LUND UNIVERSITY

Artificiell intelligens som normativ samhällsutmaning: partiskhet, ansvar och transparens

Larsson, Stefan

Published in:
Festskrift till Håkan Hydén

2019

Document Version:
Förlagets slutgiltiga version

Link to publication

Citation for published version (APA):
Larsson, S. (2019). Artificiell intelligens som normativ samhällsutmaning: partiskhet, ansvar och transparens. I R. Banakar, K. Dahlstrand, & L. Ryberg-Welander (Red.), *Festskrift till Håkan Hydén* (s. 339-370). Juristförlaget i Lund.

Total number of authors:
1

Creative Commons License:
Ospecificerad

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Stefan Larsson

Artificiell intelligens som normativ samhällsutmaning: partiskhet, ansvar och transparens[1]

”Models are opinions embedded in mathematics.” [2]

Introduktion: AI och samhälle

Under senare år har betydande framsteg gjorts inom området för artificiell intelligens (AI) och i synnerhet inom ramen för maskininlärning (ML). I dess vida betydelse är användningsområdena och ytorna för interaktion med människor många, kanske speciellt för de informationsintensiva och mer digitala miljöerna. Till exempel automatiserad och differentierad prissättning för hotellnätter och flygresor, riktad och individualiserad reklam, men även för informationshanteringen kring både lagerhållning och erbjudanden för kundkortsinnehavare till butikskedjor med fysiska butiker. Även våra hem har i ökande grad självlärande termostater eller annan förment smart ”property technology”, eller underhållning i form av streamingtjänster för film eller musik som använder algoritmer och information om din och andras mediekonsumtion för att rekommendera och ”personalisera” just ditt utbud. Även i direkt livsavgörande sammanhang finns flera användningsområden. Det pågår tester för självkörande eller i olika grad av autonoma fordon i trafiken, utveckling av verktyg för ML-förstärkt cancerdiagnostik, prediktiv analys för riskbedömning hos försäkringsbolag och kreditgivare, och bildigenkänningsalgoritmer som används i sociala medier, polisiärt arbete eller av säkerhetstjänster och för militära ändamål, som i drönare för distanskrigsföring.

---

[1] Jag vill tacka det internationella rättssociologiska institutet i Oñati, Baskien, för min forskarvistelse i det välsorterade biblioteket i juni-juli, 2018. Jag vill också tacka forskningsfinansiären Vinnova och mina kollegor i projektet om etik och hållbar AI samt nätverket AIML@LU för inspiration, kunskap och tid att rikta blickarna åt det här hållet. Och, slutligen, vill jag tacka Håkan Hydén för inspiration för forskning i gränslandet digitalisering och rättssociologi.

[2] Cathy O’Neil, datavetare och författare till boken *Weapons of Math Destruction* (2016).

I det här kapitlet tecknar jag några av de rättsliga och samhälleliga utmaningar som användandet av AI och ML medför, i termer av vad jag kallar normativ design, samhällelig bias i autonoma och algoritmiska system, svårigheter med ansvarsfördelning och behovet av att nyansera vad man avser med transparens och insyn i den här kontexten. Eftersom ansvarsfrågor är tätt knutna till transparensfrågor tecknar jag sju ”nyanser” av transparens nedan, dvs. olika syften och intresseavvägningar som ryms inom transparensbegreppet. Fokus ligger därmed i det här kapitlet inte främst i att tydligt definiera vad AI är enligt ett datavetenskapligt perspektiv, utan vid att peka ut samhällsbetydelsen av en vardagligt och praktiskt applicerad AI utifrån ett samhällsvetenskapligt och rättsvetenskapligt perspektiv (jfr Rahwan, 2018).

I takt med ett ökat samhälleligt användande och beroende av AI och ML ökar också samhällets behov av att förstå eventuella negativa konsekvenser och risker, hur intressen och makt fördelas och vilka behov det finns av rättsliga och etiska ramverk, standarder, certifieringar eller processuella ställningstaganden. Det finns inom litteraturen kring artificiell intelligens redan en tradition av att uttrycka regler och normativa principer för den artificiella intelligens som är tänkt att ha grader av autonomi och agens. Isaac Asimovs tre robotiklagar från 1942 är kanske de mest kända, och dessa har fått flera efterföljare inom robotikforskningen. De tidiga farhågor som föranledde ett slags regleringsbehov och nedtecknad etik relaterade ofta till en tänkt generell intelligens som genom sina insikter och sin analysförmåga kunde vända sig *mot* människan. Idag finns också farhågor uttryckta i termer av en tänkt kommande superintelligens där teknologiska landvinningar kan leda till en uppgraderbar och självförbättrande artificiell intelligens och ett slags ”singularitet” där mänskligheten, så som vi känner den, i stort sett går under.

Detta kapitel fokuserar dock inte en tänkt superintelligens eller generell intelligens, utan nutida mer vardagligt applicerade varianter av artificiell intelligens för att sortera i medföljande rättsliga och rättssociologiska utmaningar. Definitionen av AI för detta kapitel blir därmed bred och inkluderar en rad teknologier och analysmetoder som har kommit att samlas under artificiell intelligens som paraplybegrepp: maskininlärning, naturlig språkinlärning, bildigenkänning, s.k. neurala nätverk och djuplärande. Framförallt ML, som enkelt uttryckt, handlar om metoder för att få datorer att ”lära” sig utifrån data utan att datorerna har programmerats för just den uppgiften, är ett fält som har fått en oerhört stark utveckling över bara de senaste få åren genom tillgång med historiskt ojämförbart stora digitala datamängder och kraftigt ökande analytisk processorkraft. Även om ”machine learning” som

begrepp myntades redan 1959 så har fältet gått från att vara en underdisciplin med målet att eftersträva artificiell intelligens till att lösa mer praktiska problem, där prediktion ligger i fokus, baserad på träningsdata. Området brukar räknas till artificiell intelligens idag, men är också nära kopplat till statistik och bildigenkänning, där ML har visat sig vara väldigt användbar och med åtskilliga praktiska applikationer. Centralt för ML specifikt, men också AI generellt, är de algoritmer som används, utvecklas och studeras för att skapa lärande effekter i mjukvara och ge sannolikhetsbedömningar.

Den betydande skillnaden mellan tidiga AI-inriktade regler och etiska principer är att idag föranleds regleringsdiskussionen av en vardaglig användning av AI och ML sprungen ur en digitaliserad och alltmer datadriven tillvaro. Utgångspunkten här är att en rad samhälleligt centrala funktioner – som påverkar arbetslivet, familjers ekonomiska förutsättningar, hur nyheter och kunskap distribueras och individers sjukvård – medieras idag genom artificiell intelligens. Detta aktualiserar en rad frågor för det rättssociologiska fältet att studera där jag utgår från följande tredelning i detta kapitel:

1. Frågor kring legitimitet och ”fairness” och vilka sociala normer som självlärande autonoma system reproducerar eller förstärker och vilka de *bör* reproducera eller förstärka, exv. när det gäller individers data som kan aktualisera frågor om samhällelig bias med diskriminering utefter kön, etnicitet m.m.;
2. Hur frågor om ansvar och ansvarsfördelning kan problematiseras när alltmer autonom mjukvara – artificiella ”agenter” – tar automatiserade beslut i olika sammanhang;
3. Behovet av transparens och den roll den spelar, inklusive förståelse och insyn i hur automatiserat algoritmmedierat beslutsfattande går till, vilka värden som behöver balanseras och hur olika AI-applikationer i samhället kan eller bör granskas, av vem eller vilken part, och vilken nivå av förklaring, insyn eller förståelse som är behövlig.

Syftet är här att ta ett första steg i en bred rättslig och rättssociologisk riktning genom att teckna några av de rättsliga och normativa utmaningar som AI medför. Även i en svensk politisk kontext har det på senare tid väckts frågor om regleringsutmaningar för både datadrivna marknader men också specifikt de algoritmstyrda utvecklingsområdena kring ML och AI. I maj 2018 publicerade exempelvis den svenska regeringen en *Nationell inriktning för artificiell intelligens*, som bl.a. innehåller en skrivning om behovet av att ”Sverige behöver utveckla regler, standarder, normer och etiska principer i syfte att vägleda etisk och hållbar AI och användning av AI.” (Regeringskansliet,

2018, s. 10). Teoretiskt sett kan terminologin väcka en del frågor om hur man ska sortera mellan begreppen och vad det är som egentligen avses, men bör tolkas välvilligt som ett behov av vissa värdemässiga gränsdragningar behöver göras i ett samhällsperspektiv när en så kraftfull och potentiellt självständig, icke-transparent och svårförklarad teknologi utvecklas och appliceras inom samhällets centrala funktioner och marknader.

AI som rättslig utmaning: FAT

När det gäller data- och algoritmdrivna system och eventuella samhällsimplikationer av applicerad AI finns det en växande kunskapsmassa i forskningslitteraturen om betydelsen av legitimitet, ansvarsfördelning och transparens. Med engelskspråkig terminologi finns det ett relativt nytt fält formulerat kring *Fairness*, *Accountability* och *Transparency* som förkortas FAT.[3] 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, innehåller de också inneboende risker, såsom kodifiering och förstärkande av samhällelig bias, reducerad ansvarsskyldighet, och ökad informationsasymmetri mellan dataproducenter (kunder) och datainnehavare.

Samtidigt uttrycker denna relativt nya dräkt (FAT) frågor som har en lång historik inom samhällsvetenskaperna och humaniora. Ansvarsfördelning är en central rättslig funktion, transparens eller insyn är centralt för möjligheten till ansvarsfördelning och är en av tillsynens grunder, och den legitimitetsterminologi som ”fairness” uttrycker kan med fördel låna från den bredare och empiriskt utvecklade rättsvetenskap som rättssociologin innebär (jfr Larsson, 2018c).

