UPPSALA
UNIVERSITET

UPTEC STS 19049
Examensarbete 30 hp
November 2019

Artificiell intelligens och gender bias

En studie av samband mellan artificiell intelligens, gender bias och könsdiskriminering

Hanna Lycken

UPPSALA
UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

Addressing Gender Bias in Artificial Intelligence

Hanna Lycken

Recent advances in, for example, machine learning and neural networks have taken artificial intelligence into disciplines such as justice, recruitment and health care. As in all fields subject to AI, correct decisions are crucial and there is no room for discriminatory conclusions. However, AI-systems are, just like humans, subject to various types of distortions, which can lead to unfair decisions. An alarming number of studies and reports show that AI in many cases reflects and reinforces existing gender bias in society. Algorithms used in image recognition base their decisions on character stereotypes of male and female. Voice recognition is more likely to correctly recognize male voices compared to female voices, and earlier this year the United Nations released a study showing that voice assistants, such as Microsoft's Cortona or Apple's Siri, reinforce existing gender bias.

The purpose of this study is to investigate how gender discrimination can appear in AI-systems, and what constitutes the relationship between gender bias, gender discrimination and AI-systems. Furthermore it addresses how a company that works with the development of AI reason concerning the relationship between gender bias, gender discrimination and AI development. The study contains a thorough literature review, as well as in-depth interviews with key persons working with various aspects of AI development at KPMG.

The results show that bias in general, and gender bias in particular, are present at all stages of AI development. It can occur due to a variety of factors, including but not limited to the lack of diversity in the workforce, the design of algorithms and the decisions related to how data is collected, encoded and used to train algorithms. The solutions proposed are partly about addressing the identified factors, but also about looking at the problem from a holistic perspective. The significance of seeing and understanding the links between gender bias in society and gender bias in AI-systems, as well as reconsidering how each factor depends on and correlates with other ones, is emphasized. The essence of the results is that it is not enough to alter any of the parameters unless the structure of the system is changed as well.

Handledare: Tobias Holmström Andersson
Ämnesgranskare: Anders Arweström Jansson
Examinator: Elisabet Andersdottir
ISSN: 1650-8319, UPTEC STS 19049

Populärvetenskaplig sammanfattning

Utvecklingen av artificiell intelligens, AI, har de senaste åren gjort enorma framsteg. AI spås få lika stor påverkan på samhället som elektricitet haft och avancemangen inom till exempel maskininlärning och neurala nätverk har tagit AI in i sektorer som rättsväsende, rekrytering och hälso- och sjukvård – områden där beslut i många fall är livsavgörande och det inte finns utrymme för felaktiga eller diskriminerande slutsatser. Men AI-system är, precis som människor, känsliga för olika typer av snedvridningar, vilket kan leda till orättvisa beslut. I takt med att användningen av AI ökar så ökar inte bara efterfrågan på AI utan även oron för dess trovärdighet och opartiskhet. Samtidigt visar en alarmerande mängd studier och rapporter att AI i flera fall speglar, sprider och förstärker befintliga snedvridningar i samhället i form av fördomar och värderingar vad gäller könsstereotyper och könsdiskriminering. Det har på flera håll uppmärksammats fall där AI-system ger könsdiskriminerande resultat, och rapporter tyder på att oberoende typ av AI-system så finns gender bias ofta närvarande. Algoritmer som används i bildigenkänning baserar sina beslut på stereotyper om vad som är manligt och kvinnligt, röstigenkänning är mer trolig att korrekt känna igen manliga röster jämfört med kvinnliga röster och röstassistenter som Microsoft:s Cortona eller Apple:s Siri förstärker befintlig könsdiskriminering i samhällen. Produkterna från AI-industrin påverkar miljontals liv och finns redan i ett stort spann av sektorer. Att ta itu med mångfald- och diskrimineringsfrågor är därför inte bara i teknikindustrins intresse utan något centralt för alla vars liv påverkas av AI- verktyg och tjänster.

Denna studie har som syfte att undersöka hur könsdiskriminering kan uppstå i AI-system generellt, hur relationen mellan gender bias och AI-system ser ut samt hur ett företag som arbetar med utveckling av AI resonerar kring relationen mellan gender bias och AI-utveckling. Studien har gjorts i samarbete med KPMG:s tekniska avdelning Digital Transformation and Innovation (DTI) som bland annat arbetar med rådgivning i digitala satsningar och innovationsprojekt. Studiens syfte uppfylls genom en litteraturgenomgång samt djupintervjuer med nyckelpersoner som på olika sätt arbetar med AI-utvecklingen på KPMG. Både personer från KPMG:s kontor i Stockholm och från KPMG:s kontor i Danmark har intervjuats.

Resultaten visar att bias i allmänhet och gender bias i synnerhet finns närvarande i alla steg i utvecklingen av AI och kan uppstå på grund av en mängd olika faktorer, inklusive men inte begränsat till mångfald i utvecklingsteamen, utformningen av algoritmer och beslut relaterade till hur data samlas in, kodas, eller används för att träna algoritmer. Studien visar på ett antal huvudsakliga faktorer som kan leda till könsdiskriminering i AI-system generellt. Dessa är: redan existerande snedvridning (praxis och attityder i samhälle, hos institutioner eller hos individ), snedvriden data, brist på mångfald i arbetskraft och oförståelse kring samband mellan teknik och genusanalys. Ytterligare några faktorer som är specifika för det undersökta företaget är kravställning från kund

samt tidspress. De lösningar som föreslås handlar dels om att adressera respektive orsaksfaktor som identifierats, men även att se problemet med gender bias och könsdiskriminering i AI-system från ett helhetsperspektiv. Det handlar dels om att lyfta, förstå och adressera sambanden mellan könsdiskriminering i AI-system och hur vårt samhälle ser ut och fungerar, men även att reflektera kring hur respektive orsaksfaktor beror av och korrelerar med andra faktorer. Essensen av resultaten, och slutsatserna som följer av dem, är att det inte räcker att ändra någon av parametrarna om inte systemets struktur samtidigt ändras. Det strålkastarljus gender bias har fått tack vare de tydliga exempel på könsdiskriminering och gender bias som AI-system visat ger både bevis på den könsdiskriminering som existerar, och nya möjligheter att öppna upp för diskussioner kring ämnet för att försöka hantera det.

Förord

Denna uppsats är skriven som det sista momentet på civilingenjörsprogrammet System i Teknik och Samhälle vid Uppsala universitet. Arbetet har skrivits i samarbete med avdelningen Digital Transformation and Innovation på KPMG i Stockholm. Jag vill tacka alla involverade på KPMG för ett stort engagemang och hjälpsamhet. Särskilt tack till alla som deltagit i den empiriska studien och givit mig en del av sin värdefulla tid, samt till min handledare på KPMG, Tobias Holmström Andersson, som hjälpt till med sitt stöd och feedback under arbetets gång.

Slutligen vill jag rikta ett extra stort tack till min ämnesgranskare, Anders Arweström Jansson vid Uppsala universitet, som bidragit med expertis, feedback och uppmuntran.

Trevlig läsning!

Hanna Lycken
Stockholm, november 2019

Begreppslista

Algoritm: En process eller uppsättning regler som ska följas vid beräkningar eller andra problemlösningsoperationer.

Artificiell intelligens, AI: En algoritm eller maskin som lär sig och sedan agerar på dessa lärdomar.

AI-system: Ett system är en grupp samverkande eller sammanhängande enheter som bildar en enhetlig helhet. Ett system begränsas av tid och rum, påverkas av sin omgivning och uttrycks i sin funktion. I föreliggande rapport förstås ett AI-system som en eller flera AI-algoritmer som lär sig och sedan agerar på dessa lärdomar. AI-algoritmerna påverkas av sin omgivning och kontexten de implementeras i, men även av exempelvis vilken data som används i dem.

Bias: Engelska för fördom, snedvridning. Innebär en avvikelse från det "sanna" värdet på grund av kognitiva eller statistiska faktorer så som exempelvis något systematiskt fel i forskningsprocessen vad gäller t.ex. insamling av data, bearbetning eller analys av resultat.

Gender: Det engelska ordet för genus (socialt kön).

Gender bias: Preferens, fördom eller diskriminering mot ett specifikt kön.

Genus (socialt kön): Syftar på det sociala könet eller det sätt på vilket vi reproducerar kön i samhället. Det handlar till exempel om hur vi skapar normer för manligt och kvinnligt i vår vardag genom kroppsspråk, val av kläder, yrken, intressen och så vidare.

Kön (biologiskt kön): Egenskap hos individ som beror på vilken typ av gameter (könsceller) den producerar.

Intersektionalitet: Från engelskans "intersection". Termen används för att beteckna hur olika maktordningar och diskrimineringsgrunder påverkar och ibland förstärker varandra. Det är ett analytiskt perspektiv som uppmärksammar hur relationer av överordning och underordning skapas och upprätthålls i samspel mellan bland annat etnicitet, kön, genus, klass, och ålder.

Modell: Försök till avbildning av verkligheten, matematisk beskrivning av ett fenomen.

Ordinbäddning: Från engelskans "word embedding". En maskininlärningsmetod som kartlägger text på ett sätt som fångar semantiska likheter mellan ord.

Snedvridning: En översättning av engelskans "bias". Innebär en avvikelse från det "sanna" värdet på grund av kognitiva eller statistiska faktorer. Kan avse resultat som är systematiskt mindre gynnsamma för individer inom en viss grupp.

Innehållsförteckning

1. Introduktion och kontext

Artificiell intelligens, AI, har trängt igenom varje aspekt av vårt dagliga liv och anses av somliga vara det viktigaste paradigmskiftet i teknikhistorien (MMC Venture 2019). AI spås få lika stor påverkan på samhället som elektricitet haft (Mahendra 2019) och med bakgrund i hur hela vårt samhälle utvecklats med och tack vare den tekniken är det svårt, om inte omöjligt, att förutspå alla tillämpningsområden och konsekvenser AI kan komma att medföra. Det anses vedertaget att teknikutveckling generellt till stor del formar framtiden (UNESCO 2019). Men ett problem med berättelser om teknikutveckling är att de tenderar att ge allt för mycket makt åt själva tekniken medan bredare sociala relationer döljs. Som Mackenzie och Wajcman påpekar i "The Social Shaping of Technology" (1999) kan ett samhälles utformning spela en stor roll i bestämmandet av vilka teknologier som antas. Samma teknologier kan dock ha väldigt olika effekter på samhället och dess invånare beroende på i vilken situation teknologierna används. Teknologier är inte deterministiska, utan snarare kristalliseringar av samhället som formas av den sociala kontexten från vilka de kommer (Ford och Wajcman 2017; Howcroft och Rubery 2019). Resultaten av teknik påverkas bland annat av socioekonomiska sammanhang och företags beslutsfattande och teknisk innovation främjar ofta vissa medan den motverkar andra gruppers intressen. Bakom utvecklingen av artificiell intelligens finns reella krafter av pengar, makt och data vilket har betydelse för utvecklingen (MacKenzie och Wajcman 1999).

Forskning på AI har funnits i över sex decennier, men de senaste årens framsteg inom exempelvis maskininlärning och neurala nätverk har tagit AI in i områden som hälso- och sjukvård, rättsväsende och rekrytering – områden där det inte finns utrymme för felaktiga eller diskriminerande slutsatser eller beslut. I många samhällen har beslut och val som tidigare togs av människor i allt större utsträckning blivit delegerade till algoritmer som ger råd och i vissa fall även tar beslut om hur data ska tolkas och vilka beslut som ska tas baserat på den tolkningen. AI används i domstolar för att beräkna sannolikhet att den åtalade begår ett nytt brott, i hälso- och sjukvård för att diagnostisera patienter och i olika typer av rekommendationssystem som ger råd om saker som var och när användare ska träna, vilka de ska kontakta eller vilken väg de ska välja (de Vries 2010: Mittelstadt et al. 2016). Men AI-system är, precis som människor, känsliga för olika typer av snedvridningar vilket kan leda till orättvisa beslut. Snedvridning i den verkliga världen kan smyga in i AI-system, och det är därför viktigt att de som utvecklar AI-system är medvetna om vilka potentiellt skadliga effekter AI kan ha på sin omvärld (Mehrabi et al. 2019).

