Sjyst AI och normativ design

Larsson, Stefan

Published in:
Människor och AI

2018

Document Version:
Förlagets slutgiltiga version

12. Sjyst AI och normativ design

Av: Stefan Larsson
Stefan Larsson är jurist, dr. i rättssociologi och docent i teknik och social förändring på LTH, Lunds universitet samt programchef för Digitala samhället, tankesmedjan Fores.

Om AI och samhälle

När artificiell intelligens blir vardagligt applicerad i samhället – på konsumentmarknader, i sjukvården, trafiken och på digitala plattformar – väcks helt avgörande frågor kring hur ansvar bör fördelas och vilka normer som bör styra. I ljuset av AI som en lärande teknologi som interagerar med människor och sociala strukturer blir det också centralt att förstå vilka normer och värderingar som reproduceras av de autonoma processerna själva. Den här essän visar på några centrala rättsliga och rättssociologiska frågeställningar i gränslandet AI och samhälle, där normer och normativitet har visat sig vara en lika nödvändig som evigt svårhanterlig fråga för alla samhällen, speciellt aktualiserad i tider av snabb teknikmedierad förändring.

Sjyst AI

Vad är rättvisa? Hur ska ett samhälle formulera vad som är sjyst? – eller snarar juste, som är mer tydligt etymologiskt kopplat till latinets *ius*, rätt. Frågan är central för alla samhällen, påtagligt i historiska källor i allt från Hammurabis gammalbabylonska lagar till Platons dialoger, i den romerska rätten, via tänkare som Friedrich Nietzsche, Max Weber och Emile Durkheim och fram till dagens rättsordningar. Hur rättvisa balanseras och definieras är en fråga som ständigt behöver närvara, gnuggas och åter och åter ställas i relation till teknologiska förutsättningar, sociala normer och ideologier. Vilken etik, vilka normer bör gälla?

Teknikskiften, vet vi, kan vara så stora och samhällspåver-

kande att de tämligen drastiskt organiserar om balanser och maktförhållanden i ett samhälle så att de rentav kräver nya sätt att förstå rättvisa. Johann Gutenbergs teknik för boktryckeri bidrog till en samhällsomdanande upplysning och ny masskommunicerande världsordning. Industrialiseringen tecknade senare om de europeiska förutsättningarna att gå från feodala strukturer till stater och – ofta – demokratier. Folkets normer tilläts – krävdes – få påverka rättsordningarna. Nu digitaliseras vår värld, samhället vi lever i, med oändliga mängder information inom räckhåll. Med data om individernas alla förehavanden kan vi – eller den aktör som positionerar sig till att förvalta informationen – lära sig mänsklighetens beteenden i detalj. Också detta teknikskifte stuvar om, det möjliggör nya ordningar och omöjliggör andra.

Maskininlärning och artificiell intelligens – som teknikskifte – kommer med en betydelse och kvalitet vi inte har sett förut: Tingen lär sig. Maskinerna får agens, blir alltmer självständiga och beslutsföra. De reproducerar och förstärker redan befintliga sociala strukturer när de sätts att interagera med våra vanor, bilder och språk. Och de smyger sig in – på grund av den potential de har för det redan etablerade att effektivisera, individualisera och robotisera – i samhällenas, marknadernas och individernas vardag: i medicinskt diagnosställande, i kreditvärderingar, i nyhetsdistribution, i polisiära metoder, i riktad reklam, i försäkringsbolagens riskbedömningar, i spamfiltren, i bedrägeriförsöken, i aktiehandeln, i musikrekommendationerna, i jobberbjudandena och i fordonen.

Under detta löfte om ökade intäkter, effektivisering och individuell relevans finns en uppenbar – men ändå i sina kodade detaljer – dold strävan efter att se in i framtiden. Att prediktera. Att analytiskt sannolikhetsbedöma händelser som ännu inte har hänt. Vilket beteende är mest sannolikt? Hur förbättras det individuella kunderbjudandet? Vilken diagnos är troligast? Vilka polisiära resurser ska läggas var? Vem kommer att betala tillbaks ett lån; vem kommer att skada sig; vilken är din livslängd?

Sjystheten ställer en svår fråga: hur förhåller man sig rättvist i en tvist om en analytisk framtid?