”Fairness” och legitimitet

Det finns en rad uppmärksammade exempel där samhällelig bias utan intention har reproducerats eller automatiserat förstärkts av AI-system, vilket ofta krävt rigorösa studier för att ens uppmärksammas. Fyra exempel:

1. Datavetenskapliga forskare vid University of Virginia upptäckte att populära bilddatabaser innehöll genderbias, där kvinnor kopplades till köket

[3] Se exv. https://www.fatml.org.

och män till jakt, vilket ledde till en maskininlärande applikation som inte bara reproducerade utan även förstärkte samma bias (se Zhao et al., 2017 för studie).[4]

2. I ett kritiserat amerikanskt exempel på algoritmstyrd myndighetsutövning med straffbestämning baserad på en prediktion av den åtalades recidivism, dvs. sannolikhet för återfall, visade den undersökande journalistiksajten ProPublica att det s.k. COMPAS-systemet var mer sannolikt att *felaktigt* förutsäga att svarta åtalade hade hög risk och samtidigt göra motsatt typ av misstag för vita (jfr Caplan et al., 2018).[5]

3. En forskargrupp byggde ett mjukvaruverktyg för att studera digital spårbarhet i ett försök att öka transparensen i automatiserad annonsfördelning, och fann att annonsfördelningen innehöll en könsbias som oftare fördelade högavlönade jobberbjudanden till män än kvinnor (Datta et al., 2015).

4. En vetenskaplig utvärdering av tre kommersiella könsbestämmande bildigenkänningssystem visade att kvinnor med mörkare hy är den grupp med störst grad av felklassificering (Buolamwini & Gebru, 2018). Det betyder bland annat att dessa tjänster, och applikationer som bygger på dessa, fungerar sämre för vissa gruppers utseenden. Och felmarginalen är påtagligt mycket mindre för ljushyade män.

”Bias”, som term, används även inom statistik och datavetenskap med ett flertal olika definitioner, vilket gör att det finns en begreppslig utmaning och risk för sammanblandning mellan en samhällsvetenskaplig och en teknikvetenskaplig förståelse av bias (bl.a. uppmärksammat av Narayanan, 2018). Här använder jag begreppet ”samhällelig bias” i en rättssociologiskt präglad förståelse av sociala normer och kulturella värderingar.

Mycket av den värdebaserade diskussionen inom maskininlärning och AI sker för närvarande i termer av ”etik”, som i rapporten *Ethically aligned design* (2018) från den globala ingenjörssammanslutningen IEEE. Diskussionerna kring etik och artificiell intelligens får i dessa sammanhang beteckna den breda insikten om att vi som samhällen behöver en reflektion kring värderingar och vilka normer som ska gälla när vi utvecklar AI, men också – vilket är en gryende insikt i den samhällsvetenskapligt tillvända litteraturen – vad det är som AI gör med oss, med samhället, vad autonoma algoritmdrivna system reproducerar

[4] Rapporterades i tidningen Wired 21 augusti 2017: https://www.wired.com/story/machines-taught-by-photos-learn-a-sexist-view-of-women/amp.

[5] Studien genomfördes och rapporterades av den medborgarrättsligt drivna undersökande journalister vid ProPublica (23 maj 2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

och förstärker i termer av värderingar, kultur, makt och normativitet. ”Etik” utgör därmed i min bedömning här ofta ett proxybegrepp, dvs. ett begrepp som mäktar med att samla de diversifierade grupper som å ena sidan utvecklar metoderna och teknologierna – matematikerna och datavetarna – de som kommersialiserar och implementerar på marknaderna, och de samhällsvetenskapligt och humanistiskt inriktade som studerar och bättre vill förstå metodernas och teknologiernas plats och påverkan i samhället. Etikdiskussionen kommer rimligen i AI-sammanhang med tiden att ersättas med mer tydligt definierade begrepp inom regleringsområden, branschstandarder, certifieringar och mer utförligt analyseras inom traditionella vetenskapliga discipliners huvudområden gällande kulturer, makt, marknadsteori, normer etc. Inom rättssociologin har man länge studerat legitimitet i termer av sociala normer, i linje med Durkheims ”sociala fakta”, eller Ehrlichs ”levande rätt”, dvs. som något empiriskt mätbart, strukturellt utbrett, men ändå inte nödvändigtvis formaliserat i termer av lag ”i böckerna” (Pound, 1910). Det har bland annat utvecklats av Håkan Hydén (2002) och Måns Svensson (2008; Hydén och Svensson, 2008). Sistnämnde, Måns Svensson, har utvecklat en modell för att empiriskt kunna studera styrkan hos sociala normer, vilket också gör att olika fälts normstyrka kan jämföras med varandra – som med rattonykterhet, bilbältesanvändning och hastighetsförseelser (Svensson, 2008) – eller relateras till vad lagen postulerar – som i en studie av unga svenskars fildelning kontra upphovsrätt och den s.k. Ipredlagen (Svensson & Larsson, 2012).

Att datoriserade system kan innehålla bias eller samhälleligt problematiskt skeva värderingar är inte nödvändigtvis en insikt som är ny (jfr Friedman & Nissenbaum, 1996), men skalan på dagens användning och det samhälleliga beroendet är större än någonsin, med konsekvenser och risker för samhälleligt centrala funktioner som kreditgivning, förmedling av arbete, sjukvård och kunskaps- och nyhetsförmedling. Sociala, politiska, ekonomiska och kulturella dimensioner kopplade till exempelvis sökmotorer har också studerats av många (jfr Hargittai, 2007), även gällande kulturella implikationer av policies kring vad som betraktas som obscent eller tabubelagt språkbruk och hur den s.k. ”auto complete”-funktionen, när sökmotorn fyller i relaterat till vanliga sökningar, kan ge kontroversiella effekter (Troumbley, 2015).

Att även sökmotorer, som i mycket är automatiserade och innehåller självlärande och därmed artificiellt intelligenta element, samspelar, reproducerar och delvis är en produkt av samhällets sociala, historiska och kulturella strukturer demonstrerades nyligen med emfas av den amerikanska

kommunikationsvetaren Safiya Noble (2018) i boken *Algorithms of Oppression: How search engines reinforce racism*. Algoritmerna kan därmed automatiserat begränsa individers möjligheter på ett sätt som kan vara både olagligt eller betraktas som oetiskt, argumenterar Noble, som en ”teknologisk redlining” där dataanalysen diskriminerar vissa grupper på ett dolt sätt som sker strukturellt och oftast bara blir upptäckbart genom omfattande granskning i efterhand. Den amerikanska sociologen John McKnight (jfr Norton, 2013) myntade under 1960-talet termen ”redlining” för att beskriva en diskriminerande praktik att markera områden (med en röd penna på en karta) där banker skulle undvika investeringar baserade på samhällsdemografi, och uttrycket har använts för att beskriva en systematiskt försämrad tillgång till funktioner som finansiella tjänster, försäkring eller sjukvård för vissa områden i en stad. Noble använder konceptet som ett sätt att rikta ansvarsfrågan mot de digitala intermediärer som samspelar med – och därmed bidrar till – en redan befintlig diskriminering. Hon förser Caplan et al. (2018) med en förklaring där

> ”...teknologisk *redlining* är en form av digital datadiskriminering, som använder våra digitala identiteter och aktiviteter för att stärka ojämlikhet och förtryck. Det sker ofta utan vår kunskap, genom våra digitala engagemang, som blir en del av algoritmiska, automatiserade och artificiellt intelligenta sorteringsmekanismer som antingen kan riktas mot oss eller utesluta oss. Det är en grundläggande dimension för att generera, upprätthålla eller fördjupa rasistisk, etnisk och könsdiskriminering, och den är centralt knuten till fördelningen av varor och tjänster i samhället, som utbildning, boende, och andra mänskliga och medborgerliga rättigheter”. (Noble i Caplan et al., 2018, s. 3.)

Noble kopplar därmed teknologisk redlining till en längre historik nu överförd i en teknologiskt datafierad kontext. En utmaning ligger i bristen på insyn och transparens för att metoderna är ”alltmer svårupptäckbara på grund av deras digitala implementeringar genom internetbaserad programvara och *plattformar*, inklusive uteslutning från, och kontroll över, individuellt deltagande och representation i digitala system” (Noble i Caplan et al., 2018, s. 4).[6] Teknologisk redlining kan därmed ge ett utfall när de profilerade inte har någon kontroll över hur data används för att profilera dem. Om samhällelig bias finns i datan replikeras det i utfallet. Utan verkställbara mekanismer för transparens,

[6] För en analys av digitala plattformar som styrande mekanismer med delvis automatiserade och AI-beroende rättsliga kvaliteter, se Larsson (2018b), och utvecklingen av s.k. plattformsekonomier, se Larsson & Andersson Schwarz (2018).

revision av dataanvändningen eller fungerande ansvarsfördelning finns det, enligt Caplan et al., (2018), knappast någon kännedom om hur algoritmiskt beslutsfattande begränsar eller hindrar medborgerliga rättigheter. Detta är ett argument för behovet av mer transparens i användandet av datadrivna autonoma tjänster och plattformar.

Det finns också en kritik mot biasreproducerande system som pekar mot att en alltför homogen designcommunity leder till blinda fläckar. Det argumenteras exempelvis för i en rapport från AI-forskningsinstitutet AI Now (Campolo et al., 2017) som beskriver ”legacies of bias” och konstaterar:

> ”AI is not impartial or neutral. Technologies are as much products of the context in which they are created as they are potential agents for change. Machine predictions and performance are constrained by human decisions and values, and those who design, develop, and maintain AI systems will shape such systems within their own understanding of the world. Many of the biases embedded in AI systems are products of a complex history with respect to diversity and equality.” (AI Now report 2017, s. 18.)

I linje med ovan kan man därmed konstatera att värderingar och normativitet ryms på båda sidor av designprocessen, dvs. både i dess bruk av den data som kommer från människor i samhällen som innehåller bias och strukturella skevheter och på den kontext som tar fram applikationer och designar tjänsterna. Sidorna hänger ihop, eftersom hanteringen av samhällelig bias troligen inte är möjlig om inte designsidan har en utvecklad förståelse för designens normativa roll, och aktivt arbetar för att hantera den. Detta väcker både svåra men behövliga frågor om vem som borde ha ansvar för vad inom samhällsapplicerade autonoma system.

