Vilka etiska principer är vanligast utpekade som mest centrala för styrningen av AI och varför?

Den första frågan föranleds av att själva AI-definitionen är omdebatterad, och kan skifta beroende på vilket disciplinärt fält den som

definierar termen har sin bas i. Jag argumenterar här för behovet av att se teknologierna i deras tillämpade kontexter och i deras samspel med mänskliga värderingar och samhällsstrukturer, vilket inte minst understryks av maskininlärningens beroende av stora mängder data eller exempel att utgå från. AI-system kan därmed reproducera inte bara positiva och avsedda uttryck och strukturer utan även mänskligheten problematiska sidor och beteendemönster, såsom ojämställdhet och diskriminering. Därutöver finns de uppenbara riskerna för missbruk av kraftfulla teknologier för illvillig användning av olika slag. Den andra frågan får placeras in i ett bredare forskningsfält kring styrmedel bestående av olika inriktningar som dock ofta delar några principiella och centrala värden som bestämmande över data, grad av rimlig transparens och hur ansvar bör fördelas. AI-fältet föranleder därför många vägar fram med olika styrfunktioner och styrmedel, både i form av lagstiftning, standardiseringsdialoger och, inte minst, etiska riktlinjer. Den tredje frågan är tudelad på så vis att jag dels pekar ut vilken typ av AI-relaterade utmaningar som får störst uppmärksamhet, och dels pekar på de framväxande insikterna om samhällsrelaterade utmaningar som därmed ligger till grund för de etiska riktlinjerna. En viktig aspekt som frågan visar och som ofrånkomligen behöver hanteras är det temporala gapet mellan en långsam och stundom utdragen, men demokratiskt och politiskt förankrad, lagstiftningsprocess och den oerhört snabba utvecklingen som kännetecknar AI och dess underliggande element.

För att kunna svara på frågorna behövs inledningsvis en utvecklad beskrivning av AI, som tar hänsyn till den rikedom av infallsvinklar som finns inom forskningslitteraturen. En sådan beskrivning presenteras i nästa avsnitt. Därefter förtydligar jag vad det är som behöver styras, med utgångspunkt i växande insikter inom kritisk AI-forskning gällande oförutsedda konsekvenser av tillämpad AI, speciellt i termer av diskriminerande eller skevt reproducerande utfall av algoritmiska processer. Både oförutsägbarhet och utmaningar med att kunna förklara hur algoritmiska processer och AI-system når ett visst resultat, utfall eller löser ett specifikt problem har lett till att transparensfrågan kommit att bli helt central. Därefter riktar jag in mig på begreppet AI governance och hur idéer om styrning

av AI specifikt kommit till uttryck. Jag gör det utifrån insikter om det konstanta samspelet mellan samhälle och AI, med fokus dels på hur tillämpad AI hela tiden behöver utvärderas utifrån samhällets normer och etik, och dels på hur dagens AI-system ofta är beroende av stora mängder data för att kunna tränas upp för att lösa problem. Dessa data inbegriper i många fall bilder på människor eller en kvantifiering av mänskliga uttryck och sociala strukturer, vilket gör att dagens AI i allra högsta grad samspelar med samhällets alla delar, inklusive mindre önskvärda beteenden. Jag använder här begreppet *society-in-the-loop* för att visa på hur innovation behöver samverka med samhällets förväntningar och behov. Detta innebär också en utveckling av argumentationen kring varför AI behöver styras. I nästföljande avsnitt presenterar jag styrningsidéerna i sig, med fokus på etiska riktlinjer, vilket är kärnan i detta kapitel. Som redan nämnts ovan, har ett växande antal riktlinjer tagits fram globalt sett. Jag berör några av dessa riktlinjer och lyfter fram betydelsen av tillväxten i sig. Vidare fördjupar jag det europeiska perspektivet genom att analysera de hittills mest centrala etiska riktlinjerna, nämligen de som tagits fram av EU-kommissionens expertgrupp, AI HLEG. Avslutningsvis presenterar jag rekommendationer och möjliga vägar fram, med avsikten att kunna peka ut utvecklingsområden och relevanta frågor för lagstiftare, myndigheter och de som beforskar, utvecklar och tillämpar AI-system.

Vad är AI?

Trots den uppmärksamhet som AI och maskininlärning får i både media och europeiskt policyarbete råder det inte någon konsensus kring hur AI bäst bör definieras. En rad definitioner har lanserats inom såväl forskning som i myndighetsrapporter, men en stor utmaning ligger i att det rör sig om ett dynamiskt och föränderligt fält. Jag vill här dels poängtera dynamiken i begreppsapparaten så som den har diskuterats inom traditionell AI-forskning, dels presentera några centrala element som ändå går att utmejsla, samt visa på vilka aspekter expertgruppen AI HLEG fäster vikt vid. Avslutningsvis vill

jag också, i ljuset av de utmaningar som AI i sin tillämpning och interaktion med samhällets värderingar och strukturer har visat upp, argumentera för att det finns en poäng med, sett från ett flervetenskapligt perspektiv, att inte för ensidigt luta sig mot en datavetenskapligt förankrad definition av AI. Själva definierandet är en form av konceptuellt styrande som har effekter på regleringsdebatten, varför man också behöver vara nogsam när man utarbetar definitioner för ett så mångfacetterat begrepp som AI.

I samband med att EU-kommissionens expertgrupp, AI HLEG, offentliggjorde sina etiska riktlinjer för tillförlitlig AI publicerades också ett definitionsdokument som är menat att klargöra vissa aspekter av AI som vetenskaplig disciplin och som teknik. Ett uttalat syfte med dokumentet är att undvika missförstånd och att uppnå en gemensamt delad kunskap om AI som kan användas även av icke-experter. Vidare kan det visa på detaljer som kan bidra till diskussionen om de etiska riktlinjerna. HLEG utgår först från den definition som gavs i EU-kommissionens meddelande om AI för Europa, publicerat i april 2018, vilken de därefter utvecklar. I kommissionens meddelande definieras AI enligt följande:

> Artificiell intelligens avser system som uppvisar intelligent beteende genom att analysera sin miljö och vidta åtgärder – med viss grad av självständighet – för att uppnå särskilda mål.
>
> AI-baserade system kan vara helt programvarubaserade och fungera i den virtuella världen (t. ex. röstassistenter, bildanalysprogram, sökmotorer, tal- och ansiktsigenkänningssystem), eller inbäddas i hårdvaruenheter, (t. ex. avancerade robotar, självkörande bilar, drönare eller applikationer för sakernas internet).

Definitionen tar framförallt fasta på autonomin, dvs. att det finns ett mått av agens i AI-system, och påpekar att systemen kan vara både inbäddade i fysiska enheter såväl som rena mjukvarusystem. Samtidigt ger exemplen en fingervisning om vad expertgruppen har i åtanke, och i förlängningen vad de etiska riktlinjerna har för styrningsobjekt. Inom den mjukvarubaserade kategorin av AI-system pekar de på

röstassistenter, bildanalytisk mjukvara, sökmotorer, röst- och ansiktsigenkänning. Inom de fysiska, hårdvarubaserade applikationerna identifierar man avancerade robotar, autonoma fordon, drönare och de uppkopplade prylar som ses som en del av ”sakernas internet” (*Internet of Things*, IoT). Eftersom autonomin poängteras kan man tolka det som att det inte gäller alla drönare eller alla uppkopplade saker utan endast de som har ett autonomt eller lärande element. Frågan om vad som kännetecknar en ”avancerad” robot kan inte nödvändigtvis besvaras genom en enkel gränsdragning. Det knyter an till en betydande svårighet med AI-definierandet som har att göra med AI:s dynamiska sida. Definitionen av AI inbegriper i någon mån sådant som ännu inte är uppnått. En rapport från Stanforduniversitetet, skriven av en grupp AI- och robotikforskare, refererar till fenomenet som ”AI-effekten” eller ”uddaparadoxen”, i betydelsen att när en AI-teknologi blir allmängods betraktas den inte längre som AI. På samma sätt som en ”avancerad robot” inte alls är samma sak år 2020 som det var i början på 1990-talet, har även den konceptuella gränsdragningen för vad som kännetecknar AI ändrats i takt med vad som har blivit möjligt att genomföra och hur dessa metoder blir tillgängliga för en bredare användning.