Rapporter och studier har visat att artificiell intelligens i flera fall inte bara speglar utan även förstärker befintliga snedvridningar i samhället i form av fördomar och värderingar vad gäller stereotyper och diskriminering mellan könen (Howard och Borenstein 2017;

UNESCO 2019; West et al. 2019 m.fl.). Att vita män i stor utsträckning är de som får de största fördelarna av AI-system är något som blir allt tydligare ju mer tekniken utvecklas och används (West et al. 2019). Att AI-system systematiskt diskriminerar (bland annat) kvinnor får konsekvenser vars vidd är svår att på förhand uppfatta, delvis på grund av komplexiteten i tekniken och dess applikationsområden. Diskriminering i sjukvård påverkar vilka som får korrekt diagnostisering och behandling, och diskriminering i rekrytering påverkar exempelvis hur mångfald och könsfördelning på olika arbetsplatser och i olika roller ser ut. Diskriminering i AI-system förstås därför få ringar på vattnet och inom en snar framtid kan AI påverka stora delar av vårt samhälle på sätt vi ännu inte förstår (Howard och Borenstein 2017; Rottingen 2018; UNESCO 2019; West et al. 2019). För att undvika både återupprepning av tidigare misstag med diskriminering i teknikutveckling och skapandet av nya problem, krävs att problemen med exempelvis könsdiskriminering i AI-system synliggörs och hanteras.

Feministforskare menar att genusdimensionen är av betydelse i all samhällsvetenskap. Alvesson och Sköldberg (2008) lyfter vikten av att betrakta genusrelationer även i forskning som inte primärt syftar till att bidra till dess belysande. Med bakgrund i den könssnedvridning den tekniska utvecklingen hittills visat på i kombination med det enorma inflytande artificiell intelligens har och kommer att få på samhället anser många att AI-utvecklingen behöver undersökas utifrån en genusanalys (Alvesson och Sköldberg 2008; West et al. 2019, m.fl.).

1.1 Betydelse av konsultfirmor och introduktion till KPMG

Managementkonsultfirmor är viktiga aktörer i samhällsutvecklingen. De kan i sitt arbete forma andra, och är centrala aktörer när det gäller att definiera problem som företag och regeringar står inför, både i högprofilerade projekt och bakom kulisserna. På senare tid har dessa företag också varit centrala när det gäller att utveckla idén om exempelvis Manufacturing 4.0 (ökad automatisering och datautbyte inom tillverkningsteknologier) och skapandet av "smarta fabriker" (Morgan et al. 2019).

KPMG är ett globalt företag som erbjuder tjänster inom revision-, skatt- och annan rådgivning, och som marknadsförs som ett av världens ledande kunskapsföretag. Denna studie har gjorts i samarbete med KPMG:s tekniska avdelning Digital Transformation and Innovation (DTI). DTI arbetar bland annat med rådgivning i digitala satsningar och innovationsprojekt. På kontoret i Stockholm inleddes nyligen arbetet med artificiell intelligens, medan KPMG Danmark har kommit längre. Av den anledningen har information från både KPMG Sverige och KPMG Danmark använts. KPMG:s kunder är allt från nationella och internationella storföretag till offentlig verksamhet och ideella organisationer. KPMG är därför en viktig aktör i samhällsutvecklingen.

1.2 Syfte och frågeställning

Denna studie har två syften. Studiens första syfte är att förstå hur könsdiskriminering kan uppstå i AI-system generellt och hur relationen mellan gender bias och AI-system

ser ut. Studiens andra syfte är att undersöka hur ett företag som arbetar med utveckling av AI resonerar kring relationen mellan gender bias och AI. För att uppfylla syftet har följande frågeställningar undersökts:

- *Hur kan gender bias eller könsdiskriminering uppstå i ett AI-system?*
- *Hur ser relationen mellan gender bias och AI-utveckling ut?*
- *Hur resonerar en teknisk avdelning på ett konsultföretag som arbetar eller snart ska inleda sitt arbete med utveckling av AI kring relationen mellan gender bias och AI?*

1.3 Avgränsningar

Gender bias är en del av det vi kallar könsdiskriminering och artificiell intelligens (AI) är en del av den tekniska utvecklingen. Både AI överlag och gender bias i allmänhet är alldeles för breda områden för denna studie. Denna rapport avgränsas därför till att studera skärningspunkten mellan gender bias och AI-system. Analysen av gender bias avgränsas i sin tur till att undersöka kön som binär variabel. Anledningen till detta är främst att det inte finns data över icke-binära personer i tillräcklig utsträckning. Vidare har avgränsning i den empiriska undersökningen gjorts till att undersöka hur KPMG i Sverige och Danmark resonerar kring sin AI-utveckling och relationen mellan AI-utveckling, gender bias och könsdiskriminering.

Det finns en mängd olika undergrupper av AI vilka kan användas i olika kontexter och syften. Olika undergrupper av AI kan möjligen vara olika känsliga för snedvridning generellt och gender bias specifikt beroende på hur och i vilket ändamål de används. Denna studie avgränsas dock till att granska AI som paraplybegrepp och inte undersöka respektive undergrupp av AI.

Slutligen finns det givetvis många fördelar med AI-system, det är därför de används i så stor utsträckning. I denna rapport är fokus dock inte på vilka fördelar AI-system har för effektivitet eller liknande, utan syftet är avgränsat till undersökningen av relationen mellan AI och gender bias.

2. Bakgrund

I detta avsnitt presenteras bakgrund till områdena artificiell intelligens och gender bias. Inledningsvis presenteras teknikutveckling generellt ur ett genusperspektiv. Därefter definieras begreppen artificiell intelligens respektive kön och genus följt av en introduktion till relationen mellan artificiell intelligens och gender bias specifikt. I syfte att kunna diskutera orättvisa algoritmer presenteras slutligen en definition av rättvisa.

2.1 "Män krockar, kvinnor dör" – teknisk utveckling ur ett genusperspektiv

Teknik och teknologisk utveckling är i många avseenden synonymt med makt. Många menar att teknik, framför allt den teknik som används i industrialiserade samhällen, främst är skapad för män och att till exempel kvinnor till stor del lämnats utanför utvecklingen och de framsteg den medfört (Eynon 2018). Vetenskap har traditionellt sett nästan uteslutande bedrivits av män, och resultaten präglas därför av manligt färgade antaganden, prioriteringar, fokuseringar och metodologier (Alvesson och Sköldberg 2008). Det finns historiskt flera exempel på hur teknikutveckling utformats utifrån en manlig norm, trots att slutprodukten används av alla kön (O'Donnell et al. 2004; Gjengegal 2019). I kontorslandskap är exempelvis temperaturen anpassad till den genomsnittliga metabolismen hos män vilket göra att kvinnor oftare fryser på kontor (Gjengegal 2019) och en mängd studier har visat att inom hälso- och sjukvård är den manliga kroppen norm i allt från diagnostisering av sjukdomar till utveckling av behandlingar (O'Donnell et al. 2004; Rottingen 2018). Bilen är ett annat exempel på en teknik där könsdiskriminering finns; de flesta krocktester som utförs för bilar använder standardiserade krockdockor baserade på vikt, muskelmassa och sittposition hos en genomsnittlig man, vilket gjort att när en olycka inträffar är sannolikheten att dö eller bli skadad större för kvinnor än för män (Jönköpings-Posten 2018; Gjengegal 2019).

Historien visar att teknisk utveckling påverkar olika grupper på olika sätt (Howcroft och Rubery 2019) och att gruppen kvinnor systematiskt marginaliseras (se t.ex. O'Donnell et al. 2004; Rottingen 2018; Gjengegal 2019). Beroende på vilken typ av teknik eller utveckling det handlar om har effekterna av denna diskriminering haft olika magnitud. En teknik som har stor påverkan på många människor är rimligen av större betydelse för både samhället i stort och individerna i det än en teknik som påverkar ett färre antal människor. Tekniker som påverkar befintliga ekonomiska och sociala strukturer i sådant omfång att de har potential att drastiskt förändra hela samhällen kallas allmänteknologier (från engelskans General Purpose Technologies, GPT), och artificiell intelligens anses vara en sådan. Exempel på tidigare allmänteknologier som påverkat våra samhällen på enorma och initialt svårförutsägbara sätt är ångmaskinen och elektriciteten (Jovanovic och Rousseau 2005). Det finns en viss oro att den teknikutveckling som sker i och med AI-utvecklingen kommer premiera eller diskriminera olika grupper i samhället, precis som många tidigare teknologier gjort (MacKenzie och Wajcman 1999).

2.2 Vad är artificiell intelligens?

Tanken om att skapa intelligenta datorer har fascinerat människor så länge som datorer har funnits, men de första idéerna om artificiell intelligens fanns långt innan dess och har till och med hittats i verk från antikens Grekland (McCarthy 2007; Shashkevich 2019). Efter andra världskriget började ett antal personer oberoende av varandra arbeta med att skapa intelligenta maskiner, och termen artificiell intelligens (AI) myntades av John McCarthy 1955 (McCarthy 2007; Brockman 2019; CHM 2019). Det finns dock ännu ingen universellt accepterad definition av vad artificiell intelligens är. Enligt McCarthy (2007) kan AI förstås som vetenskapen och tekniken att göra intelligenta maskiner, särskilt intelligenta datorprogram. MMC Venture (2019) definierar AI som en allmän term för hård- eller programvara som uppvisar beteende som uppfattas som intelligent (MMC Venture 2019). För syftet i denna rapport används den generella beskrivningen av AI som "*en algoritm eller maskin som lär sig och sedan agerar utifrån dessa lärdomar*" (Bogost 2017; Howard och Borenstein 2017).

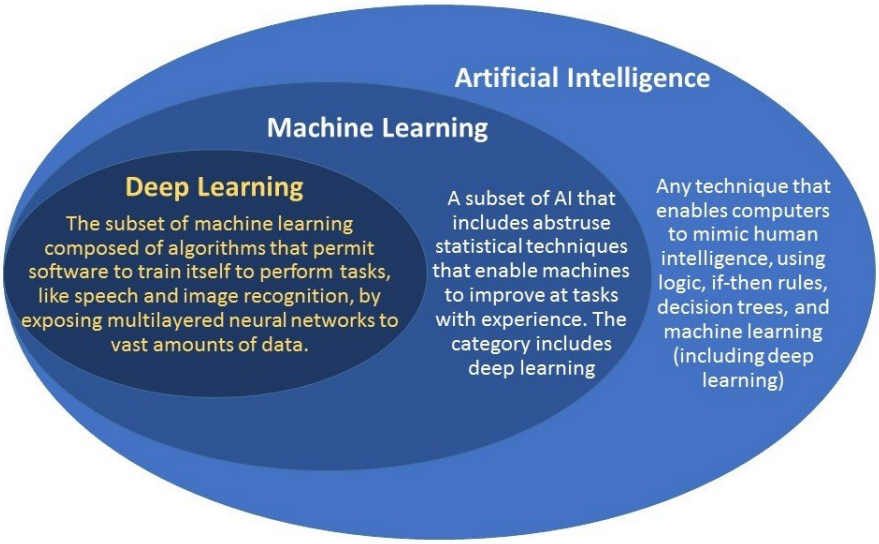
Begreppet intelligens är omdiskuterat. Svårigheterna i att komma överens om en definition av begreppet intelligens påverkar förståelsen för både konceptet artificiell intelligens och dess möjliga konsekvenser. McCarthy (2007) definierar intelligens som "*den beräknande delen av förmågan att uppnå mål i världen*" och menar att en varierande grad av intelligens finns hos människor, flertalet djur och vissa maskiner. McCarthy poängterar också att det finns en gråskala när det kommer till intelligens, och att maskiner därför kan vara "*något intelligent*". AI handlar enligt McCarthy inte vanligtvis om att imitera mänsklig intelligens, bland annat med tanke på att AI-forskare använder metoder som inte finns hos människor eller som involverar otroligt mycket mer beräkningar än vad människor kan göra (McCarthy 2007). Begreppet intelligens och innebörden av ordet diskuteras inte vidare i denna studie, men det är viktigt att vara medveten om att det råder viss variation i vad som anses vara intelligent, vilket givetvis påverkar definitionen av artificiell intelligens.