Vilken transparens behövs för att säkerställa att bedömningen är rimlig? Vilken förklarbarhet behövs för att vi ska känna tillit för den bedömning som gjorts? Frågan säger något om maskininlärningens konservativa sida, dess beroende av en gyllene standard att jämföra med, en databas som definierar korrekta beslut att lära från. Frågan ställer också krav på att förstå mer av vad denna automation ger, eller snarare hur tingens och kodens *autonoma* beslut påverkar olika intressen. När maskinerna får agens behöver vi diskutera behovet av transparens än mer. I vilken grad, för vilka parter och när behövs mer eller mindre transparens, bättre förklaringar, mer insyn och tillsyn? Vilket ansvar har den som skapar agens för vad agenten producerar? Ett strikt? Men om agenten förändras genom den eller de som den interagerar med? Hur långt och länge kan ett strikt ansvar sträckas? Ska utvecklaren av ansiktsigenkänning stå till svars för alla efterföljande användningsområden? Inga? Några?

Om vi förutsätter att samhällen innehåller värderingar, sociala normer och kulturer: Vad är det som reproduceras när autonoma processer lär sig av samhällets strukturer? Vad är en sjyst AI? Hur vet vi – hur ser och förstår vi – att den är sjyst när utfallet är en automatiserad sannolikhetsbedömning? Och hur ska dess sjysthet säkerställas i relation till ett samhälle som i sig innehåller skevheter? Är partiskt. Filtrerat i bubblor. Ojämnt maktfördelat. Ojämlikt. Ojämställt. Där stigberoende i både språk och strukturer låser oss till ett kulturellt och konceptuellt arv när vi försöker förstå nya fenomen. Hur säkerställer vi att våra traditionella sätt att förstå inte hindrar oss från att anamma potentialen i det nya? Eller vara blinda för dess risker? Att gamla metaforer inte står i vägen för bättre fungerande förståelser?[66] Och mer teknikförankrat, detaljerat: hur vet vi att underliggande

66 För betydelsen av *hur* vi förstår digitala fenomen för deras reglering, se Larsson (2017) *Conceptions in the Code. How Metaphors Explain Legal Challenges in Digital Times.* Oxford University Press.

träningsdata är lämplig för sitt ändamål? Hur säkerställer vi att designen inte främst reproducerar de värden som dess skapare representerar?

Neutrality lost – när ingen neutral position är möjlig

En oundviklig fråga för den part som designar tjänster som tränas upp på samhällets inneboende strukturella värderingar och förhållanden blir hur samhällelig bias ska hanteras. Forskargrupper av olika slag har uppmärksammat utmaningen:[67] Ska man reproducera världen som den är eller som man eftersträvar att den ska vara? Och vems framtidsvilja avser man?

Det finns rimligen flera algoritmberoende sammanhang som leder till automatiserade *normativa* beslut och som därmed behöver behandlas som just normativa. Hur hanterar man normer och normativitet när man har att göra med en självlärande applikation? Försöker man, som varit vanligt för vissa typer av digitala intermediärer, hävda neutralitet i stil med utsagor om att »vi är *bara* en plattform« och därmed riskera att inte bara reproducera samhällelig bias utan även bidra till den och förstärka den?[68] När bör man ta ansvar, göra normativa ställningstaganden och genom sin design av automatiserade processer motverka samhällelig bias? Svaret är sällan givet. Men allt fler pekar på vikten av ansvarserkännande och tydlighet i ansvarsfördelning i användandet av AI.

Styrkan med »bara en plattform«-argumentet är att man kan undvika oerhört svåra och rimligen konfliktfyllda beslut och låta teknisk och mjukvarurelaterad expertis försöka optimera de system man bygger. Svagheten ligger i att man undviker en kart-

67 Se exempelvis Zou & Schiebingers kommentar i Nature (18 juli 2018): »AI can be sexist and racist — it's time to make it fair«.

68 Se Larsson (2019) »Sju nyanser av transparens: Om artificiell intelligens och ansvaret för digitala plattformars samhällspåverkan,« i Andersson Schwarz, J. & Larsson, S. (red.) *Plattformssamhället. Den digitala utvecklingens politik, innovation och reglering.* Stockholm: Fores.

läggning av de oetiska konsekvenser som kan uppstå, d.v.s. att göra en dålig riskbedömning, och – inte minst – att man riskerar att bygga klandervärda system som diskriminerar och bidrar till samhällets obalanserade strukturer, eller rakt av inte fungerar som tänkt. Och insatsen kan vara stor. Sociala medieplattformar beskylls för att ha en roll i att manipulera politiska val och undergräva demokratiska strukturer; straffvärderingssystem riskerar att bidra till längre fängelsestraff baserade på ovidkommande grunder; medicinskt diagnosställande och autonoma fordon riskerar att döda människor.