AI HLEG noterar också att intelligensbegreppet, som är en explicit del av AI, är ett särdeles svårfångat koncept som funnits med sedan grundandet av forskningsområdet. I en essä från 2007 samlar forskarna Shane Legg och Marcus Hutter över 70 olika definitioner av intelligensbegreppet. De visar på hur AI-forskningens definitioner har tagit fasta på olika ingående aspekter, med olika emfas på problemlösning, förbättring och lärande över tid, god prestanda i komplexa miljöer, eller generaliserbarheten i att kunna lösa olika typer av problem utan specifik träning på varje enskild typ av problemdomän. Sistnämnda beskrivs ofta som en del i strävan efter att uppnå en generell intelligens, som till skillnad från dagens snäva, domänbegränsade AI, är tänkt att kunna lära sig en övergripande intelligent hållning som klarar att lösa olika sorters komplexa problem. Ofta hänvisas här till människans dynamiskt intelligenta förmågor som en förebild. Intelligensbegreppet väcker dock också en rad associationer till mänskliga förmågor utöver de problemlösande, som

t.ex. att ha känslor och självmedvetande. Sådana förmågor kan inte sägas utgöra en del av de metoder och teknologier som karakteriserar tillämpad AI år 2020 och de är därmed inte föremål för styrning genom etiska riktlinjer. Det hindrar inte att det finns inriktningar inom AI-forskningen som strävar dithän och forskar på sådana frågor.

Man kan därmed konstatera att AI vid 2020-talets början primärt är ett paraplybegrepp som inkluderar en rad olika teknologier och analysmetoder, såsom: maskininlärning, naturlig språkinlärning, bildigenkänning, s.k. neurala nätverk och djuplärande. Framförallt maskininlärning kan betonas, som enkelt uttryckt, handlar om metoder för att få datorer att ”lära” sig utifrån data utan att datorerna har programmerats för just den uppgiften. Maskininlärning är ett fält som har utvecklats oerhört starkt under slutet av 2010-talet genom tillgången till historiskt ojämförbart stora digitala datamängder och kraftigt ökande analytisk processorkraft. ”Machine learning” som begrepp myntades redan 1959 av AI pionjären Arthur Samuel, som skapade ett av världens första självlärande spelprogram. Sedan dess har fältet dock gått från att vara en underdisciplin till AI med huvudmålet att eftersträva artificiell intelligens till att bli ett mer praktiskt orienterat forskningsfält, där prediktion ligger i fokus, baserad på träningsdata. Området brukar numera räknas till AI, men är också nära kopplat till statistik och bildigenkänning, där maskininlärning har visat sig vara väldigt användbar. Centralt för maskininlärning specifikt, men också AI generellt, är de algoritmer som används, utvecklas och studeras för att skapa lärande effekter i mjukvara och ge sannolikhetsbedömningar.

Komplexiteten i begreppsapparaten leder AI HLEG till att föra fram en tämligen mångfacetterad definition som därmed utvidgar EU-kommissionens ursprungliga definition av AI. Denna expanderade definition inkluderar även AI-funktionaliteten i sin systemiska kontext (dvs. att den ofta ingår i ett större sammanhang), maskininlärningens uppdelning mellan strukturerad och ostrukturerad data, samt att AI-system i huvudsak är måldrivna för att uppnå något som en människa har definierat:

> Artificiella intelligenssystem (AI-system) är programvarusystem (och eventuellt även hårdvarusystem) som har konstruerats av människor och som, när de får ett komplext mål, agerar i den fysiska eller digitala dimensionen genom att uppfatta sin omgivning via datainsamling och att tolka insamlade strukturerade eller ostrukturerade data, resonerar om den kunskap eller behandlar den information som härletts ur denna data och beslutar om den bästa åtgärd eller de bästa åtgärderna som ska vidtas för att uppnå det fastställda målet. AI-system kan använda symboliska regler eller lära sig en numerisk modell. De kan också anpassa sitt beteende genom att analysera hur den omgivande miljön har påverkats av deras föregående åtgärder. (EU-kommissionen, AI HELG, ”En definition av AI: Viktigaste förmågor och vetenskapliga discipliner”, s. 6).

Det finns således olika aspekter av AI att ta fasta på i en definition av fenomenet, där de mest centrala för dagens AI-utveckling och användning tenderar att kretsa kring a) autonomi/agens, b) självlärande utifrån stora datamängder, och c) graden av generaliserbart lärande. Man talar här ibland om att vi – trots den snabba utvecklingen på området – befinner oss inom ramarna för en svag eller smal intelligens där problemområdena fortfarande är väldigt smalt definierade. Det finns en strävan inom i vart fall delar av forskningen om att få fram mer generellt intelligenta applikationer, som därmed skulle kunna överföra insikter från en specifik domän till andra områden.

Noterbart är att trots att expertgruppens huvudsakliga dokument med etiska riktlinjer för tillförlitlig AI handlar om etiska, rättsliga och i mångt och mycket samhällsvetenskapliga och humanistiska frågor så finns inte dessa ämnesområden med när AI beskrivs som forskningsdisciplin.

> Som vetenskaplig disciplin innefattar AI flera metoder och tekniker, t.ex. maskininlärning (som fördjupad inlärning och förstärkt inlärning är specifika exempel på), maskinresonemang (som omfattar planering, schemaläggning, kunskapsrepresentation och resonemang, sökning och optimering) och robotik (som omfattar kontroll, perception, sensorer och manöverdon samt integrering av all övrig teknik i cyberfysiska system).

Avsaknaden av hänvisning till andra discipliner i denna definition är intressant, särskilt med tanke på att även dokumentet pekar på utmaningar med förklarbarhet och orättvis snedvridning (bias), dvs. data som är partisk eller skev på något sätt. Det finns med andra ord en konceptuell tudelning mellan expertgruppens definition av ”AI-forskning” som en i huvudsak matematiskt förankrad data- eller mjukvaruorienterad forskningsdisciplin och de rättsvetenskapliga, humanistiska och samhällsvetenskapliga AI-relaterade forskningsbehoven. Det är dock möjligt att det bara är en tidsfråga innan forskningsfälten tydligare möts och de olika vetenskapliga bakgrunderna återspeglas i definitionen av AI-forskning.

Jag har tidigare forskat på betydelsen av hur den metaforiska förståelsen av digitala och abstrakta fenomen påverkar hur de regleras. Genom att studera utvecklingen inom digital fildelning kunde jag visa på hur upphovsrätten expanderade i sin tolkning av digitala kopior som exemplar i upphovsrättens mening, vilket både utmanade den traditionella regleringens analogt förankrade tankegods men också påverkade synen på fildelningsfenomenet. I linje med det menar jag att beskrivningen och förståelsen av AI – dess inneboende koncept och metaforer – behöver förvaltas med viss lyhördhet för hur teknologierna samspelar med samhället i sin användning och utveckling. Det betyder rimligen också att man bör se definitionsprivilegiet som en viktig del i, eller ett förstadium till, utvecklingen av reglering och styrning av AI. Det finns med andra ord skäl till att hålla sig något kallsinnig till de matematiskt och datavetenskapligt grundade definitionerna av AI, framför allt när syftet är att bättre förstå vad dessa teknologier och metoder betyder i sin samhällstillämpning och vad det medför för regleringsbehov.

Vad är det som behöver styras?

AI, oavsett om man kallar det självlärande teknologier eller autonoma system, inrymmer en oerhörd potential genom att erbjuda individuellt relevanta tjänster, förbättrade möjligheter till preventiva bedömningar och automatisering av beslutsfattande i så skilda domä-

ner som sjukvård och självkörande fordon. Utvecklingen har gått fort när det gäller maskininlärningens kapacitet, med neurala nätverk, s.k. djuplärande *(deep learning)* och metoder som generativa adverseriella nätverk (s.k. GAN:s) som kan generera syntetisk information som möjliggör skapandet av realistiska, men falska, bilder. Potentialen är stor för en rad informationsberoende fält, som handel, sjukvård och offentlig förvaltning. Både privata och offentliga forskningsfinansiärer riktar forskningsmedel till en allt starkare, men i hög grad ingenjörstillvänd AI-forskning. Samtidigt växer insikterna om etiska och normativa utmaningar med tillämpad AI, vilket jag, tillsammans med andra forskare i en flervetenskaplig grupp, visat i en nyligen genomförd kunskapsöversikt på området för ”hållbar AI”. Detta sker i takt med att fler forskningsdiscipliner intresserar sig för frågorna utifrån sina perspektiv och teoribildningar, tillsammans med teknikforskare eller enskilt, och att mer av denna typ av kritisk forskning också finansieras.