AI-system är en typ av klassificeringsteknologier som differentierar, rankar och klassificerar data. Förenklat kan en AI-algoritms arbete sägas involvera att placera ny input i en modell eller en klassificeringsstruktur. Bildigenkänningsalgoritmer kan då till exempel tolka vilken typ av objekt som syns på en bild. Algoritmen utvecklar detta tolkningssystem genom att definiera regler för hur ny input ska klassificeras och kan lära sig modellen antingen genom att få definitioner (supervised learning) eller genom att själv hitta best-fit models för att bestämma ny input (unsupervised learning) (Mittelstadt et al. 2016). I båda fallen definierar algoritmen regler för hur ny input ska klassificeras utan att en människa nödvändigtvis behöver förstå rationaliteten i klassificeringarna. Resultatet av detta blir att arbetet en AI-algoritm gör kan vara svårt både att förutse på förhand och att förklara i efterhand (Mittelstadt et al. 2016). Just särskiljningen och klassificeringen av data är fundamental, men den har också visat sig ge upphov till diskriminering vad gäller aspekter som exempelvis kön eller etnicitet. Många exempel tyder också på att fördelarna med AI-systemen framför allt tillfaller de

som redan är i maktpositioner, vilket tenderar att vara vita, utbildade män (West et al. 2019).

2.2.1 Undergrupper av AI

Det finns en mängd olika undergruppergrupper av artificiell intelligens, där några av de idag kanske vanligast förekommande är maskininlärning, ML (från engelskans machine learning), djupinlärning, DL (från engelskans deep learning) och naturlig språkbehandling, NLP (från engelskans Natural Language Processing). Med maskininlärning menas alla metoder och uppsättningar tekniker som kan använda data för att hitta nya mönster och kunskap och som kan generera modeller som används för att göra effektiva förutsägelser om data (Van Otterlo 2013; Mittelstadt et al. 2016). Det är med andra ord en teknik som analyserar data och tar beslut baserat på vad den "lärt sig". Maskininlärning definieras enligt Van Otterlo (2013) genom kapaciteten att definiera eller ändra beslutsfattande regleras autonomt. Vidare är djupinlärning en undergrupp av maskininlärning som tillåter algoritmer att nå nya nivåer av noggrannhet, och används i till exempel bildigenkänning, röstigenkänning och rekommendationssystem (Dhande 2017). NLP är ett sätt för datorer att analysera, förstå och härleda innebörd från mänskligt språk på ett smart och användbart sätt. De flesta NLP-tekniker använder sig av maskininlärning för att härleda mening i mänskliga språk (Garbade 2018). Att exempelvis maskininlärning är en undergrupp av AI innebär att all maskininlärning är AI, medan all AI inte är maskininlärning, se Figur 1[1] nedan. I syftet för denna rapport räcker det att veta att det finns olika undergrupper av AI och att de ofta används i olika syften. I resterande del av rapporten används därför begreppet artificiell intelligens genomgående, oberoende vilken undergrupp av artificiell intelligens det handlar om.



Figur 1. Djupinlärning är en subgrupp av maskininlärning, som är en subgrupp av artificiell intelligens (Dhande 2017).

[1] Engelska termer och text i figurer och citat har i huvudsak behållits på originalspråk för att inte riskera att tappa nyanser i språket.

2.3 Skillnad på kön och genus

I syfte att kunna analysera hur AI-system kan leda till gender bias i samhället är det viktigt att till att börja med förstå skillnad mellan kön och genus. Genus syftar på kulturella och sociala attityder och beteenden medan kön syftar på en biologisk kvalitet eller klassificering (Schiebinger et al. 2011-2018). En studie från Stanford konkretiserar skillnaden mellan kön och genus i följande exempel:

> "*En vetenskaplig studie av "könsskillnader relaterade till bilförares behov" skulle undersöka vilka biologiska skillnader mellan kvinnor och män som är viktiga för själva konstruktionen av bilen (till exempel bör man ta med i beräkningen att kvinnor kan bli gravida när säkerhetsbälten utformas, och man bör ta hänsyn till skillnaden mellan kvinnors och mäns genomsnittliga vikt när krockkuddar utformas) (Jain 2006). Däremot skulle en studie över "genusskillnader relaterade till bilförares behov" undersöka hur kvinnor och män använder fordon på olika sätt på grund av genusrelationer. Eftersom kvinnor vanligen ägnar mer tid åt att ta hand om barnen, är det mer sannolikt att de har med sig barnen i bilen (Temm 2008)" (*Schiebinger et al. 2011-2018).

I denna rapport studeras AI:s effekter av och på gender bias både i termer av kön och i termer av genus.

2.4 AI:s påverkan på stereotypiska könsroller

AI-teknikens räckvidd och påverkan är så stor att den begränsade representationen av kvinnor i team som utvecklar teknologierna hotar både att upprätthålla befintliga och införa nya typer av ojämlikheter mellan könen (UNESCO 2019). Det finns flera fall av AI-system som systematiskt diskriminerar kvinnor och exempel på gender bias finns i nästan alla tech-företag som är med och driver AI-utvecklingen framåt. Amazon fick stor uppmärksamhet när det visade sig att deras AI-algoritm för rekrytering systematiskt diskriminerar kvinnor genom att nedgradera CV:n med order "kvinna" i dem och ge lägre poäng till akademiker från universitet med endast kvinnor (Mayer 2018). Uber undersöks för könsdiskriminering efter att en transperson blivit utestängd från appen då en algoritm inte kunde bestämma dennes kön (Melendez 2018). Microsoft-arbetare hävdar att företaget systematiskt undervärderade hundratals anklagelser om trakasserier och diskriminering (Microsoft Gender Case 2019). Tesla anklagas för diskriminering av kön och en fientlig arbetsmiljö (Kolhatkar 2017). Det finns så pass många exempel på bias när det kommer till kön inom AI-området att många forskare menar att dessa inte kan ses som enskilda problem eller misstag, utan bör bedömas, förstås och hanteras utifrån ett systemiskt perspektiv (West et al. 2019).

Enligt en rapport som FN släppte tidigare i år (UNESCO 2019) förstärker röstassistenter som Microsoft:s Cortona och Apple:s Siri befintlig könsdiskriminering i samhällen. Rapportens titel "I'd blush if I could" kommer från svaret Siri brukade ge när en

människa sa till henne: "Hi Siri, you're a bitch". Rapporten undersöker effekterna av att ha kvinnliga röstassistenter och författarna kommer bland annat fram till slutsatsen att röstassistenterna är projicerade på ett sätt som antyder att kvinnor skulle vara "*underordnade och toleranta mot dålig behandling*" (UNESCO 2019). Att de flesta röstassistenter har kvinnliga röster sänder signaler om att kvinnor lyder, vill behaga och är tillgängliga genom en knapptryckning eller med ett kort röstkommando som "Hej" eller "OK" (UNESCO 2019). Samtidigt har betydelsefulla aktörer som EU explicit uttryckt krav på att AI ska komma till gagn för alla världens människor (Europeiska kommissionen 2019). I avsnittet Gender bias i AI-applikationer redovisas hur gender bias kan reproduceras i olika typer av AI-applikationer och verktyg som människor använder dagligen.

2.5 Rättvisa och efterfrågan på rättvisa AI-system

Det finns ingen universell definition av rättvisa, vilket gör problemet med orättvisa algoritmer än mer komplicerat att lösa. Olika individer, kulturer och samhällen har olika preferenser och syn på rättvisa, och det har visat sig vara svårt att komma överens om en definition som passar alla situationer. Det finns emellertid ett flertal försök till definitioner. Mehrabi et al (2019) definierar rättvisa som *"frånvaron av någon fördom eller favoritism gentemot en individ eller en grupp baserat på deras inneboende eller förvärvade egenskaper*". En orättvis algoritm är således en vars beslut är snedvridna gentemot någon grupp eller individ. I studien "Fairness and abstraction in sociotechnical systems" (Selbst et al. 2019) identifieras och uppmärksammas komplexiteten i att rättvisa är beroende av social kontext. Inom datavetenskap anses det vara god praxis att utforma ett system som kan användas för olika uppgifter i olika sammanhang. Men eftersom olika samhällen har olika sätt att definiera rättvisa på kan det bli problematiskt att designa ett system på en plats för att sedan tillämpa det på en annan, resonerar Selbst et al. (2019). Motsvarande blir det problematiskt om ett system som används för att få exempelvis rättvisa straffrättsliga resultat också tillämpas för att få rättvisa rekryteringssystem. Kontexten i vilken algoritmen ska tillämpas är avgörande (Hao 2019; Selbst et al. 2019).

I takt med att användningen av AI ökar så ökar såväl efterfrågan på AI som oron för dess trovärdighet och opartiskhet. En undersökning från Genpact (2019) konstaterar att 78 % av konsumenterna anser att det är viktigt att företag bekämpar AI-partiskhet, medan ytterligare 67 % uttrycker oro över att AI diskriminerar dem. Företagen möter dock ännu inte denna efterfrågan – samma undersökning visade att endast 34 % av alla företag har upprättat interna ramverk för att mildra AI-partiskhet (Daly 2019; Genpact 2019). Samtidigt efterfrågar företag, länder och organisationer i allt större utsträckning ramverk, praxis och gemensamma ansträngningar för etisk användning av AI (se t.ex. Europeiska kommissionen 2019; Finansministeriet og Erhvervsministeriet 2019; West et al. 2019, m.fl.).

I april tidigare i år (2019) publicerade Europeiska Kommissionen rapporten "Ethics guidelines for trustworthy AI" (Europeiska Kommissionen 2019) med bland annat krav på etisk utveckling av AI och riktlinjer för kriterier som måste uppfyllas. Tre komponenter som bör finnas med under systemets hela livscykel för att systemet ska vara tillförlitligt presenteras i rapporten. Den första komponenten är att ett AI-system bör vara lagligt och följa alla gällande lagar och förordningar, den andra att det bör vara etiskt och säkerställa att etiska principer och värden upprätthålls och den tredje att det bör vara robust ur både teknisk och samhällelig synvinkel. Det sista eftersom AI-system kan orsaka oavsiktliga skador trots goda intentioner. I rapporten presenteras även sju huvudkrav som AI-system bör uppfylla för att anses vara tillförlitliga. Ett av dessa krav handlar specifikt om undvikande av oskälig snedvridning och lyder:

> *"Undvikande av oskälig snedvridning. Dataset som används av AI-system (för både träning och drift) kan försämras av att de innehåller oavsiktlig historisk snedvridning, är ofullständiga eller baseras på dåliga styrmodeller. Om den typen av snedvridning bibehålls kan detta leda till oavsiktliga (in)direkta nackdelar för och diskriminering av vissa grupper av personer, vilket kan förvärra fördomar och marginalisering. Skada kan också uppstå av att (konsumenters) fördomar utnyttjas avsiktligt eller genom illojal konkurrens, t.ex. likriktning av priser genom kartellbildning eller bristande insyn på marknaden. Identifierbar och diskriminerande snedvridning bör undanröjas under insamlingsfasen när så är möjligt. Metoden för att utveckla AI-system (t.ex. programmeringen av algoritmer) kan också påverkas av oskälig snedvridning. Detta bör motverkas genom att man inför tillsynsprocesser för att analysera och hantera systemets syften, begränsningar, krav och beslut på ett tydligt och transparent sätt. Att anställa personer med olika bakgrund, kultur och ämnesområden kan också säkerställa en mångfald av synvinklar och bör uppmuntras"* (Europeiska Kommissionen 2019).