Det finns en växande insikt i designsammanhang, indikerat i IEEE-rapporten *Ethically Aligned Design*[69] eller i rapporter från forskningsinstitutet AI Now – att kulturella värderingar och samhällelig bias ofrånkomligen inryms i de data som relaterar till individer som därmed måste hanteras ansvarsfullt. Samtidigt kan man från ett rättssociologiskt perspektiv konstatera att det inte är någon enkel sak eller finns någon »snabbfix« att hantera normativa ställningstaganden. I avsaknad av en neutral position kommer fler AI-utvecklare ofrånkomligen att nödgas ta normativ ställning i frågor de troligen skulle vilja undvika. Det stärker argumentet om att ingenjörsutbildningar om AI, bildanalys och algoritmer behöver inkludera ansvarsfrågor i relation till samhälleliga eller etiska konsekvenser av den design de lär sig utveckla och applicera.

En databas med bias skulle kunna vara deskriptivt korrekt, d.v.s. beskriva samhället som det är med dess befintliga ojämställdhet, vare sig det gäller mellan kön eller andra grupper. Men den design som reproducerar och »lär sig« baserat på dessa förhållanden utvecklar en aktiv agens i förhållande till den befintliga ojämställdheten. Den ser mönstret och reproducerar det. Bidrar till dess spridning. Förstärker det.

Den som designar måste därmed ta ställning normativt om

69 IEEE (2018). *Ethically Aligned Design. A vision for prioritizing human well-being with autonomous and intelligent systems*. Version 2.

den vill bidra till ojämställdheten eller motverka den. Det finns för många applikationer därmed inte längre någon neutral position eftersom båda alternativen kan kräva kontroversiella ställningstaganden av normativ karaktär. Att reproducera bias och diskriminering kan vara klandervärt – man *borde* motverka – likväl som att motverka kommer att ställa en inför normativa avvägningsutmaningar – vilken balans är egentligen den rätta och vad ger det för andra effekter i ett senare skede? Brist på möjlig neutral position leder till att ett val av normer behöver göras. Ett svårt val, kan det visa sig.

Om normer och kontextualitet

En artificiell intelligens med förmågor att inte bara härma beteende och språkliga konventioner utan även med potential att själv utgöra en autonom aktör med normativ agens kommer att behöva välja normer att lära av. Val av normer är inte nödvändigtvis någon enkel fråga, även om man identifierat att AI-designen är normativ. Nämnda rapport från IEEE konstaterar också detta, där de har samlat några av de mest kompetenta personerna på området för etik och autonoma och intelligenta system. Man konstaterar att om maskiner interagerar med en mänsklig community som autonoma agenter så förväntas dessa agenter att följa de sociala och moraliska normerna i denna community.[70] Ett nödvändigt steg, därmed, för att möjliggöra detta för maskiner är att identifiera dessa normer. Men – frågade man sig helt riktigt – vilka normer?

Också destruktiva och våldsbejakande gruppers normer erbjuder »lärande« för autonoma intelligenser, vilket visar på komplexiteten och det kunskapsbehov om normer som autonom teknologi väcker. Ett exempel – nämnt på fler ställen i den här antologin – i linje med den här problematiken är den självlärande Twitterbotten Tay, som släpptes ut av Microsoft i mars 2016, och inom loppet av några timmar uttryckte rasistiska, antisemitiska

70 IEEE, 2018, s. 36.

och kvinnoförnedrande tweets för att sedan stängas ned av sin skapare. Samtidigt har rimligen experimentet inneburit ett lärande för botutvecklingen, där frågor om att filtrera språk och göra normativa ställningstaganden är en del av diskussionen.

Vid val av normer kan man konstatera att den nedtecknade lagstiftningen kan vara enklare att utröna (men inte alltid) än sociala och informella normer. Och återigen lär den rättssociologiska forskningen oss att relationen mellan lagstiftning och samhälle är en långt mer komplex relation än vad ett striktare dogmatiskt perspektiv på juridik ofta låter hävda. Det finns inte bara gränser för vad lagstiftning kan mäkta med att reglera, vilket den amerikanske rättsvetaren Roscoe Pound beskrev redan 1917, men människornas beteende och förväntningar på vad som betraktas som rätt och fel, exempelvis i trafiken, är dessutom långt ifrån någon direkt översättning av lagen, exempelvis trafikregleringen. Lagstiftning är dessutom svår att dra slutsatser från om man inte vet något om processen som aktualiserar lagen – vem får klaga/överklaga, vem dömer och vilka blir domens konsekvenser? Vissa lagar säger inte ens någon nämnvärt i sig själv utan har nästan allt sitt normativa innehåll definierat i praxis, d.v.s. i prövningarna av lagstiftningen. Det är där detaljerna mejslas ut, oavsett hur välmenande et knippe principer kan se ut att vara.