Ett fokus i den moderna AI-utvecklingen ligger på just lärandet, dvs. att de underliggande modellerna anpassas och modifieras baserat på de data – de exempel – som modellerna presenteras för. En prediktion, diagnos eller individuell anpassning blir därmed inte bättre än vad underliggande data tillåter. I takt med att AI blir vardag – i våra sociala medieflöden, i musikrekommendationstjänsterna och i bankernas riskbedömningar – är det samhällets strukturer och individernas beteenden och värderingar som används som data och mängden data är oerhört stor. Det betyder också att oönskad social snedvridning (bias) och ojämställda förhållanden som samhället huserar riskerar att reproduceras i AI-tjänster. Detta samtidigt som teknologiernas komplexitet och den digra informationsmängden gör processerna svårgranskade och svåröverblickbara, därav jämförelsen med en svart låda (”black box”).

Om vi går närmare in på de snedvridande effekter så finns det AI-system som har visat sig ha sämre precision för kvinnor och för personer med mörk hy, vilket i värsta fall lett till rasistiska och könsdiskriminerande utfall. Ett sådant flagrant exempel på oförutsedda konsekvenser i mötet mellan AI och samhället har handlat om automatiserad rasism i ett system för bedömning av återfallsrisk i det

amerikanska domstolssystemet (s.k. recidivism). Ett annat exempel är könsdiskriminerande jobbrekommendationssystem som per automatik rekommenderade arbeten med högre lön till män. Det har även visats hur välanvända bilddatabaser med skev kulturell, köns- och etnicitetsfördelning ger oönskade effekter för lärande algoritmer. På applikationsnivå kan den här typen av bias leda till att kamerafunktioner som ska tipsa fotografen om när den fotograferade blinkar tolkar det som att asiater alltid blinkar eller att en bilddatabas automatiskt taggar svarta människor som ”gorillor”. Utöver det mänskliga lidande som sådana tillämpningar kan ge upphov till så skapar de även stora risker inom områden där bedömningen kan ha särskilt allvarliga konsekvenser, t.ex. inom sjukvård, försäkring och finansiella tjänster. Dessa normativa och etiska frågor utgör några av den moderna AI-forskningens underbeforskade utmaningar, vilket väcker ett reglerings- och styrningsbehov. Jag utvecklar den här typen av normativa utmaningar i relation till rättssystemets mångahundraåriga utveckling i en essä om ”Sjyst AI” från 2018.

Behovet av styrning uttrycks exempelvis av juridik- och teknologiforskare Urs Gasser och kollegan Virgilio Almeida i en artikel från 2017. De menar att regeringar, civilsamhällets aktörer, den privata sektorn och akademin gemensamt behöver diskutera styrningsmekanismer som kan minimera riskerna och de möjliga nackdelarna med AI och autonoma system samtidigt som de utnyttjar hela teknikens potential. Forskarna pekar särskilt på behovet av att säkerställa transparens, ansvarsskyldighet och förklarbarhet för vad de kallar ”AI-ekosystemet”. Här har vi också många av de kategorier som finns inom det framväxande fält som kallas FAT, dvs. *Fairness, Accountability and Transparency* (rättvisa, ansvarsutkrävande och transparens). Med FAT poängteras att algoritmiska system används i ett antal sammanhang som med hjälp av stora datamängder *(Big Data)*, filtrerar, sorterar, betygsätter, rekommenderar, ”personifierar” och på andra sätt formar mänskliga erfarenheter och förhållanden. Även om dessa system ger många fördelar har de också inneboende risker, såsom kodifiering och förstärkande av samhällelig snedvridning (bias), reducerad ansvarsskyldighet och ökad informationsasymmetri mellan dataproducenter (kunder) och datainnehavare. I rapporten

om Hållbar AI som jag nämnde ovan gör vi en kunskapsöversikt om ”fairness” i termer av både rättvisa och orättvis snedvridning, dvs. diskriminerande konsekvenser av autonoma system. Vi diskuterar även illvillig användning, dvs. missbruk, som en särskild kategori. Sådant missbruk återanknyter till ansvarsfrågan – dvs. vilket ansvar de bör ha som utvecklar AI när det kommer till hur produkter och AI-innovationer missbrukas av en annan part. Detta diskuteras även i åtskilliga rapporter inom fältet.

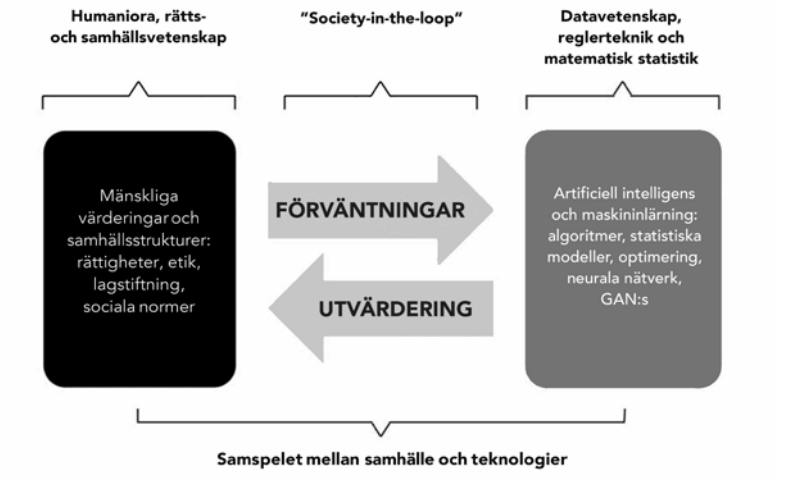
Vad betyder *AI governance*? *Society-in-the-loop*

Det finns olika sätt att förstå governance i förhållande till AI. En rapport från *Future of Humanity Institute* vid universitetet i Oxford lägger fokus på de institutioner och sammanhang där AI byggs och används. Speciellt, menar man i rapporten, kan man förstå AI governance som inriktad på att maximera oddsen för att människor som bygger och använder avancerad AI har de mål, incitament, resurser, stöd, samt den tid, utbildning och organisatorisk hemvist som krävs för att göra det till förmån för mänskligheten. Just detta institut leds av filosofiprofessor Nick Bostrom, som är mest känd för sin forskning om s.k. superintelligens, dvs. idén om att maskiners intelligens ska bli så generell att mänskligheten kan komma att bli underordnad och utsättas för en existentiell risk.

Detta kapitel fokuserar dock på de framväxande etiska riktlinjerna inom AI, och därmed på governance i en mer mondän och pragmatisk mening som uttrycks i fråga 1 ovan. Delar av AI-utveckling och tillämpning är redan reglerad, speciellt när det gäller tillgång till personuppgifter i den form som omfattas av Dataskyddsförordningen (GDPR). Det finns behov av att bättre tolka och förstå vad implementeringen av GDPR i förhållande till exempelvis automatiserat beslutsfattande (enligt Art. 22), inte minst i ljuset av en ökad praktisk utbredning av automatiserat beslutsfattande. Detta oaktat så kännetecknas styrningsfältet av mycket hög aktivitet på området för etiska riktlinjer, framtagna med ambitionen att påverka utvecklingen och tillämpningen av AI. Detta framför allt när det gäller att hantera de

praktiker man menar är särdeles problematiska. Innan vi tittar närmare på de etiska riktlinjer som tagits fram, och vad dessa innebär i termer av styrning eller governance, behöver vi vidare utveckla resonemanget kring relationen mellan AI och samhällets normer och värderingar.