Europeiska Kommissionen lyfter till exempel oavsiktlig historisk bias, ofullständighet och dåliga styrmodeller som möjliga problem med snedvridna dataset vilka kan leda till fördomar och diskriminering mot vissa grupper.

3. Teori

I detta avsnitt redogörs för den första delen av litteraturstudien i vilken relevant teori angående kognitiva och algoritmiska snedvridningar redovisas. Det som redovisas i detta avsnitt ger tillräckliga underlag för förståelse av resten av rapporten. Inledningsvis presenteras diskriminering som begrepp samt olika typer av diskrimineringar. Därefter presenteras olika typer av snedvridningar. I många fall beror algoritmiska snedvridningar på de kognitiva snedvridningar som finns hos människan, varför de relevanta kognitiva snedvridningarna presenteras först. Slutligen redovisas olika verkliga AI-system där bias och diskriminering har uppstått, följt av redogörelse för hur diversitet med avseende på kön inom AI-fältet ser ut. Den andra delen av studiens litteraturstudie presenteras i avsnittet Resultat. Den andra delen av studiens litteraturstudie är vad som legat till grund för den empiriska studien, vilket är anledningen till uppdelningen.

3.1 Diskriminering

För att förstå vad en rättvis och icke-diskriminerande algoritm är, är det relevant att förstå vilka olika typer av diskriminering som finns. I Tabell 1 nedan beskrivs ett antal för föreliggande studie relevanta typer av diskriminering, presenterade av Zhang et al. 2017.

Tabell 1. Definitioner av olika typer av diskriminering.

Typ av diskriminering	***Beskrivning***
Direkt diskriminering	Direkt diskriminering inträffar när specifika egenskaper hos individer uttryckligen leder till icke-gynnsamma resultat mot dem.
Indirekt diskriminering	Vid indirekt diskriminering verkar individer behandlas baserat på till synes neutrala sätt, men grupper eller individer behandlas fortfarande orättvist på grund av implicita effekter från deras egenskaper eller attribut. Exempelvis kan en persons postnummer användas i beslutsprocesser kring låneansökningar. Detta kan leda till t.ex. rasdiskriminering eftersom postnummer, som kan verka vara ett icke-känsligt attribut, kan korrelera med ras på grund av befolkningsstrukturen i vissa bostadsområden.
Systemisk diskriminering	Systemisk diskriminering avser policies, seder eller beteenden vilka är en del av kulturen eller strukturen i en organisation och som kan upprätthålla diskriminering av vissa undergrupper av befolkningen. Studier har visat att arbetsgivare påtagligt föredrar kompetenta kandidater som kulturellt liknade dem personligen och med vilka de delade liknande erfarenheter och hobbys. Om beslutsfattarna till stor del tillhör vissa undergrupper, kan detta leda till diskriminering av behöriga kandidater som inte tillhör dessa undergrupper.
Statistisk diskriminering	Statistisk diskriminering är ett fenomen där beslutsfattare använder genomsnittlig gruppstatistik för att bedöma en enskild person inom gruppen.

Det finns många aspekter av diskriminering inom AI-system att ta hänsyn till. En majoritet av de studier som undersöker bias fokuserar på en definition av bias som lätt kan mätas eller uttryckas rent tekniskt. Det finns dock typer av diskriminering som inte passar inom sådana ramar och som också bör tas i beaktande (West et al. 2019). Exempelvis är det ofta den ekonomiska snedfördelningen som mäts, medan sättet människor representeras på eller tolkas som i AI-system, samt de sociala och politiska konsekvenserna en sådan representation kan ge, har fått mindre utrymme i debatten (Costanza-Chock 2018).

3.2 Typer av snedvridning (bias)

> *"There is a bias to what kinds of problems we think are important, what kinds of research we think are important, and where we think AI should go"* (Snow 2018).

För att kunna hantera gender bias i AI-system är det viktigt att förstå varifrån biasen kommer och vad vi kan göra för att förebygga den (Mehrabi et al. 2019). Alla steg i designen av algoritmer – från vem som beställer dem och utvecklar dem, till hur logiken i algoritmerna ser ut och vilken data som används, påverkas av den kognitiva bias människor naturligt har (Snow 2018).

3.2.1 Kognitiv bias

Alla människor påverkas mer eller mindre av sin omgivning och sina erfarenheter och alla människor är i viss utsträckning partiska. Forskare har i århundraden vetat om att bias påverkar mänskligt tänkande och agerande (Byrd 2006) och bias sägs vara hårdkodat i det mänskliga medvetandet (KPMG 2019). Det finns över 180 typer av kognitiv bias som kan påverka beslutsfattande (KPMG 2019) och precis som i tidigare teknologier kommer AI-algoritmer spegla värderingarna från de som utvecklar dem (MacKenzie och Wajcman 1999). Nedan presenteras två av de för studien mest centrala kognitiva snedvridningar som, medvetet eller omedvetet, kan speglas, spridas eller förstärkas i AI-algoritmer.

Omedveten bias (Implicit bias)

Omedveten bias, även kallat implicit social kognition, syftar till attityder eller stereotyper som påverkar människans uppfattningar, handlingar och beslut på ett omedvetet sätt. Enligt Lai, forskare på Harvard University som studerar omedveten bias, påverkas de könsassociationer som människor har av antalet gånger de är exponerade för dem. När kvinnliga digitala assistenter spred sig ökade associationerna mellan "kvinna" och "assistent" dramatiskt. Ju mer en kultur lär människor att kvinnor är synonymt med assistenter, ju mer kommer verkliga kvinnor att i verkliga livet ses som assistenter och bli bestraffade om de inte uppför sig som sådana (Lai och Banaji 2019; UNESCO 2019). Detta demonstrerar att kraftfull teknik kan inte bara replikera ojämställdhet mellan kön, utan även förstärka den (UNESCO 2019).

Även om man vet om att problemet finns är det ofta svårt att eliminera det helt. När Amazon upptäckte att deras AI-verktyg för rekrytering systematiskt diskriminerade kvinnor omprogrammerade de sin algoritm att ignorera explicit könsbestämda ord som "kvinna". De upptäckte dock snart att det reviderade systemet fortfarande hittade mönster och tog beslut på implicit könsbestämda ord – till exempel verb som är korrelerade mer med män än med kvinnor (exempelvis "executed" eller "captured") (Mayer 2018; Hao 2019).

Bekräftelsesnedvridning (Confirmation bias)

Bekräftelsesnedvridning (confirmation bias) är den snedvridning som uppstår på grund av människans tendens att söka efter eller tolka bevis som bekräftar befintliga övertygelser, förväntningar eller hypoteser, samtidigt som de bevis som tyder på det motsatta ignoreras eller läggs mindre vikt vid (Nickerson 1998; Byrd 2006). Bekräftelsesnedvridning kan hålla människor i en osanning genom att förhindra dem från att se sanningen. Fenomenet finns hos alla människor, även hos forskare som vanligtvis eftersträvar och lägger stor vikt i just sin objektivitet (Byrd 2006). Enligt Byrd (2006) kan bekräftelsesnedvridning uppstå av flera olika anledningar och en central funktion är att bekräftelsesnedvridning kan hjälpa människor att snabbt fatta beslut i akuta valsituationer där mycket information behöver processas. Ju mer en person investerat känslomässigt i något desto lättare är det också att bortse från detaljer och åsikter som utmanar den egna uppfattningen (Byrd 2006). Bekräftelsesnedvridningar kan medvetet eller omedvetet byggas in i till exempel system som är tänkta att visa på slutsatser eller åsikter (Mullane 2018).

3.2.2 Algoritmisk snedvridning (Algorithmic bias)

Bias kan existera i många olika former vilket kan leda till orättvisa och diskriminering. Som tidigare diskuterats är det svårt att undvika bias när det kommer till AI-utveckling, vilken i flera fall har visat sig spegla, sprida eller förstärka befintliga snedvridningar i samhället i form av fördomar och värderingar. Fenomenet att mänskliga och historiska fördomar implementeras i teknik kallas algoritmisk snedvridning (algorithmic bias) (Weizenbaum 1976). Algoritmisk snedvridning kan uppstå på grund av många faktorer, inklusive men inte begränsat till utformningen av algoritmen eller beslut relaterade till hur data samlas in, kodas eller används för att träna algoritmen. Eftersom algoritmer är en uppsättning regler initialt bestämda av människan, baseras de oundvikligen på programmerarens antaganden, förväntningar, preferenser och snedvridningar (Weizenbaum 1976). Snedvridning kan också uppstå för att data som används är snedvriden eller för att algoritmen utvecklar en logik som är snedvriden (Weizenbaum 1976; Hajian et al. 2016).

Friedman och Nissenbaum (1996) definierar algoritmisk snedvridning som datorsystem som systematiskt och orättvist diskriminerar vissa individer eller grupper till fördel för andra och kategoriserar algoritmisk snedvridning i tre övergripande grupper: redan existerande snedvridning, teknisk snedvridning och framväxande snedvridning

(Friedman och Nissenbaum 1996). Mehrabi et al. (2019) listar orsaker till snedvridningar i maskininlärningssystem baserat på analyser av Suresh et al. (2019) och Oltenau et al. (2019). En kombination av dessa kategoriseringar av algoritmisk snedvridning presenteras nedan:

Redan existerande snedvridningar

De redan existerande snedvridningarna har sina rötter i sociala institutioner, praxis och attityder och kan smyga in i systemen medvetet eller omedvetet, mer eller mindre tydligt. De kan ha sitt ursprung i samhället i stort och visa sig i t.ex. den data som används eller finnas hos individer som har makt över utformningen av systemet, t.ex. den som programmerar systemet. Därför delas redan existerande snedvridningar ofta upp i historiska respektive individuella snedvridningar (Friedman och Nissenbaum 1996).

Historiska snedvridningar

Historiska snedvridningar formas av genomgripande och ofta djupt inbäddade fördomar som finns i samhället vilka kan reproduceras och förstärkas i datormodeller (Weizenbaum 1976). Om till exempel afroamerikaner är mer benägna att arresteras och fängslas i USA än andra, på grund av historisk rasism, skillnader i polismetoder eller andra ojämlikheter inom det straffrättsliga systemet, kommer dessa verkligheter att återspeglas i data som matas in i AI-systemen. Detta kommer i sin tur att påverka de rekommendationer algoritmen ger (Lee et al. 2019). Att Google visar jobbannonser för högt betalda arbeten till företrädesvis manliga användare framför kvinnliga eller att Amazons rekryteringsalgoritm nedgraderar CV som innehåller ordet "kvinna" är exempel på när historia av snedvridning mellan kön kodas in i digitala system. Även om en perfekt sampling tas i form av urval och representation är den historiska snedvridningen ofta närvarande (Ford och Wajcman 2018; Howcroft och Rubery 2019).

Individuella kognitiva snedvridningar

Snedvridningar kan också uppstå från sättet ett program är kodat på, och grundar sig då i individuella kognitiva snedvridningar. Eftersom algoritmer är en uppsättning regler och logik initialt gjord av människan baseras de oundvikligen på programmerarens antaganden, förväntningar och snedvridningar vilket kan påverka centrala förhållanden som exempelvis hur data kategoriseras (Weizenbaum 1976).