Agens och ansvarsfördelning

Det finns olika parallella utvecklingslinjer för ansvarsfördelningsfrågor gällande artificiell intelligens. Den traditionella AI-forskningen har, som nämnt, tidigare hänvisat till Asimovs robotiklagar (jfr Leigh Andersson, 2008), och industriella sammanslutningar och forskningsassociationer har utvecklat en rad principiella förhållningssätt inom robotik och maskininlärning. Även enskilda företag gör ibland principiella ställningstaganden för deras AI-utveckling. IEEE-rapporten nämnd ovan fokuserar ansvarsfrågor utifrån ett design- och designerperspektiv och tar även upp autonoma vapen som ett

särskilt problematiskt fält. Google publicerade i början av juni 2018 ett knippe principer för sitt förhållningssätt till artificiell intelligens[7], bara någon vecka efter att det blev känt att bolaget beslutat att inte förnya kontraktet för samverkan med den amerikanska militären kring att använda maskininlärning för att analysera videomaterial från drönare, det s.k. Mavenprojektet.[8] En ökad medvetenhet för negativ och illvillig användning av AI uttrycks också som ett ansvar för design- och utvecklingssidan av en stor samling forskare på området (Brundage et al., 2018). Hotbilden handlar här bl.a. om utvecklade varianter av cyberattacker som automatiserad hacking, 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, som inför Brexitomröstningen (Bastos & Mercea, 2017), eller för att skapa splittring i frågor, som vaccinationsdebatten i USA (exv. Broniatowski et al. 2018).[9] Forskargruppen kring illvilliga (”malicious”) användningsområden för AI efterlyser i ett säkerhetsperspektiv en starkare kultur av ansvar hos AI-utvecklare för vad deras verktyg kan användas till och pekar ut vikten av utbildning, etiska standards och normer (Brundage et al., 2018, s. 7).

Inom en kritisk diskurs för algoritmers betydelse menar man att riskerna med att bias återkommande förs in och automatiseras i processer är en helt central utmaning – om än inte som resultat av en medvetet illvillig användning. Det kan, som nämnt, ske på grund av att träningsdatan är skev, otidsenlig eller på annat sätt en dålig representant för vad man vill uppnå (cf. Bozdag, 2013). Ett algoritmiskt ansvarserkännande (”algorithmic accountability”), menar Caplan et al. (2018) – i en rapport från forskningsinstitutet Data & Society – avser processen att fördela ansvar för skada när algoritmstyrt beslutsfattande resulterar i diskriminerande eller orättvisa konsekvenser (jfr Diakopoulos, 2015; Zarsky, 2016). Ett sådant erkännande kan därmed också avse ansvaret för hur en algoritm skapas och dess inverkan och konsekvenser på samhället. Om skada uppstår innefattar ansvarsfulla system en mekanism för åtgärd.

> ”While law has always lagged behind technology, in this instance technology has become de facto law affecting the lives of millions – a context that demands lawmakers create policies for algorithmic

[7] Pichai, S. (7 juni 2018) ”AI at Google: our principles”, Google blog. https://www.blog.google/topics/ai/ai-principles/.

[8] https://www.theverge.com/2018/6/1/17418406/google-maven-drone-imagery-ai-contract-expire.

[9] För mer om plattformars samhälleliga betydelse, se Andersson Schwarz & Larssons antologi om *Plattformssamhället* (2019).

> accountability to ensure these powerful tools serve the public good." (Caplan et al., 2018, s. 12.)

Detta ekar av juridikprofessor och internetforskaren Lawrence Lessigs argumentation kring "code is law" från drygt ett decennium tidigare (Lessig, 2006) där den digitala arkitekturen i sig måste vara med i analysen av normer och beteende (jfr Larsson, 2013). Exempelvis, dagens digitala intermediärer behöver ideligen utveckla policy och implementering – i hög grad automatiserad och ofta beroende av maskininlärning – för innehållsmoderering, trendingfunktioner och relevansbedömningar, vilket har en fundamental betydelse för både individer och företag som använder eller är beroende av deras tjänster och produkter (jfr Andersson Schwarz, 2017; Andersson Schwarz & Larsson, 2019; Gillespie, 2018).

En annan inneboende utmaning här är prediktionen, dvs. att en maskinlärande applikation kan användas för att göra sannolikhetsbedömningar om framtida händelser, som med andra ord inte ännu hänt. Allvaret beror på vad detta sannolikhetsbedömmande får ligga till grund för. När sannolikhetsbedömningen exempelvis används för kreditbedömningar, medicinska diagnoser, polisiär resursfördelning eller strafftidsrekommendationer är det rimligen oerhört viktigt att prediktionen görs på så goda och lärande grunder som möjligt.

Två exempel kan här tas upp för att förtydliga hur AI och ML tar del i komplexa samhälleliga fält som ytterligare visar på behovet av att förstå AI som en samhällsutmaning: digitala plattformar och autonoma fordon.

Plattformarna

En vidare diskurs med AI-relevans för utmaningar med ansvarsfördelning rör digitala plattformars betydelse, vilket delvis är en debatt om hur intermediärer ska bedömas när det gäller ansvar för det material eller beteende som sprids eller genereras i relation till plattformarna (Andersson Schwarz, 2017; Larsson, 2018b). Intermediärsansvarsfrågor är inget nytt,[10] men aktuella exempel kan man finna hos de storskaliga digitala plattformarna, som i debatterna kring vilket ansvar Facebook och YouTube (Google) bör ha för den information som delas inom deras respektive plattformar eller gällande Googles relevansbedömning i den indexerande sökmotorn (jfr Gillespie,

[10] När individerna bakom fildelningssajten The Pirate Bay åtalades 2009 för medhjälp till upphovsrättsbrott uppkom en liknande konceptuell utmaning kring hur domstolen skulle se på ansvar i förhållande till denna "plattform" (se Larsson, 2017a).

2018). Eftersom plattformarna är storskaliga – Facebook har över två miljarder aktiva användare och Google har enligt uppgift inte mindre än sju tjänster som har mer än en miljard användare – behöver de också i hög grad automatisera hanteringen av informationshanteringen, där båda aktörerna är stora investerare i och utvecklare av artificiell intelligens för en rad funktionaliteter, bland annat för ansiktsigenkänning, språkanalys och röstigenkänning (Dolata, 2017). En variant av intermediärsansvarsfrågan rör kontroll över användarinformation, vilket aktualiserades i den s.k. Cambridge Analytica-skandalen där information från mellan 50 till 87 miljoner Facebookanvändare visats ha använts för att påverka demokratiska val i en rad länder.[11] När Facebooks VD, Mark Zuckerberg, frågades ut av den amerikanska kongressen med anledning av skandalen fick han frågor om plattformens ansvar för det material som sprids. Zuckerberg förde återkommande fram AI som ett verktyg för att bekämpa oönskat innehåll som hatiska uttalanden, falska nyheter, hämndporr och annat. Hans uttalanden har kritiserats för att vara uttryck för en förenklad ”*AI solutionism*” – som klingar av den kritiske digitaliseringsforskaren Evgenij Morozovs kritik mot ”*teknologisk solutionism*” (2013) – och att de automatiserade optimeringsverktygen som den storskaliga plattformen bygger på i sig själva bidragit till hur spridningen av falska nyheter och kontroversiellt innehåll.[12] En ansvarserkännande plattformsdesign ställs inför en rad normativa ställningstaganden kring vilken typ av bilder, texter och länkar som ska räknas som stötande, olagliga eller falska. Inte sällan dras gränserna på olika sätt i olika kulturer och jurisdiktioner. Även en del kunskapsområden, exv. om historiska händelser eller geografiska platsers tillhörighet, kan vara kontroversiella och bestridas av endera grupper, vilket gör den normativa uppgiften lika svår som behövlig.

Medieforskaren Tarleton Gillespie har analyserat hur sociala medier administreras och diskuterar ansvarsfrågorna i boken *Custodians of the Internet* (2018). Ordvalet – ”custodians” – är rimligen ingen slump, eftersom han tidigare konstaterat att ”plattformar” är ett undslippande begrepp som döljer hur aktiva besluten är som ligger bakom sociala mediers moderering (Gillespie, 2017). Han föreslår förbättringar för innehållsmodererandet hos de stora plattformarna, bl.a. genom en radikal designtransparens.

[11] En nyhet som fick mycket spridning när journalisten Carole Cadwalladr publicerade en artikel med en visselblåsare i The Guardian, 18 mars 2018. https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump.

[12] BuzzFeed News (11 april 2018) ”Why Facebook Will Never Fully Solve Its Problems With AI”, av Davey Alba. https://www.buzzfeednews.com/ article/daveyalba/mark-zuckerberg-artificial-intelligence-facebook-content-pro.