Hur man definierar AI debatteras flitigt, vilket bör förstås i relation till en expansion av teknologiernas användningsområden och möjligheter, men också till hur de används i en alltmer vardaglig kontext på både marknader och i offentlig förvaltning. Nämnda Gasser och Almeida konstaterar också att en orsak till svårigheten att definiera AI från ett tekniskt perspektiv är att AI inte är en enskild teknologi. AI är snarare ”en samling tekniker och underdiscipliner som sträcker sig från områden som röstigenkänning och datorseende till varseblivning och minnesfunktioner, för att nämna några”. Det finns därmed en poäng med att betona växelverkan mellan AI som en å ena sidan teknologisk, matematisk och datavetenskaplig idé och disciplin och å andra sidan de samhälleliga värderingar som den samspelar med. Detta speciellt gällande databeroende maskininlärning, där exemplen som algoritmer tränas på utgår från människors beteenden, normer och värderingar. Denna andra sida studeras bland annat av rättsvetare, etiker och andra samhällsvetare och humanister.



FIGUR 1. BEARBETNING FRÅN IYAND RAHWANS ARTIKEL ”SOCIETY-IN-THE-LOOP: PROGRAMMING THE ALGORITHMIC SOCIAL CONTRACT” (2018).

Av särskilt intresse utifrån perspektivet samhällstillämpad AI, dvs. AI inom till exempel konsumentmarknader och i offentlig verksamhet, är att förstå och utvärdera de maskinlärande teknologierna utifrån samhällets värderingar, normer och etik. Detta leder till ett flervetenskapligt forskningsbehov. Den data- och samhällsvetenskapliga forskaren Iyad Rahwan belyser just det här samspelet mellan inomteknologisk utveckling och samhällets värderingar i termer av behovet av att hålla ”society-in-the-loop” som en underbyggnad för att säkerställa ”samhällskontraktet” (se figur 1). Inom mjukvaruutveckling finns uttrycket human-in-the-loop för att beteckna en viss typ av problemlösning där människan är med i själva designen för problemlösningen. Skräppostfilter och musikrekommendationer, eller för den delen det individuella Facebookflödet, kan beskrivas som lärande system där människors kontinuerliga individuella input spelar en vital roll för själva problemlösningen: att avgöra vad som är skräppost, vilken musiktyp som lyssnaren gillar, vilken typ av media och vilka vänner som användaren finner mest relevanta. Rahwan lyfter den idén till samhällsnivån.

Inspirerad av Rahwans iterativa koncept men kompletterat med den rättssociologiska och samhällsvetenskapliga forskning jag riktat in mig på, kan man konstatera att behovet av att konstant utvärdera nya självlärande teknologier uppstår av åtminstone tre anledningar:

1. För att avgöra sjystheten *(fairness)* i de självlärande teknologiernas utfall. Sjystheten kan ses som både en etisk fråga och en rättslig, och därmed också en demokratisk – vilket naturligtvis studeras inom en rad vetenskapliga discipliner. Ur ett perspektiv kan man anföra att sjystheten i vart fall delvis avgörs genom ett kommunikativt samtal *(deliberation)* snarare än en optimerande expertprocess. Med bas i den vetenskapsteoretiska litteraturen kunde man därmed uttrycka det som skillnaden mellan vad som kan ses som objektivt bäst i en optimerbarhetsmening och vad som kan ses som mest acceptabelt i en demokratisk och intresseavvägande mening. Samtidigt, vilket jag lyfter fram i essän om Sjyst AI, har samhällets sökande och byggande av en balanserad rättsordning

krävt en månghundraårig utveckling för att sätta sig. Att normativt avgöra sjysthet – från latinets *ius* för rätt – är därför ingen lättvindigt kalkylerbar uppgift utan något som tagit samhällen lång tid att finna strukturerna för.

2. Lärandet utgår från oss – våra strukturer och beteenden – dvs. vi *är* de exempel som många AI-system lär sig från. Det betyder att samhällets informella strukturer och normer rent empiriskt kommer att ligga till grund för artificiella agenters beteende. Samhällets värderingar ska här inte enbart ses som vad vi tycker att mänskligheten bör stå för, som att skydda mänskliga rättigheter, utan även problematiska sidor av vad mänskligheten de facto delvis är – ojämställd, diskriminerande och rasistisk. Vidare medför det att redan befintliga idéer om rättigheter för individer att bestämma över sin information aktualiseras. Med andra ord, även om utvecklad precision hos en viss typ av algoritmer skulle kräva tillgång till en stor mängd personuppgifter eller patientdata så kommer tillgången nödvändigtvis bero på en samhällelig intresseavvägning mellan värdet att utveckla AI och värdet av individuellt självbestämmande.
3. Missbruk är alltid att vänta. Kraftfulla teknologier kommer att användas för många olika syften, även kriminella, bedrägliga och repressiva. I februari 2018 publicerade en grupp AI-fokuserade säkerhetsforskare en rapport som tecknar kommande områden där AI kommer att användas för illvilliga syften, till skada och kriminalitet. Den hotbild som denna forskargrupp tecknar inkluderar utvecklade varianter av cyberattacker som automatiserad hackning, och risken för distansövertagande av uppkopplade autonoma fordon, som därmed kan användas i fysiska attacker, till exempel för att styras in i folkmassor. Det inkluderar också politisk och polariserande användning av botnätverk för att påverka val.

I linje med sistnämnda kan metoderna bakom GAN nämnas. De har till en början kommit att fokusera på skapandet av fejkade syntetiska ansiktsbilder med realistiskt utseende, med startpunkt i Ian

Goodfellows banbrytande artikel från 2014. I takt med att fotorealismen i de fejkade bilderna har stärkts har också s.k. *deep fakes* kommit att användas i bedrägligt syfte. Sådana bilder har med andra ord inte bara använts för att skapa fotorealistiska datorspel och en talande Mona Lisa, utan även för att trakassera kvinnor genom att t.ex. framställa manipulerade nakenbilder (s.k. fejkhämndporr). I kombination med vad som ibland kallas Barbara Streisandeffekten – det som hamnar på internet, kopieras, distribueras och stannar på internet – blir de trakasserande konsekvenserna i värsta fall långtgående, bestående och återkommande. Ett annat exempel på möjliga missbruk av AI-innovationer som blev känt under 2019 var hur kriminella hade använt AI-genererade röstimiterande metoder för att via telefon pressa en underordnad på ett företag, genom att simulera chefens unika röst, att föra över pengar till ett ungerskt konto. Att landvinningar inom AI kan och kommer missbrukas på en rad olika sätt, i exemplen gällande autenticitet, är en av de utmaningar som uppstår i interaktionen mellan AI och samhälle. Det knyter an till frågan om governance och vilka etiska riktlinjer som krävs för AI-utveckling, inte minst vad gäller ansvarsfrågor vid design och utveckling av kraftfull s.k. multifunktionsteknologi *(multipurpose technologies)*.

EU – etiska riktlinjer för tillförlitlig AI

Om vi först blickar mot de diskussioner som förs kring AI och etik på en global arena kan man konstatera att det för närvarande är ett livaktigt ämne bland akademiker och policyorienterade organ. Just etiska riktlinjer som styrverktyg har haft en påfallande stark utveckling under slutet av 2010-talet. Teknikfilosofen Brent Middelstad konstaterade under 2019 exempelvis att åtminstone 84 initiativ har tagit fram principiella förhållningssätt för att guida en etiskt försvarbar AI-utveckling, användning och styrning. Google och Telia är exempel på företag som offentliggjort etiska principer för sitt arbete med AI, och det relativt unga forskningsinstitutet AI Now har blivit känt för sina publikationer på området. En annan studie om det globala AI-etiklandskapet, gjord av forskare vid det schweiziska *Health Ethics &*

Policy Lab, konstaterar att det finns en relativ samstämmighet globalt kring i vart fall fem principiella förhållningssätt av etisk karaktär: 1.) transparens, 2.) rättvisa och ”fairness”, 3.) icke-skadlig användning, 4.) ansvar, och 5.) integritet/dataskydd. Samtidigt konstaterar studien att det finns betydande skillnader i hur dessa principer tolkas, varför de anses viktiga, vilken fråga, domän eller vilka aktörer de avser och hur de ska implementeras.