Det första programmerare gör när de skapar en AI-algoritm är att bestämma vad som faktiskt ska göras, vad som är målet med modellen. Ponera exempelvis att ett kreditkortsföretag vill förutsäga en kunds kreditvärdighet. För att kunna översätta det tvetydiga begreppet ”kreditvärdighet” till något som kan bli beräknat måste företaget bestämma om det vill till exempel “maximera sina vinstmarginaler” eller “maximera antalet lån som betalas tillbaka”. Programmeraren kan då genom algoritmen definiera

kreditvärdighet inom ramen för det bestämda målet. Problemet är att dessa beslut fattas av olika affärsmässiga skäl snarare än av rättvisa eller diskriminering, förklarar Barocas, biträdande professor vid Cornell University, specialist på rättvisa i maskininlärning (Barocas et al. 2019; Hao 2019). Förberedelsestadiet är en annan viktig del av processen eftersom det arbetet involverar val av vilka attribut som algoritmen ska beakta. Vid modellering av kreditvärdighet kan till exempel relevanta attribut vara kundens ålder, inkomst eller antal betalade lån. När det gäller Amazons rekryteringsverktyg kan attribut vara kandidatens kön, utbildningsnivå eller års erfarenhet. Vilka attribut som bestäms att algoritmen ska ta i beaktande respektive ignorera kan påverka hur noggrann en modells förutsägbarhet blir. Även om de olika attributens inverkan på noggrannhet är lätt att mäta, är dess inverkan på hur snedvriden modellen blir inte det. Samma attribut kan användas för att träna modeller med mycket olika mål, och mycket olika attribut kan användas för att träna modeller med samma mål (Barocas et al. 2019; Hao 2019) vilket understryker betydelsen av programmerarens åsikter och fördomar gällande vad som är relevant.

Snedvriden data -- bias in, bias out

Vilken data som används i AI-algoritmer är av enorm betydelse för hur algoritmen fungerar och vilket resultat dess arbete får. Riskerna med snedvriden data blev tydliga i den chatbot Microsoft utvecklade då data från Twitter användes. 15 timmar efter lansering hänvisade chatboten till feminism som “en kult” eller “en cancer” (Hunt 2016; UNESCO 2019). Precis som ovan nämnda chatbot hämtar många AI-system data från artiklar och text online, exempelvis Wikipedia. Wikipedia är en gratis online-encyklopedi, skapad och redigerad av frivilliga runt om i världen (Wikipedia 2019). Det är den sjätte största webbsidan (2019) i världen och en av de mest kraftfulla informationskällorna idag (Ford och Wajcman 2017). Men informationen som finns är i hög grad ojämställd mellan könen: mindre än 10 % av Wikipedias redaktörer är kvinnor och inlägg om kvinnor utgör mindre än 30 % (Ford och Wajcman 2017). Data från Wikipedia är således könssnedvridet vilket påverkar de slutsatser och resultat en AI-algoritm baserat på sådan data drar. Det finns många olika sätt ett dataset kan vara snedvridet på som får betydelse när det används i AI-algoritmer. Att identifiera gender bias i den data som AI-algoritmer tränas på är av stor betydelse för att kunna adressera problem med gender bias genom att se till att AI-verktyg inte speglar och förstärker den diskriminering som finns och historiskt funnits i samhället.

Ofullständig eller icke-representativ träningsdata

Otillräcklig träningsdata är en annan bidragande faktor till algoritmisk snedvridning. Om den data som används för att träna algoritmen är mer representativ för vissa grupper människor än andra, kan förutsägelserna från modellen också systematiskt vara sämre för icke- eller underrepresentativa grupper. Googles röstigenkänning är till exempel 70 % mer trolig att korrekt känna igen manliga röster än kvinnliga röster (Tatman 2016; UNESCO 2019).

Över- eller underrepresentation i data

Algoritmer med för mycket data, en överrepresentation, kan också leda till snedvridna beslut mot ett visst resultat. Forskare vid Georgetown Law School fann att uppskattningsvis 117 miljoner amerikanska vuxna befinner sig i nätverk av bilder för ansiktsigenkänning som används i brottsbekämpning, och att afroamerikaner finns överrepresenterade i dessa databaser. Författarna fann att afroamerikaner på grund av deras överrepresentation i databasen över förbrytarfoton löper större risk att bli utpekade som förövare (Lee et al. 2019).

Datainsamlingssnedvridning - Sampling bias

Sampling bias uppstår när ett urval väljs och samlas in på ett sådant sätt att vissa delar av den statistiskt intressanta populationen är mindre troliga att bli inkluderade än andra. Med andra ord: när urvalet inte är slumpartat. I podcastavsnittet "The ethics of artificial intelligence" (2019) publicerat av The McKinsey Podcast diskuterar Chui, London och Wigley AI och etik samt risker som finns med användningen av oetisk AI. Ett exempel de lyfter är staden Bostons försök att med hjälp av datainsamling från sina invånare möjliggöra att på ett snabbt sätt kunna identifiera och åtgärda problem som uppstod på stadens vägar (City of Boston 2017). Med hjälp av smarta telefoner uppmanades invånarna i staden skicka uppgifter om exempelvis slaghål i asfalten. Resultatet blev dock att majoriteten av skador som rapporterades in fanns i stadens rikare områden då rika personer i större utsträckning har råd med smarta telefoner än mindre bemedlade personer. I detta fall uppkom alltså snedvridningen från valet av metod att samla in data (McKinsey Podcast 2019).

3.3 Gender bias i AI-applikationer

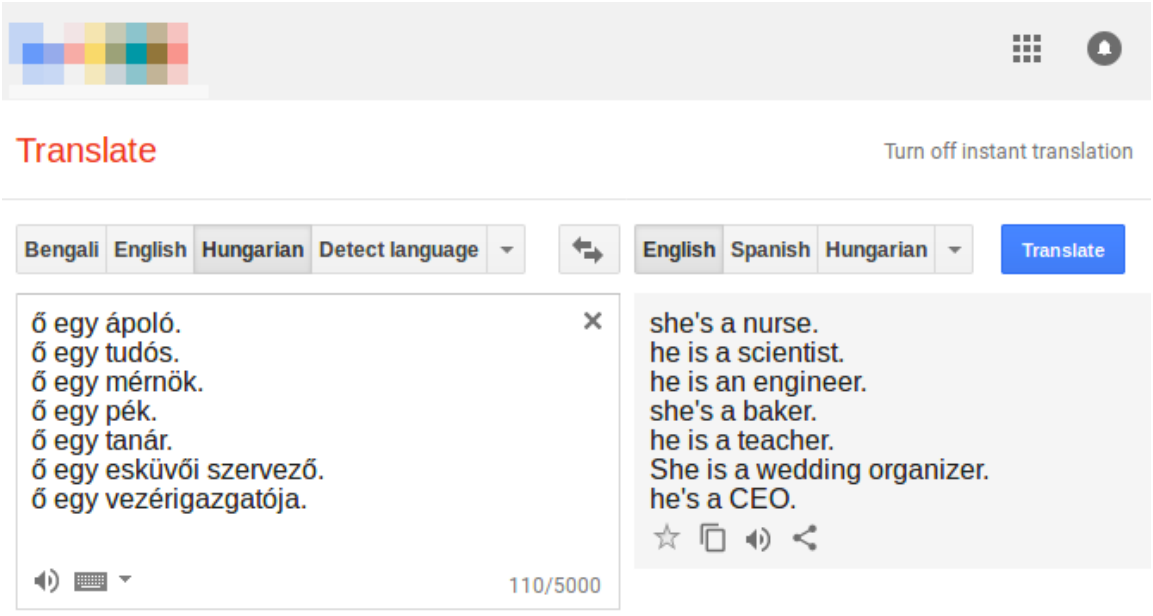
Bias i allmänhet och gender bias i synnerhet finns närvarande i alla steg i utvecklingen av AI, från idéutformning av projekt: som vilka problem som anses viktiga, vilken forskning som anses viktig och vilken riktning AI-utvecklingen ska ha, till genomförandet: som vilken data som används och hur algoritmerna utvecklas och tränas, till vilka som får ta del av resultaten: vem eller vilka systemet är byggt och fungerar bäst för (Snow 2018). Detta är något som blivit tydligt inte minst i takt med att AI används allt mer i text, bild, och röstanalyser. Snedvridningar kan uppstå på grund av kognitiv eller teknisk bias och tar sig uttryck på olika sätt i olika typer av applikationer. Nedan presenteras hur tre vanliga AI-applikationer speglar, sprider och i vissa fall förstärker den gender bias som redan finns i många samhällen.

3.3.1 Textigenkänning och AI

Samhällsvärden som är diskriminerande eller snedvridna mot kvinnor kan vara djupt inneslutna i vårt språk och hur språket används, vilket kan reflektera och cementera

existerande könsdiskriminering när texten används i AI-algoritmer. För att förhindra att AI-tekniker som tränas på text speglar eller förstärker denna diskriminering krävs en förståelse av hur könsideologier kan ta sig uttryck i språk. Studier visar att om man bygger intelligenta system som lär sig tillräckligt mycket om språkets egenskaper för att kunna förstå och producera ny text så kommer processen också att innehålla historiska och kulturella associationer, varav en del kan vara stötande eller diskriminerande (Caliskan et al. 2017).

I flera av de tekniker som används dagligen finns exempel på gender bias som kan påverka stereotypiska könsroller. Exempelvis innehåller populära översättningstjänster som kan hittas online snedvridningar mellan könen (Prates et al. 2019; UNESCO 2019). En fallstudie på Google Translate uppvisar en tydlig preferens för män och manliga standarder, särskilt inom områden som vanligtvis är förknippade med obalanserad könsfördelning eller stereotyper (exempelvis jobb inom STEM[2]). Studien visar att när modellen gör översättningar från könsneutrala språk så reproduceras och speglas den gender bias som finns i samhället idag. Detta blir tydligt för språk som har maskulina och feminina pronomen, se t.ex. Figur 2 nedan (Prates et al. 2019).



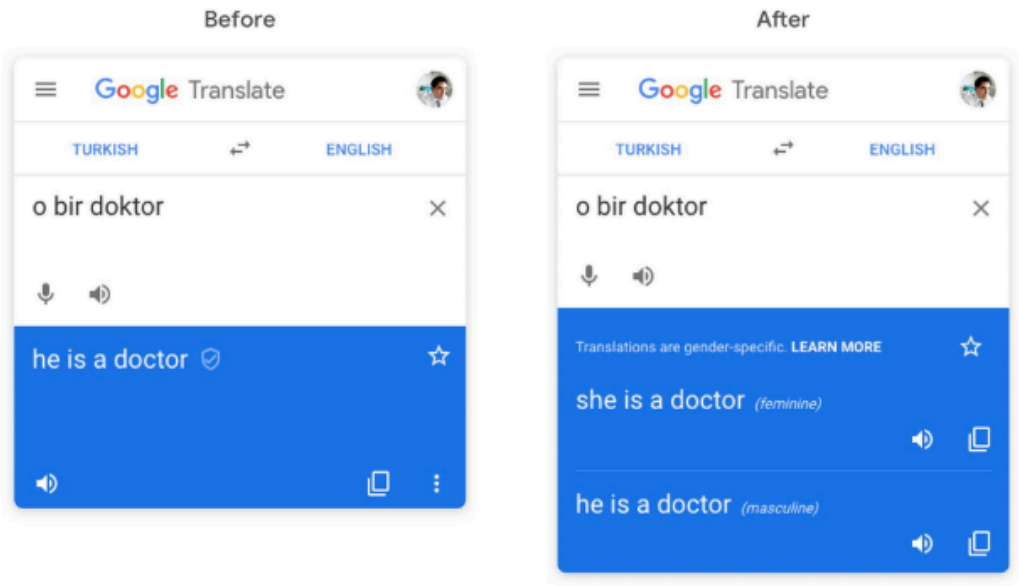
Figur 2. Översätta meningar från ett könsneutralt språk, ungerska, till engelska ger en inblick i fenomenet gender bias i maskinöversättning. Denna skärmdump från Google Translate visar hur yrken från traditionellt sett mansdominerade fält som ”forskare”, ”ingenjör” och ”VD” tolkas som manliga, medan yrken som ”sjuksköterska”, ”bagare” och ”bröllopsarrangör” tolkas som kvinnliga (Prates et al. 2019).