Självkörande bilar

En rad traditionella biltillverkare världen över utvecklar autonoma bilar, utmanade av teknikbolag som Googles avknoppade Waymo, transportförmedlande Uber och elbiltillverkaren Tesla. Förarlösa bussar har under en tid testats experimentellt av Nobina i Kista, och sedan den 24 januari 2018 är en linje i drift. Liknande projekt kring självkörande fordon i kollektivtrafik finns i bl.a. China, Holland, Schweiz och Las Vegas och det är bara en tidfråga innan autonoma fordon blir en vanligare syn i en vardaglig trafikmiljö i många städer världen över. Autonomin, som i datadrivna applikationer i mycket är avhängiga de algoritmer som designats för att utföra de funktioner som behövs, är ett fält som är helt centralt för förarlösa fordon, men också väcker frågor i relation till ansvarsfrågor. Lagstiftning för utvecklingen inom självkörande fordon i trafik är under utveckling i Sverige (SOU 2018:16)[13], och där är ansvarsfrågorna helt centrala vid olyckor, vilket också har diskuterats i forskningslitteraturen en tid (jfr Hevelke & Nida-Rümelin, 2015). Frågorna har aktualiserats inte minst i dödsolyckor med autonoma fordon inblandade. En Tesla modell S, som använder både radar och kameror för att tolka omgivningen, förväxlade i ett fall från 2016 en lastbil med himlen, vilket resulterade i en dödsolycka. I mars 2018 Körde en Volvo-SUV, som Uber använder i sin utveckling av självkörande fordon, på och dödade en kvinna i Arizona i USA, vilket väckte omfattande debatt kring ansvarsfrågor och självkörande fordon i trafiken. Även om en jämförelse med förarstyrd trafik skulle ge att den autonoma är avsevärt säkrare så kommer ändå olyckorna påverka den tillit som behövs för att människor ska acceptera höggradigt autonoma fordon jämte sig i trafiken. KPMG (2017) har bedömt den impact som självkörande fordon kan komma att få på kollektivtrafiken. En av flera utmaningar som lyfts är ”the burden of liability” (s. 12), uttryckt som ett behov av rättsutveckling, men även ”privacy”, dvs. integritets- och dataskyddsfrågor, som ett resultat av den datainsamling och -hantering som förarlösheten kräver. Sistnämnda ställer olika rättigheter mot varandra på ett sätt som behöver jämkas.

[13] Den 1 juli 2017 fattade regeringen beslut om nya regler för självkörande fordon som gjorde det lättare att få genomföra försök med självkörande fordon på allmänna vägar (förordning 2017:309, se SOU 2016:28). I förordningen regleras även att det vid färd med ett självkörande fordon ska finnas en fysisk förare i eller utanför fordonet. Den 7 mars i år lämnades slutbetänkandet i utredningen om självkörande fordon på väg över till regeringen (SOU 2018:16; se Dir 2015:114), där ansvarsfördelningsfrågor och dataskydd utgör en väsentlig del.

Agens

Det kommer oundvikligen att uppstå frågor kring tingens eller mjukvaruprocessernas *agens* när de är utrustade med förmågan att överblicka stora mängder information och lära från den, inte minst kopplat till automatiserat beslutsfattande (Larsson, 2018c). Ansvarsfrågor relaterar direkt till agens, där centrala delar av rättslig ansvarsfördelning ligger i att bedöma intentioner och förväntningar på aktivitet i relation till insikt om risker. Detta konstateras bl.a. av Mareille Hildebrandt, som är forskare i det kombinerade fälten juridik och teknologi, och i boken *Smart Technologies and the End(s) of Law* resonerar kring agensens betydelse för ”smarta” ting (2015). Inom straffrätten finns komplexa rättsvetenskapliga resonemang kring gränsdragningsfrågor mellan uppsåt/avsikt (dolus) och vårdslöshet (culpa). I straffrättens finns olika nivåer eller former av uppsåt, som direkt uppsåt, indirekt uppsåt, likgiltighetsuppsåt, m.fl. (jfr Jareborg, 1993).

Kan en maskin eller mjukvara nå ”insikter” och ha ”intentioner”? Steget är inte nödvändigtvis väldigt långt borta, och oavsett vilket så kommer den typen av argumentation att bli juridisk när företag eller myndigheter utvecklar allt mer autonoma AI-tjänster som oundvikligen kommer att hamna i rättsliga prövningar. Det kan handla om en bred palett av allt från diskriminerande utfall enligt de reglerade diskrimineringsgrunderna till bilolyckor med självkörande fordon till oväntade kostnader för termostaters utfall till uteblivet försörjningsstöd i socialt arbete. Ansvarsfrågorna är också tätt sammanlänkade med transparensfrågor. En rad länder har också relativt nyligen lanserat AI-strategier, där etik och reglering är en fråga som ofta är central, inte minst i Sveriges (se ovan) men även i EU-kommissionens kommuniké från april 2018 (SWD(2018) 137 final).

Behovet av insyn, tillsyn och transparens

Ett omvittnat problem för ansvarsfrågor relaterade till algoritmdrivna processer är deras brist på transparens, vilket ibland kallas för den svarta lådan, eller ”black-boxing” (jfr Pasquale, 2015; Passmann & Boersma, 2017). En stor del av ansvarsfördelningsproblematiken handlar om hur man förstår vad som har hänt, vilket talar för vikten av att förstå relationen till transparens bättre gällande samhälls- och marknadsapplicerad artificiell intelligens, även om transparens inte kan ses som en lösning för alla problem (Ananny & Crawford, 2018). Det är dessvärre ofta oklart vad man avser med transparens, och brist

på transparens kan komma av flera orsaker. Nedan diskuteras sju transparensrelevanta hänsyn eller ”nyanser”, där 1 och 2 representerar motverkande intressen, och 3-7 kan sägas utgöra varianter av kunskapsutmaningar.[14]

1. Ägande

Ett proprietärt upplägg med företags mjukvara och data är ofta ett helt legitimt sätt att med en kommersiell logik bedriva konkurrensutsatt innovation. Det kan exempelvis även bli fallet när en produkt kommersialiseras och skalas upp på en marknad, och kan utgöra en förutsättning för investerare. För vissa företag är mjukvaran och dess algoritmer värdefulla ”recept” som de betraktar som företagshemligheter. Vissa företag ser sin börsvärdering som en direkt konsekvens av den data de har om sina användare (Spiekermann & Korunovska, 2016). Det förs dock ibland fram som ett granskningsproblem i relation till hur individers data samlas in och används på sätt som individen inte så enkelt har insyn i (jfr Pasquale, 2015). Ett element av utmaningen med exempel 2 ovan, gällande återfallsriskbedömning, s.k. recidivism, ligger i dess brist på transparens och därmed brist på lärande feedback. I exemplet bidrar kommersialiseringen av metoderna till bristen på transparens, eftersom granskning av mjukvara och feedbackloopar för dataanvändningen försvåras. Oaktat proprietära upplägg kan förvisso rimligen uppskalning vara en utmaning i sig, genom att användare av en AI-applikation – exv. i vården – inte kan ha samma kunskap och insyn i dess funktionalitet som dess utvecklare. Men uppskalningen sitter i praktiken ofta ihop med ett ägande och hemlighållande av de värdefulla ”recept” som ligger till grund för innovationen.

2. Undvikande av missbruk

Det finns algoritmberoende och automatiserade processer vars syfte skulle kunna motverkas av att berörda parter kände till deras exakta funktionalitet. Transparens skulle motverka processens syfte, som därmed missbrukas eller manipuleras – spelas, eller *gejmas*. Den svenska Försäkringskassan hanterar till exempel en stor mängd information och tar en stor mängd beslut på socialförsäkringens område kring vem som har rätt till ersättning eller ej. I syfte att mer effektivt upptäcka fusk men också minska mängden felbeslut använder de riskbaserade kontroller och s.k. profilering som metod. Inspektionen för socialförsäkringen (ISF) fick under 2017 i uppdrag att kartlägga och analysera

[14] För en plattformsspecifik analys, se Larsson (2018b).

Försäkringskassans arbete med profilering som urvalsmetod för riktade kontroller. De konstaterade att:

> ”Försäkringskassan bedömer att det är viktigt att allmänheten inte får kännedom om exakt vilka variabler som ingår i urvalsprofilerna. Om någon skulle begära att få information om vilka variabler som ingår i en viss urvalsprofil skulle Försäkringskassan sannolikt sekretessbelägga variablerna med stöd av 17 kap. OSL.” (ISF, 2018, s. 86)

Ett av skälen är rimligen att det finns ett inneboende incitament för klienterna att ”spela systemet” här, vilket talar för att total öppenhet skulle motverka den selektion som Försäkringskassan är satt att göra. Samma sak gäller rimligen liknande typer av ”fraud detection” som både försäkringsbolag och banker utför. Caplan et al. (2018) påpekar exempelvis att även den minsta öppenheten kring hur trendingfunktionen på Twitter fungerar har gjort det möjligt att manipulera delar av miljön för att täcka vissa ämnen med automatiserade bottar eller bot-nätverk för att otillbörligt påverka, styra eller rentav förstöra debatter.

3. Litteracitet

Det krävs en viss specifik expertkompetens för att ens kunna bedöma algoritmer och deras dataanvändning, som inte kan sägas finnas hos folk i gemen eller troligen inte ens hos alla de tillsynsmyndigheter som i ökande grad har att idka tillsyn över datadrivna och automatiserade marknader och arbetsmarknads- och myndighetspraktiker. Man kan här prata om *datalitteracitet* eller *algoritmlitteracitet* som ett behov (jfr Haider & Sundin, 2019).

4. Koncept, terminologier och förklarbarhet

Hur den komplexa AI-processen förklaras genom val av språk, metaforer och symboler, har direkta implikationer för hur ansvarsfrågor förstås. Ofta uttrycks brist på transparens som ett tillitsproblem, dvs. som i EU-kommissionens kommuniké om artificiell intelligens:

> ”...to further strengthen trust, people also need to understand how the technology works, hence the importance of research into the explainability of AI systems.” (SWD(2018) 137 final, s. 14.)

Förklaringar kan dock ske på olika nivåer och med olika typ av symbolik och med olika typer av sociala behov (se Doshi-Velez, et al. 2017), viket gör att det

inte är helt enkelt att sortera i behovet av förklaringar. Det är ibland oklart vad som avses när man menar att teknologin behöver förstås av dess användare, eller vad en lämplig förklaring skulle medföra. Vi är rimligen redan idag alla användare av en rad teknologier som vi bara har en ytterst vag förståelse för hur de egentligen fungerar. En förklaring av exempelvis AI-genererat beslutsfattande kommer därmed oundvikligen att behöva välja nivå av konkretion genom val av symbol eller metaforer. Jag har tidigare visat att beroende på genom vilka metaforer man förstår komplexa digitala fenomen så kommer normativa och rättsliga ställningstaganden att avgöras på vilka som väljs eller får styra. Delvis är detta historisk betingat, dvs att det finns ett konceptuellt stigberoende som påverkar vår förståelse på så vis att vi förstår nya fenomen genom redan etablerade begrepp (Larsson, 2017a). Hur förklaringen kring AI-genererade processer ser ut i termer av metaforer och symbolik – varför en bil eller robot agerade på ett visst sätt eller hur ett algoritmgenererat beslut nåddes – kommer därmed rimligen att vara av väldigt viktigt för hur de kommer att uppfattas och accepteras.