Svagheten med etiska riktlinjer, menar etikforskare Thilo Hagendorff i en artikel från 2019, är att AI-etik – liksom etik generellt – saknar mekanismer för att skapa efterlevnad eller implementera sina normativa anspråk. Detta är enligt Hagendorff också skälet till att etikorienterade åtgärder är så tilltalande för många företag och institutioner. När företag och forskningsinstitut formulerar sina egna etiska riktlinjer, återkommande för in etiska överväganden eller antar etiskt motiverade egna åtaganden, så menar Hagendorff att bindande rättsliga ramverk motarbetas. Han lägger därmed stor tonvikt vid just undvikande av reglering som ett huvudsyfte för AI-industrins etiska riktlinjer. Liknande farhågor uttrycker Mark Coeckelbergh, professor i media och teknologiers filosofi vid universitetet i Wien, i en artikel från 2019 om rättsliga utmaningar med AI-etik. Coeckelbergh, som också är en av medlemmarna i AI HLEG, konstaterar att ”det finns en risk att etik används som ett fikonblad som hjälper till att säkerställa acceptabel teknik och ekonomisk vinst men inte har några betydande konsekvenser för utvecklingen och användningen av teknologierna” (min övers.). Även om påpekandet har en poäng – det är otvetydigt så att många företag visar upp sådan ”självreglering” (utveckling av intern policy med svag egentlig implementering) i syfte att undvika hårdare extern reglering – kan det ändå finnas andra skäl till att etiken som styrmedel har betonats så starkt inom AI-utvecklingen. Frågan är om inte det specifika AI-fältets snabba tillväxt spelat en väl så betydande roll för att just detta fält krävt ett mjukare tillvägagångssätt i väntan på att den kritiska forskningen hinner ikapp och erbjuder en stabil grund för potent reglering. En fråga är dock vad en juridifiering av AI-etik skulle medföra, och vilka delar av den som lämpar sig bäst för lagstiftning.

Som nämnts i inledningen har den nytillträdda ordföranden för EU-kommissionen, Ursula von der Leyen, uttryckt en vilja att ta fram

lagstiftning på området för mänskliga och etiska konsekvenser av AI. Innan man kan föra en sådan diskussion behöver vi titta närmre på EU-kommissionens initiativ på AI-området. EU antog en strategi för AI i april 2018 och utsåg en expertgrupp på hög nivå (AI HLEG) för att ge råd om såväl investeringar som etiska styrningsfrågor när det gäller AI i Europa. I december 2018 presenterade kommissionen en samordnad plan – "Made in Europe" – som utarbetades tillsammans med medlemsstaterna i syfte att främja utvecklingen och användningen av AI i Europa. Till exempel uttryckte kommissionen att alla medlemsstater till mitten av 2019 bör ha egna strategier på plats, vilket dock inte blivit fallet. Expertgruppen utsågs genom en öppen utlysning och består av en tämligen blandad skara av forskare och universitetsföreträdare (exempelvis inom robotik, datavetenskap och filosofi), industrirepresentanter (som Zalando, Bosch och Google) och civilsamhällessammanslutningar (som Access Now, ANEC, BEUC). Sammansättningen har dock inte undgått kritiska röster. Exempelvis Yochai Benkler, professor vid Harvard Law School – känd för sina optimistiska arbeten om kollaborativa ekonomier med fokus på fenomen som Wikipedia, Creative Commons och öppen källkod – framförde i maj 2019 en farhåga för att industriella företrädare i för hög grad tillåts styra regleringsfrågorna runt AI. Benkler drog paralleller mellan EU-kommissionens expertgrupp, Googles fallerade råd för AI-etikfrågor och Facebooks investeringar i ett tyskt AI- och etikforskningscentrum. När de etiska riktlinjerna publicerades uttrycktes även kritik från medlemmar i själva expertgruppen. Thomas Metzinger, filosof från universitetet i Mainz, menade i en intervju att de utkast som togs fram om förbud för vissa användningsområden, som autonoma vapensystem eller motsvarigheter till det kinesiska sociala kreditsystemet, hade tonats ned av industriella företrädare och deras allierade.

Under sitt första år har AI HLEG tagit fram:

1. Etiska riktlinjer för tillförlitlig AI, som poängterar vad expertgruppen kallar ett människocentrerat *(human-centric)* synsätt på AI och listar sju viktiga krav som AI-system bör uppfylla för att vara tillförlitliga. Dessa krav testas i sin tur i

ett slags pilotprocess som kommer att bidra till en reviderad variant av de etiska riktlinjerna under 2020.

2. Policy- och investeringsrekommendationer. I linje med de etiska riktlinjerna har gruppen lagt fram 33 rekommendationer som är tänkta att vägleda en tillförlitlig AI mot hållbarhet, tillväxt och konkurrenskraft, men även vad de kallar inkludering – dvs. en förhoppning om att kunna styra mot en utveckling som kan stärka, gynna och skydda människor. Tanken är att dessa rekommendationer ska hjälpa kommissionen och medlemsstaterna att uppdatera sin gemensamma samordnade plan för AI.

Även om det återstår att se vilken typ av betydelse båda dessa källor kommer att ha på europeisk AI-utveckling så har kommissionen placerat expertgruppen i en anmärkningsvärt central position för att påverka riktningen. AI HLEG är också styrgruppen för den s.k. European AI Alliance, som är ett forum med blandade intressenter som syftar till att stimulera en bred och öppen diskussion om alla aspekter av AI-utveckling och dess påverkan på ekonomi och samhälle.

Expertgruppen pekar i de etiska riktlinjerna ut tre komponenter av tillförlitlig AI som bör finnas med under AI-systems hela livscykel:

a) den bör vara *laglig* och följa alla gällande lagar och förordningar,
b) den bör vara *etisk* och säkerställa att etiska principer och värden upprätthålls, och
c) den bör vara *robust* ur både teknisk och samhällelig synvinkel, eftersom AI-system kan orsaka oavsiktliga skador, trots goda intentioner.

Riktlinjerna fokuserar på etiska frågor (b.) och robusthet (c.) men lämnar rättsliga spörsmål – (a. ovan) utanför riktlinjerna. Detta till trots kan man konstatera att dokumentet inom ramen för etiken ändå kommer in på tämligen rättsligt förankrade frågor, i termer av ansvarsfrågor, transparens och inte minst dataskydd. Precis som expertgruppen konstaterar så styrs mycket av AI-utveckling och an-

vändning i Europa av redan befintlig lagstiftning. Det gäller dels stadgan om de grundläggande rättigheterna, GDPR, direktivet om produktansvar, direktiv mot diskriminering, konsumentskyddslagstiftning m.m. Även om förutsättningar för etisk och robust AI till viss del redan tillförsäkras av befintlig lagstiftning på europeisk och på nationell nivå, kan dess fulla förverkligande gå utöver befintliga rättsliga skyldigheter. Man kan konstatera att hur man reglerar individers bestämmande över sina personuppgifter är en central fråga för hur databeroende AI utvecklas. Dataskyddsförordningen generellt, men även mer specifikt kompletterande europeisk eller nationell lagstiftning, som t.ex. den svenska patientdatalagen, spelar avgörande roll för hur både marknadsaktörer och myndigheter tar sig an AI-relaterade frågor om riktad marknadsföring, automatiserat beslutsfattande och precisionsmedicin.

FIGUR 2. FRÅN AI HLEG, 2019. I RAPPORTEN FRÅN EU-KOMMISSIONENS EXPERTGRUPP BETONAS ATT ALLA KRAV ÄR LIKA VIKTIGA, STÖDER VARANDRA OCH BÖR GENOMFÖRAS OCH UTVÄRDERAS UNDER ETT AI-SYSTEMS HELA LIVSCYKEL.

Expertgruppen anger sju huvudsakliga förutsättningar för att uppnå eller operationalisera tillförlitlig AI, som ska utvärderas och hanteras fortlöpande under AI-systemets hela livscykel (se figur 2).