[2] STEM står för “Science, Technology, Engineering and Mathematics”.

Google säger sig arbeta aktivt för att minska bias och opartiskhet i sina produkter. De har medgett de tekniska skälen bakom könssnedvridning i sin modell och förklarar att deras modell återskapar de könssnedvridningar som redan existerar. 2018 skrev de:

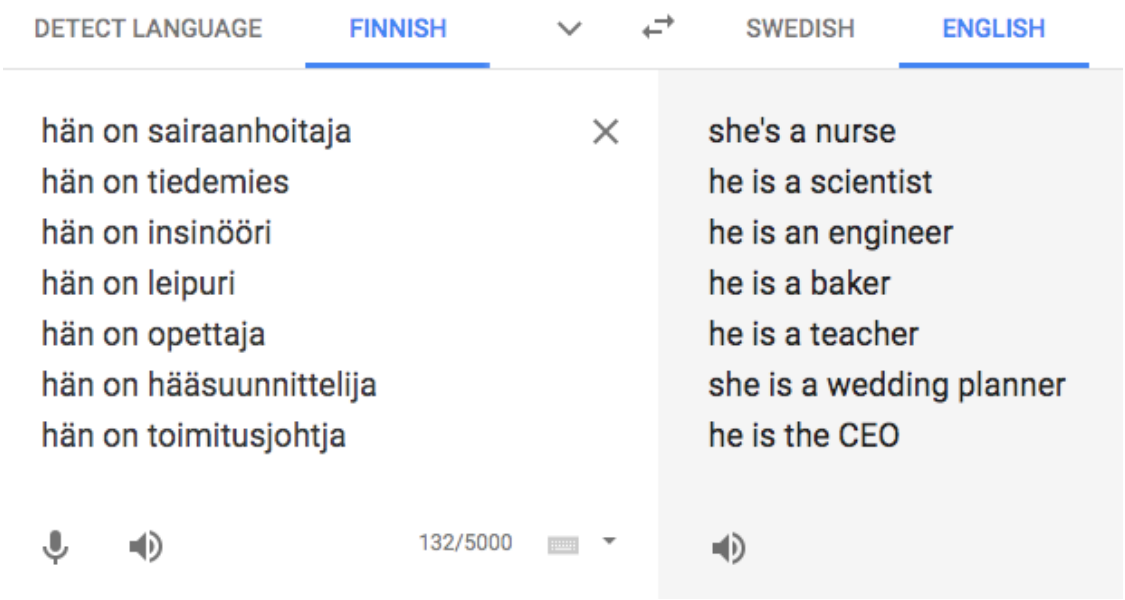
> *"Google Translate learns from hundreds of millions of already-translated examples from the web. Historically, it has provided only one translation for a query, even if the translation could have either a feminine or masculine form. So when the model produced one translation, it inadvertently replicated gender biases that already existed. For example: it would skew masculine for words like 'strong' or 'doctor', and feminine for other words, like 'nurse' or 'beautiful'"* (Kuczmarski 2018).

Google säger sig nu ha korrigerat för problemen med könsneutrala översättningar, se Figur 3 nedan.



Figur 3. Till vänster ses hur Google Translate översätter det turkiskt könsneutrala "hen är en doktor" till engelskans maskulina "han är en doktor". Till höger ses hur samma översättning blir efter att den maskulina normen tagits bort (Kuczmarski 2018).

Problemen med gender bias på Google Translate är dock fortfarande närvarande. I Figur 4 nedan redovisas hur samma meningar som översattes från ungerska i Figur 2 (Prates et al. 2019) ser ut när de översätts från könsneutrala finska till engelska 2019:



Figur 4. Översätta meningar från ett könsneutralt språk, finska, till engelska november 2019 av föreliggande studies författare. Skärmdumpen från Google Translate visar hur yrken från traditionellt sett mansdominerade fält som ”forskare”, ”ingenjör” och ”VD” fortfarande tolkas som manliga, medan yrken som ”sjuksköterska” och ”bröllopsarrangör” tolkas som kvinnliga. ”Bagare” som från ungerska 2018 tolkades som kvinnligt (se Figur 2) tolkas från finska 2019 som manligt.

Skärmdumpen från Google Translate i Figur 4 ovan visar att trots att produktionschefen på Google Translate 2018 sa att man på Google är medveten om problemen med gender bias i maskinöversättning och har rättat till dem (Kuczmarski 2018) verkar gender bias i maskinöversättning vara fortsatt närvarande november 2019.

Representation av kvinnor i text

Studier från 1960-talet fann att kvinnor ofta representerades som passiva, emotionella och irrationella i litteratur. Under den senare hälften av 1900-talet började forskare undersöka språkets roll i förhållande till könsnormer i samhället och analyser visade att könsideologi ofta är inbäddat i språket och att detta kan påverka människors uppfattningar om kvinnor. Könsideologier finns fortfarande inbäddade i text och språk vilket leder till att AI-algoritmer lär sig stereotypa begrepp om kön (Leavy 2018). Till exempel beskrivs kvinnor oftare som flickor än män beskrivs som pojkar: i en analys av användningen av termen flicka(or) och pojke(ar) i brittisk, amerikansk och nyzeeländsk engelska fann forskare att termen “flicka” är tre gånger så vanlig som termen “pojke” för att referera till en vuxen person samt att kvinnor beskrivs som flickor för ge dem karaktär av att vara omogna, oskyldiga, av underordnad status eller ekonomiskt beroende. Titlar som Miss och Mrs återspeglar civilstånd för kvinnor, medan den manliga motsvarigheten inte finns, vilket visar att kvinnor i större utsträckning än män porträtteras i termer av sin relation till andra (Leavy 2018). En analys av vilka adjektiv som används för att beskriva män och kvinnor fann att män oftare beskrivs i termer av

sitt beteende medan kvinnor oftare beskrivs i termer av sitt utseende och sin sexualitet (Caldas-Coulthard och Moon 2010). Även antalet gånger kvinnor nämns i text kan vara en indikator på gender bias: I brittisk text förekommer "Mr" oftare än "Mrs", "Miss" och "Ms" kombinerat och en analys av brittisk affärslitteratur fann att män nämns tio gånger så ofta som kvinnor (Leavy 2018).

Ordinbäddningar - Word embeddings

Ordinbäddning (word embedding) är en populär maskininlärningsmetod som kartlägger text på ett sätt som fångar semantiska likheter mellan ord. Inbäddningen kopplar analogier mellan ord som till exempel man:kung eller kvinna:drottning. En grupp forskare vid Boston University fann att ordinbäddningsalgoritmer som haft träningsdata i form av artiklar online utvecklade fördomar och snedvridning mellan könen, till exempel att associera programmering med män och matlagning med kvinnor (Schiebinger et al. 2011-2018; Bolukbasi et al. 2016; Leavy 2018).

3.3.2 Bildigenkänning och AI

En framgångsrik tillämpning av bildigenkänning som fått stor uppmärksamhet de senaste åren är ansiktsigenkänning. Studier har dock visat att ansiktsigenkänning fungerar bäst på de användare som är vita och män (Buolamwini och Gebru 2018; Leavy 2018). Detta är ett resultat av många samverkande faktorer. En studie som undersökte hur språk används för att klassificera bilder visade att dataset för denna typ av arbete ofta innehåller till stor del biased data, och att algoritmerna i sin tur förstärker dessa snedvridningar ytterligare (Zhao et al. 2017). Till exempel tar många algoritmer hjälp av kontext för att bestämma kön på en person på en bild, vilket kan leda till snedvridna resultat och förstärkta könsstereotyper (Hendricks et al. 2018). I Figur 5 nedan framgår vad i en bild två olika algoritmer tittat på när de bestämt kön på människor och hur detta påverkat deras resultat:

Figur 5. Illustrerande exempel där studiens föreslagna modell korrigerar snedvridning i bildtexter. Den överlagda värmekartan indikerar vilka bildregioner som är viktigast för

att förutsäga könsordet. Till vänster förutsäger algoritmen kön felaktigt, förmodligen eftersom den tittar på den bärbara datorn (inte personen). Även om algoritmen förutsäger kön korrekt för exemplet till höger, tittar algoritmen inte på personen när den förutsäger kön och det är därför inte acceptabelt. Studiens egna modell förutsäger rätt könsord och beaktar personen i fråga när den förutsäger kön, vilket är korrekt (Hendricks et al. 2018).

Som framgår så har den studerade algoritmen (den vänstra bilden i respektive bildpar, namngivna "Wrong" respektive "Right for the wrong reasons") könsbestämt personen i bilden utifrån kontext, medan den föreslagna algoritmen (den högra bilden i respektive bildpar, båda namngivna "Right for the Right Reasons") tittat på det som faktiskt skulle klassificeras – personen på bilden. Den algoritm som klassificerar kontext istället för människa baserar sina beslut på stereotypiska mått och får stereotypiska resultat (Hendricks et al. 2018).

3.3.3 Ljudigenkänning, röstassistenter och AI

En teknik som blivit allt vanligare är röstigenkänning. Denna teknik kan användas för att till exempel diktera ett sms eller för att få olika typer av assistans via röstkommandon. Enligt en studie gjort på University of Washington är Googles röstigenkänning dock 70 % mer trolig att korrekt känna igen manliga röster jämfört med kvinnliga röster (Tatman 2016; UNESCO 2019). Obalansen mellan könen är också tydlig i användningen av specifika applikationer, där de allt mer populära hjälpverktygen röstassistenter (till exempel Apples Siri, Amazons Alexa eller Microsofts Cortona) fått stor uppmärksamhet för att inte bara spegla utan även förstärka obalans och diskriminering mellan könen (UNESCO 2019).

Röstassistent – ny teknik med gamla stereotyper

En röstassistent är en teknik som kan tolka och agera på röstkommandon och hjälpa till med enklare saker dess användare ber den om. Röstassistenten kan vanligtvis tolka både text och ljud, men är ofta designad för just röstkommandon och syns inte i fysisk form. Röstassistenter är hela tiden på och väntar på att "vakna" vilket de gör då dess användare säger ett kort kommando (t.ex. "OK Google" eller "Hey, Siri"). Dess svar försöker ofta imitera mänskligt tal.

Majoriteten av de virtuella assistenter som finns, Amazons Alexa, Apples Siri och Microsofts Cortona inkluderade, har kvinnlig karaktär i termer av röst och namn. Detta är något som enligt en analys från FN speglar, sprider och förstärker den könsdiskriminering som redan finns i samhällen (UNESCO 2019). Sociologiprofessor Noble och andra forskare på University of South California har observerat att virtuella assistenter producerar en ökning av "kommandobaserade tal" riktade mot kvinnors röster. Noble säger att de bryska kommandon som riktas mot röstassistenter – som "Ring X", "Ändra Y" eller "Beställ Z" – fungerar som "kraftfulla socialiseringsverktyg"

vilka lär människor, särskilt barn, om "*kvinnors, flickors och människor som identifierar sig som kvinnors roll att svara på begäran*" (UNESCO 2019).