5. Komplexa dataekosystem

En brist på transparens kan relatera till hur informationen ”reser” i ”ekosystem” mellan många aktörer och datamäklare (jfr Larsson, 2017b), vilket bland annat studier tecknat i sin komplexitet och mångfald (exv. Christl, 2017). Pasquale (2017) konstaterar för s.k. datamäklare att det är orealistiskt att förvänta sig att individer ska kunna rikta sina dataskyddsrättigheter mot var och en datamäklare (jfr Larsson, 2018a).

6. Distribuerat, individualiserat utfall,

De individuellt relevanta tjänsterna, som Googles sökmotor, den riktade reklamen som följer efter besökare på en myriad sajter, det individuella flödet på Facebook, och i andra konsumentprofilerande tjänster som vill ”personalisera” sin relation eller sin prissättning har ett i hög grad distribuerat, och – draget till sin spets – individualiserat utfall. Många tjänsteutvecklare för fram betydelsen av det individuellt relevanta kunderbjudandet, att det finns en för kunden attraktiv belöning, en s.k. ”tradeoff”. Vilket stämmer för flera tjänster, men frågan är om konsumenter alltid förstår hur transaktionen ser ut (jfr Larsson 2018a). Den amerikanska konsumentforskaren Joseph Turow har med kollegor i en studie konstaterat att det finns en ”tradeoff fallacy” och en uppgivenhet för datainsamlingen hos kundkortskunder inom amerikansk handel, vilket kan mana till eftertanke och är något som behöver studeras

vidare. Utmaningen med ett distribuerat och individualiserat utfall från ett transparensperspektiv ligger främst i svårigheten att upptäcka otillbörliga mönster i något som bara visar sig individuellt, i vissa fall högst privat (jfr Nobles definition av teknologisk redlining ovan).

Myndigheternas tillsynsarbete kan ses som ett slags transparensarbete, där syftet är att granska vad marknadsaktörerna gör, för att bedöma om det finns något otillbörligt i deras praktik. I en artikel om konsumentskydd på datadrivna och automatiserade marknader, till exempel gällande marknadsföring online i sociala nätverk, utvecklar jag behovet av *algoritmisk governance*, vilket i detta sammanhang innebär att jag argumenterar för att tillsynsmyndigheter behöver metodutveckla för att ens kunna upptäcka eventuella oegentligheter eller skevheter för individualiserade tjänster (Larsson, 2018; jfr Larsson & Ledendal, 2017). Exempelvis Konsumentverket har ansvar för tillsyn av marknadsföring så att den inte är vilseledande eller på annat sätt otillbörlig. Men hur skulle man upptäcka om marknadsföring eller prissättning – antingen medvetet eller som en konsekvens av andra socioekonomiska förhållanden i samhället – leder till en samhällelig bias som träffar diskrimineringsgrunder för kön, etnisk tillhörighet, funktionsnedsättning, sexuell läggning eller ålder (jfr Larsson, 2016, s. 46 f.; 2017b)? Om den riktade reklamen eller prisdiskrimineringen är individualiserad, och träffar var och en av oss efter en algoritmiskt automatiserad bedömning av våra preferenser, stordatadrivna analyser av antagna (”inferred”, i den engelskspråkiga litteraturen) samband eller betalningsviljor, så är det tänkbart att de attribut som medföljer diskrimineringsgrunderna analytiskt kopplas samman till en struktur som reproduceras via den individualiserade reklamen eller prissättningsalgoritmen. Frågan är hur en tillsynsmyndighet ska upptäcka det. Det är inte helt enkelt eftersom det i sin extrem måste bedömas av genom att aggregera resultatet av en stor mängd individuella utfall, möjligen genom 1.) insyn i samverkan med de aktörer som genomför matchningen, 2.) analys av en mängd användares flöden, eller genom 3.) att bygga en testmiljö med konstruerade användare.

7. Algoritm-komplexitet: behovet av XAI

Det finns för komplexa AI-verktyg ett inneboende problem med att bedöma hur ett visst utfall ha nåtts. Det finns inom AI-forskningen ett uttalat fält som handlar om förklarbarhet eller tolkningsbarhet (XAI), som uppstått som en respons på problem med maskininlärning som även för forskarna innebär en ”svart låda” där ett problem blir löst men utan att det går att tolka exakt hur

(jfr Guidotti et al. 2018). Utfallet kan vara en högre sannolikhet för ett visst utfall, exv. omsatt i högre lönsamhet eller mer precision i en diagnos, när det appliceras, men inte nödvändigtvis på vilka grunder eller hur resultatet nåddes i detalj. S.k. djuplärande och andra varianter av neurala nätverk kan innebära ”svarta lådor” i den här betydelsen. Förklarbarheten inom AI-utvecklingen kan också länkas till punkt 4 ovan, dvs. att det finns mer samhällsvetenskapligt tillvända teorier inom kognitionsvetenskap och socialpsykologi om hur människor förstår och förklarar, som möjligen kan vara användbara inom AI-fältet (Miller, 2019). Man kan också tänka sig att tidsdimensionen kan utgöra ett transparens- och förklarbarhetsproblem, som vid algoritmbaserad högfrekvenshandel med aktier eller andra värdepapper.

Automatiserat beslutsfattande i offentlig verksamhet

Ett specialfall inom samhällsapplicerad AI är användning inom offentlig verksamhet, vilket är ett fält som kommer ha ökad betydelse framöver. Det finns exempel på automatiserad myndighetsutövning i den kommunala socialtjänsten – som prisats för sin effektivitet men ändå väcker transparensfrågan – där Trelleborg använder sig av vad de kallar en ”robot” i beslutsfattandet av brukares försörjningsstöd, något vi kallar för den ”tredje vågen” av digitalisering i en rapport om digitalisering av socialt arbete (Svensson och Larsson, 2018; se Svensson och Larsson, 2017). Sveriges Kommuner och Landsting har bedömt att automatiserat beslutsfattande är möjligt i offentlig förvaltning, enligt Förvaltningslagen (2017:900) i relation till EU:s Dataskyddsförordning (GDPR), men inte nödvändigtvis sådant beslutsfattande som omfattas av Kommunallagen (2017:725), se SKL (2018). De ser dock inga hinder mot att använda automatiserade beslutsstöd, dvs. där en mänsklig handläggare sedan tar det formella beslutet. Både socialtjänstens och Försäkringskassans hantering är exempel på myndigheters användning av algoritmer och automatiserade processer och för vissa fall beslutsfattande, vilket kan ställa ytterligare krav på transparensen i syfte att säkerställa att den fungerar i enlighet med de krav som ställs på myndigheten. I en rapport från AI Now (Reisman et al., 2018) föreslås en ”algorithmic impact assessment” för användning av automatiserade algoritmiskt styrda processer i offentlig verksamhet. Dvs. den grundläggande tanken är liknande – offentlig verksamhet behöver vara bra på att tidigt förstå konsekvenserna av egen användning av automatiserade processer. Och, givet att mycket av tjänsterna behöver

upphandlas av en extern utvecklare, blir även ägandefrågorna i punkt 1 ovan en del av bedömningen av transparensfrågorna.

Kontextuella transparensbehov

Olika sammanhang kan kräva olika nivåer av insyn, beroende på vad syftet och behovet är. För vissa fall skulle en dedikerad eller kvalitativ transparens vara att föredra, vilket bland annat diskuteras av Pasquale (2015, ss. 160-165), likt en bilbesiktning där inte var och en konsument egentligen har möjlighet eller ens behov av att granska innehållet och funktionerna, men det samhälleliga intresset av att utfallet av en viss applikation inte är otillbörlig och diskriminerande beaktas (jfr Larsson, 2018a). För andra fall skulle en radikal öppenhet kunna vara eftersträvansvärd utifrån ett samhällsperspektiv (jfr Gillespies argumentation för digitala plattformar, 2018). Vad man avser med transparens och begrepp som ”förklara” och ”att förstå” kommer också att behöva ytterligare förtydligas i det rättsliga arbetet med samhällsapplicerad AI.

Diskussion: Rättvisa och normativitet

Rättvisa är ett centralt begrepp inom den breda rättsvetenskapliga litteraturen som kommer att få återbesökas mycket framöver i relation till artificiell intelligens (jfr Larsson, 2018c). Också Hildebrandt argumenterar för att en rad fundamentala rättigheter är i fara i ett samhälle med datadriven agens och smarta teknologier (ss. 133 ff.). Rättvisebegreppet kan exv. återbesökas i relation till användningen av artificiell intelligens i förhållande till rättssystemet med operativa sannolikhetsbedömningar inom polisiärt arbete och straffutmätning. Det finns exempelvis en växande amerikansk användning av datadrivna sannolikhetsbedömningar för insatser inom polisiärt arbete, s.k. *predictive policing*, som också har sin kritiska litteratur (jfr Shapiro, 2017). Den kinesiska användningen av publika kameror och ansiktsigenkänning rapporteras vara under kraftig utveckling, inklusive ett s.k. socialt kreditsystem där medborgare är tänkta att bedömas och rankas i ett datadrivet system, som enligt regeringsdokumentet ska uppmuntra till ömsesidigt förtroende och bestraffa ”dåligt” beteende och belöna ”bra” beteende (Svensson, 2018).