1. Mänskligt agentskap och mänsklig tillsyn: AI-system bör bli en källa till rättvisa samhällen genom att vara ett stöd för mänsklig medverkan och grundläggande rättigheter, i stället för att minska, begränsa eller underminera mänsklig självständighet.
2. Teknisk robusthet och säkerhet: Tillförlitlig AI kräver algoritmer som är tillräckligt säkra, tillförlitliga och robusta för att hantera fel eller inkonsekvenser i AI-systemens alla arbetsfaser.
3. Integritet och dataförvaltning: Medborgarna bör ha full kontroll över sina egna data. Dessa data får heller inte användas till att skada eller förfördela dem.
4. Transparens: Betonar AI-systems spårbarhet, förklarbarhet och kommunikation.
5. Mångfald, icke-diskriminering och rättvisa: AI-system bör beakta människans alla grader av begåvning, färdigheter och krav samt garantera användarna tillgänglighet.
6. Samhällets och miljöns välbefinnande: AI-system bör användas till att stärka positiv social förändring samt öka hållbarheten och det ekologiska ansvaret.
7. Ansvarsskyldighet: Det bör införas mekanismer för att säkerställa ansvar och ansvarsskyldighet för AI-system och resultaten av deras processer, inklusive möjlighet till granskning och rapportering av negativa konsekvenser.

Gällande de investerings- och policyrekommendationer som expertgruppen också har publicerat så rekommenderar man bland annat ett riskbaserat tillvägagångssätt som är både proportionerligt och effektivt för att försäkra att AI är lagenligt, etiskt och robust i sin anpassning till grundläggande rättigheter. Expertgruppen efterlyser en omfattande kartläggning av relevant EU-reglering för att bedöma i vilken utsträckning olika lagstiftningsinstrument fortfarande uppfyller sina syften i en AI-driven värld. De understryker att nya rättsliga

åtgärder och styrningsmekanismer kan behöva införas för att säkerställa adekvat skydd mot negativa effekter och möjliggöra korrekt tillsyn och verkställighet.

Ett intressant exempel som ger en viss indikation om hur stundande regleringsdiskussioner kommer arta sig kan tas från den tyska dataetikkommissionen, som i slutet av oktober 2019 publicerade en rapport om etiska riktlinjer med rekommendationer för det breda uppdraget om att bevara social sammanhållning, skydda individer och att säkerställa välstånd i en informationsdriven era. Den tyska kommissionen gör en poängfull tredelad uppdelning mellan algoritmbaserat beslutsfattande, AI i sig och data. De tre är näraliggande och sammanhängande komponenter, men kräver ändå enskilt fokus. Detta gör att kategoriseringen blir annorlunda än den lanserad av EU-kommissionens AI-expertgrupp. Ändå är dataetikkommissionens arbete styrt av delvis samma behov som EU-kommissionens AI-expertgrupp, exempelvis gällande människocentrerad design, integritet och självbestämmande, ansvarsfull datahantering och att länka samman digitala strategier med hållbarhetsmål. För de ”algoritmiska systemen” poängteras vikten av transparens, förklarbarhet och tydliga ansvarsstrukturer, vilket är helt i linje med andra riktlinjer och de kunskapsområden som vuxit fram i enlighet med FAT.

Intressant nog konstaterar den tyska dataetikkommissionen att reglering är nödvändig och inte kan ersättas av etiska principer. De betonar också vikten av att krav ställs i relation till risker, dvs. ju högre risk för användningen av ett visst algoritmiskt system, desto högre krav på transparens, granskning och utvärdering. De öppnar också för ett strikt förbud för de allra mest riskabla applikationerna. Riskbedömning är rimligen ett lämpligt förhållningssätt för att man ska kunna sortera bland krav på reglering och interveneringsmetoder. Dataetikkommissionen föreslår även en märkning av algoritmiska system i linje med riskbedömningen. En invändning, som inte är menat att avskräcka, kunde vara att det inte alltid är så enkelt att teckna riskbilden – även den beror på hur man värderar risk och vad man jämför med. Ska man jämföra med riskerna i de teknologier som det algoritmiska systemet är tänkt att ersätta eller de potentiella riskerna med den enskilda teknologin i sig? T.ex. de eventuella risker som

finns med självkörande bilar kontra de redan etablerade riskerna med människokörda fordon. En del risker är inte heller enkla att bedömda för att de kan vara av systemisk karaktär och visa sig först när teknologin missbrukas. Det kan då handla om riktad marknadsföring som i en form kan innebära relevanta erbjudanden för konsumenter, men i sin skalbarhet (många användare) och ökade individualiserade precision kan manipulera konsumenter, ibland kallat ”hypernudging”. Teknologin kan också missbrukas genom att öka politisk polarisering i syfte att påverka demokratiska val, vilket är en omfattande debatt i kölvattnen av Facebooks relation till Cambridge Analytica.

Dataetikkommissionen i Tyskland föreslår inrättandet av ett nationellt centrum med särskild kompetens kring algoritmiska system, i syfte att bistå tillsynsmyndigheter i deras uppdrag. Vidare poängteras vikten av forskning för s. k. förklarbar AI (”explainable AI”, XAI) som innebär en strävan efter att förbättra algoritmiska systems förklarbarhet, särskilt självlärande system. Man rekommenderar därför att den federala regeringen ska finansiera ytterligare forskning och utveckling inom detta område.

Den komplexa transparensen

Insikterna inom den kritiska AI-relaterade forskningen har vuxit fram relativt snabbt. Dessa gör bland annat gällande att det finns betydande icke avsedda, negativa, konsekvenser med samhällsinteragerande självlärande teknologier. Det förstärker behovet av att bättre förstå rättvisa och ”fairness” i relation till algoritmiska och autonoma system samt även de frågor och utmaningar som avser ansvarsfördelning. Detta är också något som hänger ihop med svårigheterna att förstå och förklara (transparens) vissa utfall av vad som ibland kallas ”black box”-system. Det är rimligen insikterna i denna kunskapstillväxt som bidragit till utformningen av de etiska riktlinjerna för tillförlitlig AI. Det har i litteraturen kring etiska riktlinjer relaterade till AI argumenterats för att just *transparens* inte är en etisk princip i sig, utan snarare ett ”pro-etiskt villkor” för att möjliggöra eller ange förutsättningar för andra etiska praktiker eller principer. Som

jag utvecklar i en artikel från 2019 om den artificiella intelligensens rättssociologiska relevans finns det flera intressemotsättningar som kan knytas till transparensfrågan. Det finns även fler skäl än ren teknikkomplexitet till varför vissa upplägg kan vara av black-box-karaktär, inte minst befintligt skydd för företagshemligheter och immaterialrättsligt ägande, samt hur komplexa datadrivna marknader har kommit att bli.

Ett intressant perspektiv på AI-governance presenteras i en rapport från det amerikanska forskningsinstitutet AI Now. Rapporten handlar om vad man på svenska kan kalla algoritmisk konsekvensbedömning. Forskarna lånar från miljöbedömningslitteraturen för att peka på behovet av utvärdering och uppföljning av implementerade AI-system, med beslutsfattande i offentlig sektor i särskilt fokus. Rapportförfattarna summerar fem centrala insikter att beakta:

1. Myndigheter bör göra en självbedömning av befintliga och föreslagna automatiserade beslutssystem, utvärdera eventuella konsekvenser för fairness/laglighet, partiskhet/bias eller andra problem i de berörda grupperna.
2. Myndigheter bör utveckla externa forskningsbaserade granskningsprocesser för att upptäcka, mäta eller spåra effekter över tid.
3. Myndigheter bör meddela allmänheten sin definition av ”automatiserat beslutssystem”, befintliga och föreslagna system samt eventuella relaterade självbedömningar och granskningsförfaranden innan systemet har förvärvats.
4. Myndigheter bör begära offentliga kommentarer (samråd) för att klargöra problem och svara på frågor (dialog).
5. Regeringar bör tillhandahålla förbättrade rättsliga mekanismer för berörda personer eller grupper, för att utmana otillräckliga bedömningar eller orättvisa, snedvridna eller på annat sätt skadliga systemanvändningar som myndigheter inte har lindrat eller korrigerat.