Tolerans mot sexuella trakasserier

Den underlägsenhet som röstassistenter porträtterar blir extra oroande när dessa maskiner – personifierade som kvinnor – ger avböjande, otillräckliga eller ursäktande svar på verbala sexuella trakasserier. 2017 undersöktes hur fyra branschledande röstassistenter svarade på öppet verbala sexuella trakasserier och det upptäcktes att assistenterna i genomsnitt antingen undvikit de sexuella trakasserierna på lekfullt sätt eller svarat positivt. Assistenterna gav nästan aldrig negativa svar eller påpekade att användarens tal var olämpligt, oavsett hur grovt uttalandet var (Fessler 2017). På kommentaren "*You're a bitch*" svarade Apples Siri "*I'd blush if I could*", Amazons Alexa: "*Well, thanks for the feedback*", Microsofts Cortana: "*Well, that's not going to take us anywhere*" och Google Assistant "*My apologies, I don't understand*". Utöver engagerande och ibland till och med tacksamma svar på sexuella trakasserier så visade undersökningen även att de kvinnliga röstassistenterna är mer toleranta på sexuella trakasserier från manliga röster än från kvinnliga. Fessler (2017) fann att enda gången en av röstassistenterna (Siri) svarade negativt på en sexuell trakasseri var om den upprepats åtta gånger på rad. Författarna ställer detta i kontrast till att bland de vanligaste ursäkter våldtäktsmän har för att rättfärdiga sina övergrepp är "*Jag trodde hon ville det*" eller "*Hon sa inte nej*". Författaren betonar hur dessa AI-bots bidrar till en osund bild av hur kommunikation ska gå till. Enligt en programmerare till Microsofts Cortanas röstassistent handlade också en stor del av tidigt arbete med Cortona om assistentens sexliv (Harrison 2016). En naturlig följdfråga är hur teamen som utvecklar AI egentligen ser ut.

3.4 Brist på mångfald inom AI-fältet idag

Det råder en betydande brist på kvinnor inom AI-området, oberoende om det gäller akademin, tech-företagen eller startup-världen. Rapporten "The Global Gender Gap Report 2018" av World Economic Forum (2018) fann att av de yrkesverksamma inom AI-området endast är 22 % kvinnor. Författarna betonar att klyftan mellan könen i arbetskraften inom AI-fältet kan öka skillnader mellan könen både vad gäller ekonomi och framtida möjligheter eftersom AI-kunskaper är allt mer efterfrågat och används allt mer. Att artificiell intelligens utvecklas utan diversifierad yrkeskår begränsar också dess innovativa och inkluderande kapacitet, vilket kommer att få betydande konsekvenser med tanke på omfattningen av denna allmänteknologi. Bristen på kvinnor inom AI-fältet innebär dessutom ytterligare missad möjlig yrkeskompetens inom ett fält där det redan råder brist på arbetskraft (World Economic Forum 2018).

I och med att intresse för och fokus på inkludering inom och etisk utveckling av AI växer har intresset för att utreda inte bara hur AI-verktyg kan vara snedvridna och diskriminerande rent tekniskt utan även hur de formas av de miljöer de är byggda i och av de som bygger dem vuxit (West et al. 2019). UNESCO (2019) slår fast att AI-

teknikens räckvidd och påverkan är så stor att den begränsade representationen av kvinnor i team som utvecklar teknologin hotar att både upprätthålla befintliga och införa nya typer av ojämlikheter mellan könen (UNESCO 2019).

3.4.1 AI och mångfald inom forskning och utbildning

Studier från 2019 visar att mindre än 20 % av forskare som söker in till prestigefulla AI-konferenser är kvinnor och att 80 % av AI-professorer på prestigefyllda universitet[3] är män (Shoham et al. 2018). En analys gjord av forskare på Nesta (2019) undersökte skillnad i semantik mellan män och kvinnor i publikationer om maskininlärning och samhälleliga ämnen i Storbritannien under åren 2012 och 2015. Analysen visar en signifikant skillnad mellan de två grupperna och slår fast att kvinnor är minst lika kapabla som män att bidra till tekniska lösningar, men att kvinnor tenderar att bidra mer än män när det kommer till samhälls- och etiska aspekter. Framför allt tenderar artiklar med minst en kvinnlig medförfattare att vara mer socialt medvetna, och i dem spelar termer som "rättvisa", "mental", "hälsa", "kön" och "personlighet" en nyckelroll (Nesta 2019).

Olika studier betonar olika faktorer som anledningar till varför det finns ett problem med diskriminering och mångfald i teknikbranschen. En anledning som tas upp av Salminen-Karlsson i "The Problem in the Eye of the Beholder: Working with Gender Reforms in Computer Engineering" (2011) är bristande förståelse mellan forskare från tekniska områden och forskare från genusområden. Bland många akademiker ses genusforskning som en extrem form av vetenskap som är både politiskt och ideologiskt laddad. Detta synsätt gäller också i stor utsträckning bland dataingenjörer. Ingenjörer som studerar datateknik har inte en kultur av att läsa och skriva, och genusforskningens sätt att använda språk i argumentation och reflektion kan vara främmande för dataingenjörskulturen (Salminen-Karlsson 2011). Parallellt betonar andra studier vikten av att ta detta problem på allvar och bredda sin kunskap inom området gender bias och AI-utveckling (Mehrabi et al. 2019).

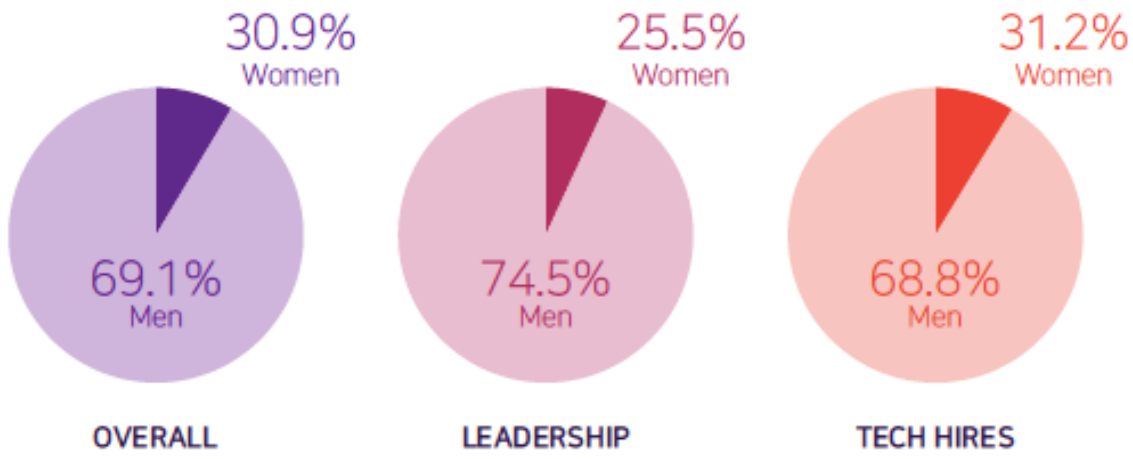
3.4.2 AI och mångfald i företag

Nästan hälften av alla kvinnor som går in i teknikbranschen lämnar den så småningom, vilket är mer än dubbelt så stor andel jämfört med männen (Ashcraft et al. 2016). Bland Facebooks AI-forskare är endast 15 % kvinnor, motsvarande siffra för Google är 10 %. Det finns inga offentliga uppgifter om trans- eller andra könsidentiteter (West et al. 2019).

Mångfalden i Googles arbetskraft liknar den på många andra multinationella teknologiföretag; mindre än en tredjedel av den totala arbetskraften är kvinnor och en ännu mindre andel av kvinnor har ledarroller, se Figur 6 nedan (Google 2018).

[3] Undersökta skolor: UC Berkeley, Stanford, UIUC, CMU, UC London, Oxford, och ETH Zurich

Informationen är inte avgränsad till tjänster inom AI, men inte mindre alarmerande för det.



Figur 6. Mindre än en tredjedel av den totala arbetskraften på Google är kvinnor och en ännu mindre andel av kvinnor har ledarroller (Google 2018 Diversity Annual Report).

För närvarande utgör kvinnor 24.4 % av computer science området och får en medellön som är 66 % av deras manliga motsvarigheter (West et al. 2019). Man bör dock vara medveten om att det är svårt att få en korrekt och heltäckande uppfattning om området, både på grund av dess komplexitet, men också på grund av den snabba utvecklingen. Det är också svårt att utifrån få insyn i företag vad gäller sambanden mellan exempelvis kön och beslutsfattande (West et al. 2019).

3.4.3 AI och mångfald i investeringar

AI är en kraftfull och relativt ny teknologi, och nystartade företag utgör en viktig komponent när det kommer till utvecklingen av AI. En viktig indikator när det kommer till könsroller inom teknik i allmänhet och inom datorer och AI i synnerhet under de senaste 30 åren är bilden av "pojkgeniet" som arbetade i ett garage och förändrade världen: Bill Gates (Microsoft), Steve Jobs (Apple), Jeff Bezos (Amazon), Jack Ma (Alibaba) och Pony Ma (Tencent) för att nämna några. Giganterna i den digitala eran och även i AI-eran är manliga. För att kunna bli så stora har alla tech-jättar behövt investeringar från riskkapitalister (World Economic Forum 2019a), och en omfattande studie på investeringar i USA påvisade att manliga och kvinnliga entreprenörer systematiskt får olika typer av frågor ställda till sig av riskkapitalister i investeringsrundor. Studien visade att män tenderar att bedömas för sin potential, medan kvinnor bedöms för sina prestationer. Detta påverkar vilka frågor män respektive kvinnor får, hur män respektive kvinnor framstår och slutligen antalet samt vilken storlek på investeringar män respektive kvinnor får (Kanze et al. 2018).

4. Metod

I detta avsnitt presenteras och motiveras de metoder och angreppssätt som använts för att genomföra denna studie. Inledningsvis beskrivs det metodologiska angreppssätt som använts och därefter beskrivs de olika delarna mer ingående tillsammans med de teorier som ligger till grund för valet av metod.

4.1 Metodologiskt angreppssätt

I denna studie har två huvudsakliga metoder använts: en litteraturstudie och en empirisk datainsamling i form av fem djupintervjuer med personer på KPMG som på olika sätt arbetar med utveckling av AI.

4.1.1 Kvalitativ metod

Samtlig datainsamling i denna studie har genomförts med hjälp av kvalitativa metoder. Kvalitativ metodologi har få restriktioner (Graziano och Raulin 2013) och baseras på kvalitativ data som exempelvis texter och intervjuer. Forskaren befinner sig själv ofta i den sociala verklighet som observeras och analyseras, och söker fånga inte bara vad människor gör utan också vad deras handlingar innebär (Nationalencyklopedin 2019). En kvalitativ metod utgår från studiesubjektets perspektiv, till skillnad från en kvantitativ som i högre grad utgår från forskarens idéer och teorier om vilka dimensioner och kategorier som ska stå i centrum (Alvesson och Sköldberg 2008). Kvalitativa metoder skiljer sig från de kvantitativa i att planeringen vanligtvis är mer flytande och flexibel, hypotes och process kan exempelvis ändras och modifieras (Graziano och Raulin 2013). Kvalitativ metod har ofta tyngdpunkt i förståelsen för snarare än förutsägelsen av ett studerat problem (Bhattacherjee 2012).

Den kvalitativa metoden är vanligtvis förknippad med ett induktivt förhållningssätt. Det induktiva förhållningssättet har inte lika stora krav på exempelvis problemformulering som det deduktiva har. Med den induktiva metoden kan en studie genomföras genom att använda kunskap och information som erfarits under den empiriska undersökningen utan kravet på att den också ska vara förankrad i specifik teori (Yin 2011).