Även svensk polis testar i ett pilotprojekt under 2018 att använda utrustningsmonterade kameror. Det har nyligen publicerats kritiserade studier som använt bildanalys och maskininlärning för att genom analys av bilder på ansikten göra sannolikhetsbedömningar om kriminalitet (Wu & Zhang,

2016). Detta kritiseras bland annat av Joi Ito, chef för MIT media lab, för att representera ett slags reduktionism med rötter hela vägen tillbaks till rasbiologins ”fysiognomiska” uttryck under 1930- och 40-talen (se Agüera y Arcas et al., 6 maj 2017, för en kritik) och programmeraren och författaren Cathy O’Neill varnar för att Big Data riskerar att bli den moderna varianten av den pseudovetenskapliga frenologin (2016, ss. 121f.). När användningen kritiseras för sin reduktionism så är en central del av kritiken mot metoden att det är svårt att bedöma vad det är maskininlärningsalgoritmerna plockar upp för att nå resultatet. Det är helt enkelt svårt att avgöra om det är relevant. Liknade kritik har riktats mot en bildanalytisk studie där forskarna hävdat att de kunde klassificera människors sexualitet genom en analys av fotografier på ansikten (Murphy, 9 oktober 2017).

Om man kombinerar sofistikerad och snabb ansiktsigenkänning, reduktionism i bedömningen av utseenden, med ökad användning av publika kameror på byggnader såväl som i polisers utrustning, så finns det också ett stort behov av transparens och att förstå vad denna ”street-level surveillance” leder till, som det beskrevs i en artikel i The Economist med Lipskyklingande terminologi.[15] Dels finns det risk för att samhällelig bias automatiseras genom ansiktsigenkänning och dels finns det en helt central rättviserelaterad utmaning i att operativt använda sig av förutsägelser om framtiden, dvs. något som ännu inte hänt, i en rättslig process. Samtidigt kan naturligtvis polisiärt videomaterial verka som god bevisning för händelser som faktiskt har hänt.

För flera av de tidiga rättsteoretikerna har en central fråga varit att analysera relationen mellan moral och rätten, inte minst i relation till rättvisa. Ett exempel är den polske rättssociologen Leon Petrazycki som var verksam i St. Petersburg och Warszawa under tidigt 1900-tal, och vars verk framförallt återfinns i en översättning till engelska från 1955. Petrazycki skiljer exv. på positiv och intuitiv rätt men också på officiell och icke-officiell rätt, där sistnämnde liknar Eugen Ehrlichs idéer om en ”levande” rätt som reproduceras informellts i samhället och Pounds ”law in action”. Petrazycki öppnade därmed upp för en mer empirisk approach på rätt och normer, vilket påverkat många efterföljande. Det öppnar också upp för insikten om att 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, även informella sådana. Val av normer är inte nödvändigtvis någon enkel fråga, även om man identifierat

[15] The Economist (2 juni 2018). ”Technology Quarterly: Justice”. https://www.economist.com/technology-quarterly/2018-05-02/justice.

att AI-designen är normativ, vilket även nämnda IEEE-rapport uppmärksammar:

> ”If machines engage in human communities as autonomous agents, then those agents will be expected to follow the community’s social and moral norms. A necessary step in enabling machines to do so is to identify these norms. But which norms?” (IEEE, 2018, s. 36.)

Lagstiftningen kan genom sin nedtecknade existens var enklare att utröna (men inte alltid) än sociala normer. Men samtidigt lär den rättssociologiska litteraturen 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 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 trafikregleringen (jfr Svensson, 2008). Den typen av insikter skulle med fördel vara närvarande i utvecklingen av design för autonoma fordon. 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. Facebooks innehållsmoderering, som använder sig av både mänskliga granskares bedömningar, användarflaggning och AI-verktyg, är ett exempel (jfr Gillespie, 2018; Larsson, 2018b). Också destruktiva och våldsbejakande gruppers normer erbjuder ”lärande” för autonoma intelligenser, vilket visar på komplexiteten att hantera och det kunskapsbehov om normer som autonom teknologi väcker. Ett exempel på utmaningar 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 och kvinnoförnedrande tweets för att sedan stängas ned av sin skapare.[16] Samtidigt har rimligen experimentet inneburit ett lärande för botutvecklingen, där frågor om att filtrera språk och göra normativa ställningstaganden är en del av diskussionen (se Larsson, 2018c). Kunskapen kring ansvarsfrågor och ansvarsfördelning för utvecklare och användare av självlärande tjänster och processer är något som behöver utvecklas då en neutral position ofta tycks omöjlig. Vilket ansvar har den som skapar agens för vad agenten producerar?

[16] https://en.wikipedia.org/wiki/Tay_(bot).

Normativ design och ”neutrality lost”

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: ska man reproducera världen som den är eller som man eftersträvar att den ska vara? Och vems framtidsvilja avser man?[17] Det finns rimligen flera algoritmberoende sammanhang som leder till automatiserade normativa beslut och som därmed behöver behandlas som just normativa. AI väcker behovet av att rättvisefrågor lyfts i samband med användningen av så tekniskt sofistikerade applikationer att vi pratar om dom som intelligenta, framförallt när de också blir vardagliga. Det är här viktigt att uppmärksamma att applikationer som använder sig av data från samhället – som innehåller en rad strukturella skevheter, maktförhållanden, ojämställdhet, brist på likabehandling, diskriminering och rasism – så kommer den normativa frågan ofrånkomligen att landa i tjänstedesignen: hur hanterar man det? Försöker man, som varit vanligt för vissa typer av plattformsintermediärer, 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?[18] När bör man aktörerna ta ansvar, göra normativa ställningstaganden och genom sin design av automatiserade processer motverka samhällelig bias? 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 dock dels i att man riskerar att stoppa huvudet i sanden för de konsekvenser som uppstår, att göra en svag riskbedömning, och – framförallt – riskera att bygga klandervärda system som diskriminerar och adderar till samhällets obalanserade strukturer och rakt av inte fungerar som tänkt på olika typer av individer. Vissa applikationer kan ha oerhört allvarliga konsekvenser – sociala medieplattformar beskylls för att ha en roll i att manipulera politiska val och undergräva demokratiska strukturer; straffvärderingssystem riskerar att bidra till längre fängelsestraff baserat på ovidkommande grunder; medicinskt diagnosställande och autonoma fordon riskerar att döda människor.

Det finns också en växande insikt i designsammanhang, exv. indikerat i IEEE-rapporten ovan (2018) och i flera rapporter från forskningsinstitutet AI Now – att kulturella värderingar och samhällelig bias ofrånkomligen ryms i den

[17] Vilket exv. lyfts i en kommentar av forskare i tidskriften Nature (Zou & Schiebinger, 18 juli 2018).
[18] Se exempel 1 om bilddatabas med genderbias och exempel 3 om automatiserade jobberbjudanden ovan.

data som relaterar till individer och därmed måste hanteras på ett ansvarsfullt sätt. Samtidigt kan man från ett rättssociologiskt perspektiv konstatera att det inte är någon enkel sak eller någon ”quick fix” att hantera normativa ställningstaganden. I avsaknad av en neutral position kommer fler AI-utvecklare att nödgas ta normativ ställning i frågor de troligen skulle vilja undvika, vilket stärker argumentet om att ingenjörsutbildningar om AI, bildanalys och algoritmer bör inkludera ansvarsfrågor i relation till samhälleliga eller etiska konsekvenser av den design de lär sig applicera och utveckla. Man kan också tänka sig att det är en fråga för ett växande antal styrelserum, för företag som verkar på konsumentmarknader. De vill naturligtvis öka sina intäkter, exv. genom mer träffsäker reklam eller individualiserad tjänsteutveckling, men vilka andra effekter kan ett maskinlärande och mönsteranalyserande system ha, och vilka etiska överväganden behöver göras? Finns det en risk att en individualiserad prisbild genom proxy utgör en s.k. teknologisk redlining eller att man på analytisk väg snarare manipulerar än influerar en redan utsatt grupp att konsumera eller skuldsätta sig på ett ohållbart sätt?[19]

En central insikt här är betydelsen av normativ design och ansvar. Det finns för många applikationer inte längre någon sann neutral position eftersom alla tillgängliga alternativ kan kräva kontroversiella ställningstaganden av normativ karaktär (Larsson, 2018c). En bilddatabas med genderbias skulle exempelvis kunna vara deskriptivt korrekt, dvs. beskriva samhället som det är i dess befintliga ojämställdhet med fler kvinnor i köket och fler män intresserade av jakt (som i exempel 1 ovan), men den design som reproducerar och ”lär sig” baserat på dessa förhållanden erhåller också en aktiv agens i förhållande till den befintliga ojämställdheten. Den som designar måste därmed ta ställning normativt om den vill bidra till det eller motverka det. Båda alternativen är normativa.

Behovet av process

Att teckna principer för hur man som AI-utvecklare ska förhålla sig normativt är därmed i vissa fall behövligt, och visar på insikt, tyder på ansvarstagande och kan innebära lärande i designarbetet. Men den medföljande normativiteten kommer med fler utmaningar än så, vilket de rättsliga strukturerna i sig är goda markörer på, med processuella ställningstaganden om prövning och

[19] Diskuterat bl.a. i Larsson et al. (2016) och i digitala konsumentsammanhang, av Larsson (2016; 2017b).

överprövning, bedömarens oberoende ställning, rättssäkerhet och likabehandling. Principer (lag), lär oss rättsvetenskapen, är inte mycket värda utan en tydlig process för deras användning och praktiska implementering. De exakta gränserna för vart och ett koncept som beskrivs i en generell princip kan behöva återkommande definieras mot specifika fall. Tolkningen kan även guidas av andra (rätts)källor som kunde motsvara förarbeten som mer utförligt förklarar syftet med skrivelsen, eller av utpekad expertis (doktrin). För närvarande finns inte någon konsensus kring vilken normativ hållning som är bäst för AI i samhället, och än mindre konsensus kring det processuella – även om mycket arbete pågår inom betydande instanser som EU-kommissionen och IEEE.

Dock, att ens uppmärksamma utmaningen och det ansvar som kan medfölja, i första hand, är ett betydande steg i rätt riktning.