Vidare kommer insikten om normativ *kontextualitet* att vara viktig för utvecklingen av självlärande artificiella intelligenser, och det kommer uppstå situationer där olika gruppers normer kommer att kollidera och behöva jämkas. Eller, situationerna är redan här. Facebooks innehållsmoderering, för att ta ett exempel, som använder sig av både mänskliga granskares bedömningar, användarflaggning och AI-verktyg, behöver ideligen jämka mellan alla de rättsordningar och kulturella normer som deras över två miljarder användare utgår från. Det är inte *en* community. Och kontexterna befinner sig ofta i konflikt, normativt, om vilken typ av nakenhet som är okej att visa upp, vilken typ av kunskap som bör få spridas och rentav vilken typ av infor-

mation som bör räknas som faktabaserad kunskap. Vad som är hatfylld propaganda och vad som är åsikter som ska vara fria att uttrycka. Samma utmaning finns i modereringen av YouTubes uppemot två miljarder användare där 400-500 timmar material laddas upp varje minut. Skalan och snabbheten kräver automatiserad implementering av policy jämte annan, trögare, mänsklig granskning. Vilket åter pekar på ansvarsfrågan, här specifikt för de digitala plattformarnas moderering.[71] Modereringen, tycks det, har blivit en kärna för vad det är plattformarna erbjuder.[72] Eller, annorlunda uttryckt, *måste* erbjuda, givet den samhällsroll de har kommit att spela. Och AI är en del av verktyget i den normativa verksamheten.

Summa: normativ design

Även om tidslinjerna ser olika ut på olika områden ser utvecklingen ut att gå mot att artificiell intelligens, framför allt maskininlärning, kan ses som en förvisso matematiskt sofistikerad men ändå högst *vardaglig* företeelse, i konstant interaktion på områden med funktioner som är fundamentalt viktiga för samhället. Vår världsbild kureras i sociala medier, diagnoser underbyggs i sjukvården; värdering av kredit, skaderisk och livslängd inkorporeras i bankers och försäkringsbolags metoder; »personalisering« erbjuds i tjänster och prissättning även från storskaliga plattformsaktörer i både digital och fysisk handel som på samma gång kan skapa betydande individuell kundrelevans *och* potentiellt oöverblickbar kapacitet till strukturell diskriminering utan möjligheter till insyn och extern granskning.

Riskerna behöver tecknas tydligare och fler typer av kompetenser behövs i de datavetenskapliga och matematiska designmiljöerna. Vilka data som används och vad den representerar avgör mycket. Maskininlärningens konservativa sida tydliggörs i dess

71 För en utveckling, se Larsson (2018).

72 Se exv. Gillespie (2018) *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.

beroende av tillgängliga och etablerade databaser, som riskerar att utgå från historiska värderingar. Vidare, eftersom samhällets strukturer redan bär på bias, partiskhet och intressekonflikter, kommer den intelligens som interagerar med samhället och lär sig från det, också att färgas, reproducera och möjligen förstärka dessa strukturer. I förlängningen blir designen normativ, och frågan om ansvar riktas mot den som designar, både för oväntade olyckliga utfall men också för svårlösta normativa val som uppstår när ingen neutral position längre är möjlig: vilka normer ska gälla, vad är sjyst AI, vad vore rättvist?

Det är svåra frågor, men nödvändiga, och i kontinuerlig rörelse. Principiella förhållningssätt behöver tydliggöras. Samtidigt är principer inget värt utan en utvecklad process för deras tillämpning. Det rättsordningarnas utveckling lär oss genom historien, genom de teknologiska innovationernas samhälleliga utbredning, är att rättvisa – den formaliserade sjystheten – är en föränderlig organism i ständigt samspel med kulturer, teknologier och marknader – och att den behöver sin process. Sjysthet är ingen historisk seger, det är en evig kamp. Som nu tar plats i utvecklingen av artificiell intelligens.

Referenser

IEEE (2018). Ethically *Aligned Design. A vision for prioritizing human well-being with autonomous and intelligent systems*. Version 2.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Larsson, S. (2017). *Conceptions in the Code. How Metaphors Explain Legal Challenges in Digital Times*. Oxford University Press.

Larsson, S. (2019). »Sju nyanser av transparens: Om artificiell intelligens och ansvaret för digitala plattformars samhällspåverkan,« i Andersson Schwarz, J. & Larsson, S. (red.) *Platt-*

formssamhället. Den digitala utvecklingens politik, innovation och reglering. Stockholm: Fores.
Pound, R. (1917). The limits of effective legal action, *International Journal of Ethics* 27: 150-167.
Zou, J. & Schiebinger, L. (18 juli 2018) »AI can be sexist and racist — it's time to make it fair«, *Nature*, comment.