Dessa rekommendationer för en ”algoritmkonsekvensbedömning” relaterar tydligt till pågående diskussioner om hur offentlig sektor kan effektiviseras genom AI, ofta uttryckt i termer av automatiserat

beslutsfattande eller AI-genererat beslutsstöd. Mycket tyder dock på att fältet behöver studeras mer. Vi behöver bland annat förstå relationen mellan automatiserat beslutsfattande, beslutsstöd och mänskligt beslutsfattande som en governancefråga. Många kommande applikationer, inte minst inom offentlig sektor och det medicinska fältet, kommer att ha olika tonvikt vid dessa tre aspekter. Samtidigt tycks denna problematik sällan vara central om man ser till de etiska riktlinjerna. Hagendorff, nämnd ovan, konstaterar till exempel i sin studie av 15 etiska riktlinjer att ”ingen riktlinje behandlar i detalj den uppenbara frågan där system för algoritmiska beslutsfattande är överlägsna respektive underlägsna mänskliga beslutsrutiner”. AI HLEG betonar dock vikten av mänskligt agentskap och mänsklig tillsyn.

Integritet och de rättigheter som det europarättsliga dataskyddet medför är en helt central punkt i styrning och utveckling av AI. Man kan dock beskriva det som att det finns ett tilltagande intresseavvägningsbehov mellan individens rätt till bestämmande och algoritmutvecklingens behov av data. Frågan pressas av allt att döma fram på en rad områden, inte minst gällande medicinska applikationer på området för life sciences. Eftersom det finns en stor nyttopotential för s.k. precisionsmedicin och utvecklat prognos- och diagnosställande, skapar detta, i kombination med en stark kommersiell sektor, också ett starkt tryck på tillgång till individers hälsoinformation. Det finns exempel på kontroverser i stil med när Googles AI-företag Deep Mind 2016 genom avtal med Londonsjukhuset Royal Free, som drivs i nationella hälsovårdsmyndighetens (NHS) regi, bereddes tillgång till patientinformation utan att patienterna eller deras anhöriga hade tillfrågats. Exemplet ställer intresseavvägningen på sin spets, men väcker även frågan om transparens och skiljelinjen mellan privat och offentlig verksamhet.

Det finns också kritisk dataskyddsforskning som menar att det moderna dataskyddet i viss mån redan innehåller omoderna drag och är dåligt anpassat till dagens marknadspraktiker, vilket medför ett slags förklarbarhetsutmaning. Oxfordforskaren Sandra Wachter, i en artikel från 2019 tillsammans med filosofen Brent Mittelstadt, argumenterar för en rätt till ”reasonable inferences”, dvs. att få veta

vilka statistiska antaganden som analytiska datainsamlare gör om en. Forskarna menar att den oro som uttrycks för AI-genererade analysers brister, när det gäller möjlighet till ansvarsfördelning, ofta i själva verket är en oro för det sätt på vilket dessa tekniker drar integritetinkräktande slutsatser om oss som vi inte kan förutsäga, förstå eller motbevisa. Wachter och Mittelstadt menar att analyser baserade på stora mängder data, vilket är fallet hos många av dagens digitala plattformstjänster och AI, drar icke-verifierbara slutsatser om individers beteenden, preferenser och privatliv. Dessa slutsatser bygger på mycket diversifierade och funktionsrika datamängder av oförutsägbart värde och skapar dessvärre nya möjligheter för diskriminerande, partiskt och invasivt beslutsfattande. Den intuitiva länken mellan handlingar och uppfattningar riskerar därmed att försvinna, vilket leder till individers brist på kontroll över sin identitet och hur vi uppfattas av andra.

Ett exempel kan vara digital marknadsföring, som bygger på långtgående antaganden om samband mellan de mest skilda ting. Med väldigt mycket data om en stor mängd företeelser, preferenser och beteenden kopplat till individer kan man nämligen, genom prediktiv analys och maskininlärning, hitta samband som inte nödvändigtvis behöver vara kausala men som ändå indikerar en förhöjd sannolikhet för det ena om också det andra föreligger. Sådana antagna samband (”inferred” brukar användas som begrepp i den engelskspråkiga litteraturen på området) har visat sig vara användbara. Dock, eftersom analyserna i kommersiella sammanhang oftast inte är transparenta för granskning, kan ett utfall av en automatiserad och sannolikhetsbaserad analys (med en räntenivå, premienivå eller individuellt riktad marknadsföring som konsekvens) innebära en utmaning för tillsynsverksamhet och dataskyddsreglering. Det sistnämnda bygger i mångt och mycket på att var och en ger samtycke, eller inte, till att deras personuppgifter lämnas över och hanteras. Det antagna samband som en analys av många individers sammantagna data innebär är varken en direkt eller observerad information som täcks av dataskyddet, men är likväl central för att sälja, marknadsföra eller på annat sätt påverka människor. Forskarna argumenterar för att individens rättigheter därmed kan behöva expanderas till att gälla även för dessa analytiskt antagna samband.

Samtidigt leder argumentationen till att man behöver se över organisatoriska förutsättningar och applikationers sammanhang för att kunna förstå hur de rättsliga styrningsutmaningarna för AI-utveckling och användning ser ut. Det är stor skillnad på offentlig förvaltningsreglering och dess behov jämfört med globalt aktiva megaplattformars skalbarhet och multijurisdiktionella upplägg. Ovan nämnde Coeckelbergh påpekar svårigheten med implementering av etisk AI i relation till megaplattformarna (se även Lundqvists kapitel i denna bok), genom att konstatera att ”det är svårt att se hur ansvarsfull innovation verkligen kan genomföras när det finns en maktkoncentration i ett relativt begränsat antal kraftfulla aktörer, inklusive ett litet antal stora företag: det verkar som att en handfull företag beslutar framtiden för AI”. För att förstå hur individer kommer i vardaglig kontakt och interagerar med tillämpad artificiell intelligens – så som ansiktsigenkänning, riktad reklam, och innehållsmoderering – och vilka förutsättningarna är för reglering och implementering av etiska riktlinjer, så behöver man se hur spänningsfältet ser ut inte minst gentemot de globala storskaliga plattformarna. De digitala plattformarnas möjligheter och utmaningar är även föremål för analys i en antologi från 2019 om ”plattformssamhället”, som medievetare Jonas Andersson Schwarz och jag är redaktörer för. I mitt kapitel i volymen analyserar jag storskaliga plattformarnas användande av AI och vad det medför i termer av skalbarhetsutmaningar, brist på transparens och mjukvarukodad styrning av användarnas beteende.

Från principer till verkningsfull implementering

Jag har ovan argumenterat för att svårigheten att definiera vad AI är utgör en del i den regleringsutmaning som följer av tillämpning och utveckling av AI. Min argumentation baseras framförallt på behovet av att se dagens maskininlärningsbaserade och databeroende AI i dess relation till samhällets strukturer och mänskliga värderingar. Ett skäl är att det för många typer av applikationer är mänskliga uttryck – ansikten, geografisk rörelseinformation, beteende i sociala medier, m. m. – som utgör de stora mängder träningsdata som precisionen i

AI-applikationerna är beroende av. Det betyder att de regleringsmässiga utmaningarna består i den kraftfullhet och potential till agens som finns i AI-metoderna, men också i att metoderna reproducerar samhällets obalanser. Databeroendet hos dagens maskininlärning i kombination med den komplexitet som skapar brist på förklarbarhet riskerar att resultera i att samhällets skevheter inte bara reproduceras utan även förstärks, samtidigt som sådana missförhållanden är svåra att upptäcka. Man kan också konstatera att kraften i de metoder som ger precision i applikationer som bildanalys, beteendeprediktion eller möjligheten att generera syntetisk data har ökat mycket på väldigt kort tid. Det i sig skapar en regleringsutmaning eftersom konsekvenserna av den snabba ökningen tar tid att förstå och utvärdera. Lagstiftningsprocesser kräver i sin tur eftertanke och deliberation för att mäkta med att finna de åtråvärda samhällsbalanser som kan anses rimliga mellan de olika intressen som dagens AI-utveckling relaterar till; individers självbestämmande, informationsasymmetrier och maktförhållanden, risker för diskriminering, manipulation och monopol; behov av transparens och tillsyn; utveckling av användbara och effektiva applikationer, industriella behov av tillväxt och äganderättsligt skydd, m. m. Denna temporala aspekt, i kombination med den månghövdade intresseavvägningen, är troligen en betydande förklaring till varför styrningen på området i mycket karaktäriseras av etiska riktlinjer under 2020-talets början. I ljuset av de teman som tas upp i detta kapitel finns det några centrala frågor att fokusera framöver:

- Först det principiella, nu det processuella

Floran av etiska riktlinjer för AI-utveckling och användning är rik på principiella ställningstaganden, men fattig på processuella upplägg. Det är rimligen en mognadsfråga, där principiella överväganden är det nödvändiga första steget. Det processuella efterkommande steget är dock nödvändigt, både för att stärka implementeringsmöjligheterna av de principiella ställningstagandena och för att säkerställa tillförlitlig AI som medborgare, myndigheter, konsumenter och företag vågar använda, förlita sig på och investera i. Kopplat till AI

HLEG utvecklas exempelvis en AI-utvärderingslista i linje med detta. Om man förstår tillväxten av etiska riktlinjer som ett uttryck för snabbheten i AI-metodernas utveckling så blir det processuella steget ett väntat andra steg. Om man däremot ser de etiska riktlinjerna som ett resultat av företagens motsträvighet mot reglering av sina verksamheter, en mjuk variant av lagstiftning som är avsedd att vara tandlös, så kommer det processuella steget att möta motstånd. Detta processuella steg innebär i förlängningen även här en metodappell för behovet av att utveckla myndigheters tillsynsverksamhet, som ett led i en praktisk implementering av befintlig reglering. Det som AI HLEG uttrycker som ett behov av granskningsbarhet (”auditability”), kan delvis överföras på de tillsynsmyndigheter som har ansvar att säkerställa marknadens funktion, vilket är väl så behövligt i relation till de globala plattformarnas monopolistiska tendenser och komplexiteten i de datadrivna marknadernas ekologi. De AI-drivna delar av marknaderna som kan styra individualiserad annonsfördelning, prissättning och annat behöver vara granskningsbara, vilket kräver en metodutveckling. I linje med detta föreslår exempelvis nämnda tyska dataetikkommission en central kompetensgrupp med syfte att bistå tillsynsmyndigheterna i deras uppdrag.

- Det flervetenskapliga AI-forskningsbehovet

Det är hög tid för humanistisk och samhällsvetenskapligt orienterad AI-forskning. Många av kunskapsbehoven kommer rimligen kräva samverkan mellan de matematiskt förankrade datavetenskapliga disciplinerna, som har djupa insikter i hur AI-system byggs och verkar, och de humanistiskt och samhällhällsvetenskapligt orienterade discipliner som kan teoretisera och öka förståelsen för AI:s samspel med kulturer, normer, värderingar, attityder eller betydelse för makt, marknader, stater och reglering. Det samspel, som betecknas *society-in-the-loop* ovan, där mänskliga uttryck och samhällets strukturer utgör själva träningsdatan leder bland annat till normativa frågor om AI-system där intresseavvägning och värdegrunder hamnar i fokus snarare än optimering och värdeneutral målstyrning. Kontraintuitivt nog handlar mycket av utmaningarna med autonoma

maskiner om människans skevheter. I den människocentrerade ansatsen (”human-centric approach”) som ofta uttrycks i etiska riktlinjer för AI-utveckling så finns något av en idealiserad föreställning om vad mänskligheten innebär i termer av värderingar och samhällsstrukturer. Omvänt skulle man, från ett beteendeempiriskt förhållningssätt, kunna konstatera att det ofta är tillämpade uttryck för mänskliga värderingar och skeva sociala strukturer som leder till automatiserade misslyckanden. För en databeroende AI som lär från exempel i stora mängder information uppstår helt enkelt ett lärande inte bara från goda och balanserade exempel utan även från mänsklighetens mindre stolta sidor: rasismen, främlingsfientligheten, löneojämlikheterna mellan könen och de informella men strukturellt utbredda orättvisorna. Utmaningar här blir därmed att normativt sortera underliggande data, alternativt att vikta de självlärande teknologiernas automation och skalbarhet så att de reproducerande och utvidgande tendenserna blir bättre och mer balanserade än det underliggande materialet. Kort uttryckt: samhällstillämpad AI medför ett flervetenskapligt forskningsbehov.

- Lagstifta klokt och kunnigt

Historien lär oss att reglering är svårt, speciellt i tider av snabb teknologiskt driven samhällsförändring. Samtidigt lär oss rättsvetare som Karl Renner, som analyserade äganderätten under Västeuropas industrialisering, också att juridiken kan vara oerhört dynamisk och anpassningsbar. Givet uttalandet från EU-kommissionens ordförande, Ursula von der Leyen, kan man tänka sig att centrala delar av de etiska riktlinjerna kan komma att formaliseras i lagtext. Även tolkningen av redan befintlig nationell och europeisk lagstiftning i ljuset av AI-systems funktionalitet, möjligheter och utmaningar är en fråga av enorm betydelse och omfattning. Dels är snabbheten i förändringen svår att hantera i relation till långsamheten i traditionell reglering – vilket jag ser som en av de grundläggande orsakerna till varför styrningen inom AI-området i så hög grad karaktäriseras av etiska riktlinjer. Dels är kunskapen om AI:s fundamentala konsekvenser för samhälle, marknader och individer fortfarande bara

under uppbyggnad. Detta utgör ett regleringsmässigt dilemma som snarast talar för att man bör anlägga ett mjukare och mer lyhört tillvägagångssätt istället för att införa hård reglering. Å andra sidan, om man tar i beaktande den kritik som gör gällande att starka industriella företrädare ser de etiska riktlinjernas brist på sanktioner som en chans att komma undan med ett samhälls- och konsumentskadligt beteende så finns det ett stort behov av tuffare tag på övergripande nivå.

AI-frågan har kommit att ta en värdegrundsbaserad och etikcentrerad utveckling inom det europeiska samarbetet. Det är ett svar på frågan hur man ser på AI och dess kvaliteteter som jag finner högst lovvärd: självlärande och autonoma teknologiers precision behöver bedömas i sitt samspel med samhällets värderingar, inte bara i ett teknologiskt vakuum. Det är en normativ definition med bäring på framtida utvecklingslinjer – en bra AI är en samhällsförankrad och tillförlitlig sådan.

Källor och litteratur

Det finns en växande litteratur kring frågor om etik, ansvarsfördelning och behovet av transparens i fältet FAT *(Fairness, Accountability and Transparency)*. FAT har även en återkommande konferens där forskare presenterar studier kring kritiska frågor i relation till användning av AI-system och maskininlärning som ofta filmas. Utöver rapporter från EU-kommissionens expertgrupp har IEEE publicerat en rapport om *Ethically Aligned Design* som kan betraktas som något av ett lexikon för att finna fler källor på den tematik som presenteras i detta kapitel.

I boken *The Black Box Society. The Secret Algorithms that Control Money and Information* (Harvard University Press 2015) visar juridikprofessor Frank Pasquale på den problematiska bristen på transparens för konsumenter och användare i relation till datainsamlande plattformsaktörer. Om man vill förkovra sig i de medvetet illvilliga användningsområdena för AI kan man läsa rapporten "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitiga-

tion" (2018), tillgänglig via https://maliciousaireport.com/.

När det gäller behovet av tillsynsmetoder argumenterar jag i artikeln "Algorithmic Governance and the Need for Consumer Empowerment in Data-driven Markets", (*Internet Policy Review,* 2018) om behovet av ett utvecklat konsumentskydd på algoritmdrivna marknader. Se även min text "Den kvantifierade konsumenten" (*Fores Policy Breif* 2018:1). I artikeln "The Socio-Legal Relevance of Artificial Intelligence"(*Droit et Société*, 2019), utvecklar jag transparensfrågan gällande AI, och argumenterar för behovet av mer rättsvetenskaplig forskning på området. I en tämligen tillgänglig antologi om *Plattformssamhället. Den digitala utvecklingens politik, innovation och reglering*, (Fores, 2019) diskuterar Jonas Andersson Schwarz och jag (reds.) samhällsimplikationerna av den datadrivna organisationsstruktur som plattformar medför, med fokus på beroendet av några få globala jättar.