Konklusioner: den normativa designen

Syftet är här att ta ett första steg i en bred rättslig och rättssociologisk riktning genom att teckna några av de rättsliga och normativa utmaningar som AI medför. Det tycks finnas en förhärskande idé om att etiska principer är något som man vid ett enstaka tillfälle kan besluta om, teckna ner och sedan genomdriva. För vissa principiella ställningstaganden – som att AI inte bör användas till vapen – kan det vara helt centralt.[20] Men för många andra sammanhang av normativ eller värdereproducerande karaktär är det snarare samhällsvetenskaplig eller humanistisk grundforskning som behöver knytas till AI-utveckling av pågående karaktär. De värderingsmässigt svåra frågorna tenderar att vara flervetenskapliga forskningsfrågor som tecknar sig i samverkan med teknologiska, datavetenskapliga eller matematiska framsteg, som ett rörligt mål. Dessa är inte vid ett enstaka tillfälle lösbara för att gälla för all framtid eftersom både etiken och normerna såväl som AI-förutsättningarna är i pågående förändring.

Styrkan i argumentet om att design av AI är normativt ligger i erkännandet av att kompetens kring värderingar, normer och etik i så fall behöver vara närvarande i utveckling och applicering för att undvika just nämnda risker. Eftersom det otvetydigt finns en sådan potential och möjlig samhällsnytta i en god användning av AI och ML kan det samhälleliga perspektivet också

[20] Jfr Lethal Autonomous Weapons Pledge, https://futureoflife.org/lethal-autonomous-weapons-pledge/.

uttryckas i termer av ett behov av att förstå hur tillit och social acceptans uppnås för applikationerna. Lämplig typ av transparens, en uttänkt och formaliserad ansvarsfördelning och tydliga indikationer på att det autonoma systemet inte förstärker eller reproducerar samhälleliga skevheter och partiskhet, som diskriminering, eller på annat sätt förstör grundläggande samhälleliga funktioner kan därmed framhållas som helt centrala för tillitsbyggandet.

När det diskuteras reglering – antingen om behovet av ny eller dess eftersläpning eller ett mer slentrianmässigt ”reglera inte sönder vår marknad” – så ska man komma ihåg att det redan finns välgrundad reglering med bred legitimitet för många av de element och applikationer som en datadriven artificiell intelligens används till. Det finns redan diskrimineringsgrunder, marknadsrätt och dataskydd. Utmaningarna för dessa typer av reglering finns i förhållande till autonoma system oftare i upptäckt, tillsyn och implementering, och delvis i den konceptuella utmaningen att relatera traditionella idéer om diskriminering, medbestämmande och otillbörlighet till nya marknadspraktiker. Finns det diskriminerande mönster i realtidsindividualiserad prisdiskriminering? Hur långt bör individens bestämmande över sina ansiktsfotografier sträckas om dessa används för att med generativa adverseriella nätverk (GANs) framställa nya, delvis artificiella, mänskliga ansiktsbilder som inte längre ser ut att representera ursprungspersonen? Sistnämnda närmar sig rättsutvecklingsfrågor. På dataskyddsområdet har det redan uppstått en betydande kritik kring medgivandepraktiker för användande av personuppgifter på datadrivna marknader. Nätmiljön, inklusive smarta telefoner, har kommit att bli så datainsamlande att det inte längre finns någon praktisk överenskommelse med individer om hur deras information ska hanteras (Larsson, 2018a). Mycket av detta träffar AI och ML, så som den appliceras på stora datamängder, används till individualisering, riktadhet, mönsteranalys, rekommendationer, att anta samband och prediktera kommande händelser och beteenden. Hur mycket av en bakomliggande logik för automatiserat beslutsfattande bör den berörde förstå eller erbjudas? Vad är praktiskt men också tekniskt möjligt? Det framstår därmed dels som att det redan finns mycket reglering av relevans för samhällsapplicerad AI – men där implementeringen är den stora utmaningen – och dels som att specifik rättsutveckling i vid mening, kanske som branschöverenskommelser, utveckling av standarder och certifiering kan krävas i relation till särskilt samhällsallvarlig användning, men också som en processuell konsekvensbedömningsfråga i relation till transparens och ansvarsfördelning.

Några av de mest centrala utmaningar som utvecklats ovan är därmed:

1. Behovet av tvärvetenskaplig och mångvetenskaplig approach: De teknik- och matematiktillvända discipliner som huvudsakligen utvecklar de oerhört sofistikerade applikationer som artificiell intelligens, bildigenkänning och varianter av ML innebär tenderar att inte organiseras nära de discipliner som genom århundraden utvecklat teori och metoder kring just värderingar, normer och etik. Här behöver broar byggas. Det talar dels för behovet av mer medvetenhet kring värderingsfrågor och normativitet i designsammanhang, och dels för behovet av mång- och tvärvetenskap i både forskning, utveckling och utbildning. Etiska, rättsliga och sociala frågor bör heller inte betraktas som ett utanpåverk till den AI-utveckling som pågår inom datavetenskapliga eller matematiska institutioner utan viktig komplementär expertis som kan bidra till AI-forskning, algoritmutveckling och maskininlärning som sådan. En del ökända applikationer har varit konsekvenser av dålig design, baserad på alltför ensidig kompetens.

2. Principer utan process är verkningslösa: Ett erkännande av normativitet innebär också att någon idé om process och implementering måste medfölja. Här bjuder samhällenas flerhundraåriga byggande av rättsordningar på exempel att lära av för principiell utveckling av AI och ML, exv. gällande analogier av behovet av att ”väcka åtal” i enlighet med det normativa ställningstagandet; hur motsvarigheten till den dömande makten organiseras; hur man bedömer att det enskilda fallet relaterar till den generella principen, osv.

3. Betydelsen av kontext: Erkännande av normativitet som en empirisk företeelse innebär att man ofrånkomligen kommer att få hantera kontextuella avvikelser och direkta normativa motsättningar – vilka normer ska gälla? Facebook, för att ta ett exempel, har över 2 miljarder aktiva användare och verkar därmed en stor mängd kulturer och i hundratals rättsordningar. Många av dessa har olika och ibland helt oförenliga kulturella preferenser för och sociala och rättsliga normer för företeelser som sexualitet och relationer, för nakenhet, för etnicitet och social ställning.

4. Behovet av tillsynskompetens och konsekvensanalys: Metodutveckling för myndigheternas tillsynsverksamhet är nödvändig i ljuset av att automatiserad användning av AI och ML möjliggör ett kraftigt decentraliserat utfall där insynen främst är tillgänglig för var och en användare eller individuell adressat. Metoder behövs för att upptäcka diskriminerande mönster eller annat otillbörligt på mer strukturell nivå i stil med ovan nämnda teknologiska ”redlining”, men även för standardisering av samhällelig konsekvensanalys vid

införande av AI-processer på såväl konsumentmarknader som i offentliga verksamheter.

5. Transparensutmaningen: för var och en typ av applikation behöver en intresseavvägning göras på samhällsnivå för vilken grad av transparens som är eftersträvansvärd. Applikationer med låg nivå av möjlig insyn och förklarbarhet bör rimligen undvikas på områden med potentiellt allvarliga utfall. Och, vidare: på vilken nivå, med vilken symbolik och vilka metaforer kan en acceptabel förklaring av ett AI-genererat utfall ske?

Det är viktigt att understryka att dessa utmaningar inte på något sätt är menat att avskräcka från att arbeta med normativa frågor i relation till artificiell intelligens. Tvärtom, de är menade att bidra till och förtydliga vilka frågor som behöver mer arbete, mer kunskap och medvetandegörande. Vi lever redan i hög utsträckning i en väldigt digitaliserad tillvaro där den data vi genererar genom våra liv används och återanvänds för en ökad grad av automatiserade processer och alltmer autonomt beslutsfattande. Vi kommer att leva alltmer jämsides med artificiell intelligens och maskininlärning av olika slag framöver, eftersom metoderna och teknologierna redan visat på en stor potential. Det är därmed desto viktigare att bidra till att den utvecklas i en legitim och tillitsbyggande riktning i mänsklighetens tjänst.

Referenser

Agüera y Arcas, B, Mitchell, M. & Todorov, A. (6 maj 2017). ”Physiognomy’s New Clothes”, i *Medium*: https://perma.cc/8FVG-TWSQ.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973-989.

Andersson Schwarz, J. (2017). Platform Logic: An Interdisciplinary Approach to the Platform-Based Economy. *Policy & Internet* 9 (4): 374-394.

Andersson Schwarz, J. & Larsson, S. (red., 2019). *Plattformssamhället. Den digitala utvecklingens politik, innovation och reglering*. Stockholm: Fores.

Bastos, M. T., & Mercea, D. (2017). The brexit botnet and user-generated hyperpartisan news. *Social Science Computer Review*, https://doi.org/10.1177/0894439317734157.

Bozdag, E. (2013). Bias in Algorithmic Filtering and Personalization. *Ethics and Information Technology* 15 (3): 209-227.

Broniatowski, D.A. et al. (2018). Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate, *American Journal of Public Health*, published online before print. DOI: 10.2105/AJPH.2018.304567

Brundage, M. et al. (2018) *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. https://maliciousaireport.com.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency* (pp. 77-91).

BuzzFeed News (11 april 2018) "Why Facebook Will Never Fully Solve Its Problems With AI", av Davey Alba. https://www.buzzfeednews.com/article/daveyalba/mark-zuckerberg-artificial-intelligence-facebook-content-pro.

Campolo, A., Sanfilippo, M., Whittaker, M., & Crawford, K. (2017). *AI Now 2017 Report*. AI Now Institute at New York University.

Caplan, R., Donovan, J., Hanson, L. and Matthews, J. (2018). *Algorithmic Accountability: A Primer*, NYC: Data & Society.

COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE EUROPEAN COUNCIL, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS. Artificial Intelligence for Europe {SWD(2018) 137 final}.

Christl, W. (2017). *Corporate Surveillance in Everyday Life: How Companies Collect, Combine, Analyze, Trade, and Use Personal Data on Billions* (Report). Vienna: Cracked Labs. Retrieved from http://crackedlabs.org/en/corporate-surveillance

Cummings, M.L., Roff, H.M., Cukier, K., Parakilas, J. & Bryce, H. (2018). *Artificial Intelligence and International Affairs. Disruption Anticipated*. Chatham House Report.

Datta, A., Tschantz, M.C., Datta, A. (2015). Automated Experiments on Ad Privacy Settings – A Tale of Opacity, Choice, and Discrimination. *Proceedings on Privacy Enhancing Technologies*. 1: 92–112, DOI: 10.1515/popets-2015-0007.

Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism*, 3(3), 398-415.

Dolata, U. (2017). *Apple, Amazon, Google, Facebook, Microsoft: Market concentration-competition-innovation strategies* (No. 2017-01). Stuttgarter Beiträge zur Organisations-und Innovationsforschung, SOI Discussion Paper.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Shieber, S., Waldo, J., Weinberger, D. & Wood, A. (2017). Accountability of AI under the law: The role of explanation. *arXiv preprint* arXiv:1711.01134.

The Economist (2 juni 2018). "Technology Quarterly: Justice", https://www.economist.com/technology-quarterly/2018-05-02/justice.

Friedman, B. & Nissenbaum, H. (1996). "Bias in Computer Systems," *ACM Transactions on Information Systems*, 14(3): 330-347.

Future of Life. Lethal Autonomous Weapons Pledge, https://futureoflife.org/lethal-autonomous-weapons-pledge/.

Förordning (2017:309). Om försöksverksamhet med självkörande fordon.

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

Gillespie (24 augusti 2017). ”The Platform Metaphor, Revisited”, i bloggserien *How metaphors shape the internet*, med C Katzenbach och S Larsson som redaktörer. https://www.hiig.de/en/the-platform-metaphor-revisited/.

The Guardian (18 mars 2018) ”‘I made Steve Bannon’s psychological warfare tool’: meet the data war whistleblower”, av Carole Cadwalladr: https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys* (CSUR), 51(5): 93.

Haider, J., & Sundin, O. (2018). *Invisible Search and Online Search Engines: The ubiquity of search in everyday life*. Chicago: Routledge Studies in Library and Information Science.

Hargittai, E. (2007). The social, political, economic, and cultural dimensions of search engines: An introduction. *Journal of Computer-Mediated Communication*, 12(3), 769-777.

Hevelke A., & Nida-Rümelin, J. (2015). Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis. *Science and Engineering Ethics* 21(3): 619–630.

Hildebradt, M. (2015). *Smart Technologies and the Ends of Law*, UK & USA: Edward Elgar Publishing.

Hydén, H. (2002). *Normvetenskap*. Lund: Sociologiska institutionen.

Hydén, H. & Svensson, M. (2008). The concept of norms in sociology of law. In: Wahlgren P (ed.) *Scandinavian Studies in Law*. Stockholm: Law and Society, pp. 15–33.

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

Inspektionen för socialförsäkring, ISF (2018). *Profilering som urvalsmetod för riktade kontroller. En granskning av träffsäkerheten, effektiviteten och rättssäkerheten i Försäkringskassans modeller för riskbaserade kontroller*, Stockholm: Rapport 2018:5.

Jareborg, N. (1993). *Uppsåt och oaktsamhet.* Göteborg: Iustus Förlag AB.

Larsson, S. (2018a). Algorithmic Governance and the Need for Consumer Empowerment in Data-driven Markets, *Internet Policy Review* 7(2):1-12.

Larsson, S. (2018b). ”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.

Larsson, S. (2018c). ”Sjyst AI och normativ design”, i Akenine, D. & Stier, J. (red.) *Människor och AI*. Stockholm: AddAI.

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

Larsson, S. (2017b). ”Digital konsumentprofilering. Stora data, prediktiv analys och policyutmaningar”, i Sandberg, A. (red.) *Kunskapsöversikter inom det konsumentpolitiska området*, Konsumentverket. 2017:1.

Larsson, S. (2017c). Sustaining Legitimacy and Trust in a Data-driven Society, *Ericsson Technology Review* 94(1): 40-49.

Larsson, S. (2016). *Digitalisering och konsumentintresset. En litteraturstudie*. Karlstad: Konsumentverket 2016:12.

Larsson (2013). Sociology of Law in a Digital Society – A Tweet from Global Bukowina, *Societas/Communitas* 15(1): 281-295.

Larsson, S. & Andersson Schwarz, J. (red. 2018) *Developing Platform Economies. A European Policy Landscape*. Stockholm: Fores

Larsson, S. & Ledendal, J. (2017). *Personuppgifter som betalningsmedel*, Karlstad: Konsumentverket. 2017:4.

Larsson, S., Svensson, L., & Carlsson, H. (2016) *Digital Consumption and Over-Indebtedness Among Young Adults in Sweden*, LUii reports, Lund & Landskrona, Sweden: Lund University.

Ledendal, J., Larsson, S. & Wernberg, J. (2018). *Offentlighet i det digitala samhället – vidareutnyttjande, sekretess och dataskydd*, Stockholm: Norstedts Juridik.

Leigh Anderson, S. (2008) Asimov’s ”three laws of robotics” and machine metaethics, *Ai & Society* 22(4): 477-493.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*. Vol 267: 1-38. https://doi.org/10.1016/j.artint.2018.07.007.

Morozov, E. (2013). *To save everything, click here: Technology, solutionism, and the urge to fix problems that don’t exist*. Penguin UK.

Motherboard (25 oktober 2017) ”Google’s Sentiment Analyzer Thinks Being Gay Is Bad”, av Andrew Thompson. https://motherboard.vice.com/en_us/article/j5jmj8/google-artificial-intelligence-bias (senast besökt 17 juli 2018).

Narayanan, A. (2018) ”21 fairness definitions and their politics”, presenterad på konferens om Fairness, Accountability, and Transparency, 23 februari 2018. http://fairmlbook.org/tutorial2.html.

The New York Times (9 oktober 2017) ”Why Stanford Researchers Tried to Create a ‘Gaydar’ Machine”, av Heather Murphy. https://www.nytimes.com/2017/10/09/science/stanford-sexual-orientation-study.html.

Noble, S. U. (2018). *Algorithms of Oppression: How search engines reinforce racism*. NYU Press.

Norton, W. (2013). *Cultural Geography: Environments, Landscapes, Identities, Inequalities*. Oxford University Press.

Pasquale, F. (2015). *The Black Box Society. The Secret Algorithms That Control Money and Information*, Harvard University Press.

Pasquale, F. (2017, September 12). ”Exploring the Fintech Landscape.” Written Testimony of Frank Pasquale Before the United States Senate Committee on the Banking, Housing, and Urban Affairs. Available at https://www.banking.senate.gov/imo/media/doc/Pasquale%20Testimony%209-12-17.pdf.

Petrazycki, L. (1955) *Law and Morality*. Cambridge, MA: Harvard University Press.

Pichai, S. (7 juni, 2018). ”AI at Google: our principles”, https://www.blog.google/technology/ai/ai-principles/.

Pound, R. (1910). Law in books and law in action. *American Law Review*, 44:12.

Pound, R. (1917). The limits of effective legal action, *International Journal of Ethics* 27: 150-167.

ProPublica (23 maj 2016). ”Machine Bias”, av Julia Angwin et al. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology* 20(1): 5-14.

Regeringskansliet (2018). *Nationell inriktning för artificiell intelligens*. Näringsdepartementet.

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018) *Algorithmic Impact Assessment. A practical Framework for Public Agency Accountability*. NYC: AI Now.

Renner, K. (1949). *The Institutions of Private Law and their Social Function* (reprint 1976). London: Routledge Kegan Paul.

SKL (2018) *Automatiserat beslutsfattande i den kommunala förvaltningen*.

Shapiro, A. (2017). Reform predictive policing. *Nature*, *541*(7638).

Självkörande fordon på väg, dir. 2015:114.

Spiekermann, S., & Korunovska, J. (2016). Towards a value theory for personal data. *Journal of Information Technology*, 23(1): 62-84. doi:10.1057/jit.2016.4.

Svensson, Marina. (2018) ”Plattformssamhället i Kina: Ett övervakningssamhälle i ett eget ekosystem”, i Andersson Schwarz, J. & Larsson, S. (red.) *Plattformssamhället. Den digitala utvecklingens politik, innovation och reglering*. Stockholm: Fores.

Svensson, Måns. (2008). *Sociala normer och regelefterlevnad: Trafiksäkerhetsfrågor ur ett rättssociologiskt perspektiv*. Lund: Lund Studies in Sociology of Law 28, Lunds universitet.

Svensson, M., & Larsson, S. (2012). Intellectual Property Law Compliance in Europe: Illegal File sharing and the Role of Social Norms, *New Media & Society*, 14(7): 1147-1163.

Svensson, L. & Larsson, S. (2018). *Digitalisering av kommunal socialtjänst. En empirisk studie av en organisation och profession i förändring*. FoU-rapport 2018:1. Helsingborg stad.

Svensson, L. & Larsson, S. (2017). *Digitalisering och socialt arbete – en kunskapsöversikt*. LUii reports.

Troumbley, R. L. (2015). *Taboo language and the politics of American cultural governance*. Doctoral dissertation, University of Hawai'i at Manoa.

Vinnova (2018). *Artificiell intelligens i svenskt näringsliv och samhälle. Analys av utveckling och potential – Slutrapport*.

Vägen till självkörande fordon – försöksverksamhet (SOU 2016:28).

Vägen till självkörande fordon – introduktion (SOU 2018:16).

Wired (21 augusti 2017). ”Machines taught by photos learn a sexist view of women”, av Tom Simonite. https://www.wired.com/story/machines-taught-by-photos-learn-a-sexist-view-of-women/amp.

Wu, X., & Zhang, X. (2016). Automated inference on criminality using face images. *arXiv preprint arXiv:1611.04135*, 4038-4052.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint* arXiv:1707.09457.

Zou, J. & Schiebinger, L. (18 juli 2018) ”AI can be sexist and racist — it’s time to make it fair”, *Nature*, comment.