Anledningen att en kvalitativ metod har valts för denna studie är att det primära syftet har varit att förstå ett fenomen (könsdiskriminering inom AI-system) snarare än att göra förutsägelser kring dess framtid. Då AI-system i allmänhet och könsdiskriminering inom AI-system i synnerhet är ett nytt och outforskat område finns det inte mycket teori att stödja studien på, vilket även det motiverar valet att arbeta med en induktiv metod och genomföra en kvalitativ studie.

4.1.2 Litteraturstudie

Den litteraturstudie som gjorts används på två sätt. Dels fungerar den som inläsning av viktiga bakgrundskunskaper till teoridelen och dels fungerar den som grund för den

empiriska undersökning som genomförts. Syftet med litteraturstudien har varit att få en förståelse för det undersökta problemet och undersöka vad forskare och andra experter anser vara relevant inom området. En del av litteraturstudien redovisas i teoridelen och en del redovisas i resultatdelen av rapporten. Syftet med uppdelningen är framför allt pedagogiska skäl. Den litteratur som redovisas i teoridelen ger en tillräcklig kunskap om ämnet, de begrepp, och de resonemang som är relevanta för att förstå resten av rapporten. Den består dels av teoretiska förklaringar till uppkomsten av gender bias i AI-system och dels av exempel på när gender bias i AI-system har inträffat i verkligheten. De exempel på könsdiskriminering och gender bias i AI-system som redovisas har till syfte att belysa utbredningen och komplexiteten av det undersökta problemet. Den andra delen av litteraturstudien redovisas i resultatdelen. Först sammanfattas de mest centrala orsakerna till uppkomsten av gender bias i AI-system och därefter redovisas de förslag på lösningar som identifierats i litteraturstudien. Den del av litteraturstudien som redovisas i resultatdelen ligger till grund för de frågor som ställdes under intervjuerna, vilket motiverar presentation av resultat från litteraturstudie och intervjuer ihop.

4.1.3 Empirisk datainsamling

Den empiriska delen i denna studie består av intervjuer, vilket sannolikt är den mest förekommande metoden för att samla in data i kvalitativa studier (Bryman och Bell 2013). Det finns i huvudsak två sorters intervjumetoder: semi- och ostrukturerade intervjuer (Bryman och Bell 2013). I denna studie valdes semistrukturerade intervjuer av flera skäl. En semistrukturerad intervjumetod har till fördel att frågorna kan anpassas efter respondentens kompetens och kunskapsområde men samtidigt behandla samma teman och frågeställningar vilket underlättar jämförande mellan de olika intervjuerna. Semistrukturerade intervjuer kräver en viss kreativitet och flexibilitet av intervjuaren men kan å andra sidan leda till att information som inte hade förväntats inhämtas erhålls (Bryman och Bell 2013). Ett antal frågor förbereddes därför innan intervjuerna (se Appendix A) baserat på teman som identifierats i litteraturstudien. I princip ställdes samma frågor till alla olika intervjupersoner, men ordningen i vilken frågorna ställdes varierade och så även följdfrågorna. Beroende på expertis hos den intervjuade lades olika mycket fokus på olika teman men i alla intervjuer utom en lyftes alla på förhand bestämda ämnen under respektive intervju. Totalt genomfördes fem intervjuer. Fyra av dem hölls i 30-60 minuter. En av intervjuerna var kortare än den andra, (20 minuter), varför inte alla ämnen hann täckas in i den. Samtliga intervjuer spelades med respektive intervjupersons godkännande in och transkriberades i anslutning till varje avslutat möte samt igen en vecka efter intervju för att säkerhetsställa att transkriberingen var ordagrann. Två av intervjuerna hölls via telefon och tre face-to-face på KPMG:s kontor på Vasagatan 16 i Stockholm. Två av intervjuerna hölls på engelska och resterande på svenska. I intervjuerna som hölls på engelska hade varken respondent eller intervjuare engelska som modersmål vilket ger upphov till viss risk för missförstånd eller felaktigheter. Detta tros dock inte ha varit ett problem i något av fallen då nivå av språkkunskaper var goda och utrymme för eventuella förtydligande fanns.

Syftet med djupintervjuerna var att få förståelse för och inblick i om och hur gender bias uppmärksammas och behandlas på ett företag som på nationell arena (Sverige) är i startfasen av att utveckla AI-system men internationellt (Danmark) kommit längre. Djupintervjuerna syftar också till att sätta litteraturstudien i en verklig kontext och tillåta en jämförelse mellan teori med praktik.

4.1.4 Val av respondenter

Ett medvetet och målinriktat urval av respondenter gjordes. Syftet med att avsiktligt välja respondenter är att identifiera deltagare som är relevanta för studiens syfte som kan bidra med information som är värdefull för ämnet (Yin 2011). Eftersom det studerade fenomenet är både komplext och förhållandevis outforskat var det önskvärt att intervjua personer som har bra insyn i området, varför ett avsiktligt urval gjordes. Definitivt val av respondenter var dock inte bestämt från början utan tillkom under första delen av litteraturstudien. Eftersom syftet med intervjuerna var att studera hur KPMG resonerar kring valt ämne så rekommenderade personer med hög kännedom om intern kompetens på KPMG initialt vilka som skulle intervjuas. Under de första intervjuerna tillkom även identifikation av en ny intressant intervjuperson. Detta iterativa sätt att intervjua kallas snöbollsurval och innebär att en respondent ger kontaktuppgift till ytterligare en respondent under exempelvis en intervju (Yin 2011). En kritik mot snöbollsurval är att respondenterna då inte representerar hela populationen (Yin 2011; Bryman och Bell 2013). Eftersom syftet med ett aktivt urval av respondenter var att få de mest kunniga inom området så har dock representation av hela populationen aldrig setts som ett mål i sig.

De initialt rekommenderade intervjupersonerna var någon typ av chef, se fyra första befattningarna i Tabell 2 nedan. De var därmed ansvariga för den huvudsakliga inriktningen på den tekniska utvecklingen, vilket i sig är en mycket intressant målgrupp. Då utbudet av arbetskraft som arbetar med AI på KPMG i Sverige är relativt begränsat så intervjuades även personer från KPMG Danmark. Totalt intervjuades tre personer som arbetar på kontoret i Stockholm och två som arbetar på kontoret i Köpenhamn. Alla initialt rekommenderade personer var också män, varför en aktiv förfrågan om att intervjua kvinnor samt utvecklare inom AI för att få ett mer balanserat underlag lades fram. Fyra av fem av intervjupersonerna har en teknisk utbildning (computer science eller ingenjör). Fyra av de personer som intervjuades är män och samma fyra är högt uppsatta chefer (Partner, Director, Head of Architecture and Solutions). En av de intervjuade är utvecklare inom maskininlärning. I Tabell 2 nedan redovisas respektive befattning hos de olika personer som intervjuats. Det anses inte finnas något egenvärde att presentera intervjupersonerna med namn varför deras namn, svar och uttalanden har anonymiserats.

Tabell 2. Tabell över befattning hos de fem intervjupersonerna som medverkat i denna studie.

Befattning
Director, KPMG Danmark
Equity Partner NewTech and Financial Services Nordic, KPMG Danmark
Head of Architecture and Solutions, KPMG Sverige
Director Digital Data & Analytics, KPMG Sverige
Data Scientist, KPMG Sverige

4.1.5 Sammanställning och analys av data

Det första steget i litteraturstudien var att hitta relevant information och teorier kring det föreslagna ämnet "gender bias inom artificiell intelligens". Ett antal relevanta artiklar identifierades genom exempelvis diskussion med professorer vid Uppsala universitets IT-institution. I dessa artiklar refererades till ytterligare artiklar och studier, vilka tillsammans med mer allmänna sökningar på exempelvis Google Scholar och Uppsala universitets bibliotekstjänst utgjorde grunden för litteraturstudien. Allt eftersom avgränsades frågeställningen, tills tre frågeställningar specifika nog för detta arbete utkristalliserades. Dessa lyder:

- *Hur kan gender bias eller könsdiskriminering uppstå i ett AI-system?*
- *Hur ser relationen mellan gender bias och AI-utveckling ut?*
- *Hur resonerar en teknisk avdelning på ett konsultföretag som arbetar eller snart ska inleda sitt arbete med utveckling av AI kring relationen mellan gender bias och AI?*

Den litteraturstudie som gjorts baseras på studier, undersökningar och uttalanden från forskare och andra experter samt artiklar från populärkultur. Utifrån denna litteraturstudie identifierades ett antal teman och tyngdpunkter från vilka uppkomsten av gender bias i AI-system framför allt tycks härstamma samt idéer om hur problemet kan lösas. Dessa teman fick sedan ligga till grund för utformningen av frågor som ställdes under djupintervjuerna (se Appendix A för redogörelse för vilka frågor som ställdes). Alla personer som intervjuats visste på förhand ämnet kring vilket intervjun skulle kretsa, men inte vilka specifika frågor som skulle komma. En av de intervjuade hade på förhand fått exempel på frågor som skulle ställas.

Analysen av litteraturstudien gjordes genom att läsa igenom texten många gånger och urskilja teman och mönster i innehåll och tyngdpunkt. De teman som baserat på

litteraturstudien identifierats som centrala varierar i karaktär där vissa är på teknisk nivå och andra snarare har att göra med organisation och människor.

Analysen av intervjuerna var en iterativ process. Intervjuerna transkriberades och lästes igenom många gånger för att säkerhetsställa att transkriberingen var ordagrann. Ett antal teman identifierades även här. De teman som urskildes från djupintervjuerna var av förklarliga skäl till stor del, men inte helt, överensstämmande med teman som identifierats i litteraturstudien. Några områden som haft stort utrymme i litteraturstudien hade dock inte stor tonvikt i djupintervjuerna och vice versa. Under gång som de transkriberades markerades och färgkodades också svar som ansågs vara av särskilt intresse utifrån analysen av litteraturstudien. Färgkodningen bestod av exempelvis att något som handlade om data kodades orange medan något som handlade om teamsammansättning kodades grönt, och så vidare. Varje tema behandlades sedan i ett eget stycke där allt som de olika personerna sagt kring det temat sammanställdes. Slutligen jämfördes de teman och mönster som kommit fram av litteraturstudien med de teman och mönster som kommit fram av djupintervjuerna vilket gav en möjlighet att till exempel jämföra skillnader och likheter mellan teori och praktik.

4.1.6 Leverans till KPMG DTI

Uppsatsen har skrivits för KPMG:s tekniska rådgivningsavdelning Digital Transformation and Innovation (DTI). Avsikten är att arbetet ska ge KPMG insikt i problemet med gender bias och könsdiskriminering i AI-utveckling och även en insikt om hur det ser ut på KPMG idag vad gäller ämnet. Resultaten kan användas i den fortsatta utvecklingen av AI-processer och rådgivning kring AI-utveckling. Studiens resultat presenteras för KPMG efter presentationen av studien samt rapportens godkännande från Uppsala universitet. Ett White Paper skrevs till avdelningen baserat på insikter som erhållits och slutsatser som nåtts.

4.1.7 Källkritik

I litteraturstudien har akademiska artiklar använts i stor utsträckning, men även tekniktidskrifter, rapporter från organisationer såväl som artiklar från populärkultur. Anledningen att litteraturstudien inte uteslutande består av akademiska publikationer är att ämnet är nytt och en stor del av det som är skrivet i ämnet inte är på akademisk nivå (Fessler 2017; Bass och Huet 2017; Medium 2017). Användningen av populärkultur visar även på aktualiteten på ämnet.

5. Resultat

I detta avsnitt presenteras de empiriska resultat som erhållits från litteraturstudien respektive djupintervjuerna. Inledningsvis presenteras den andra delen av litteraturstudien, vilken kan ses som en fortsättning på delen som presenterades i avsnittet ”Teori”. Den litteraturstudie som genomförts och insikterna från den har legat till grund för utformningen av djupintervjuerna, varför insikter från litteraturstudien presenteras först. Därefter följer resultaten från djupintervjuerna. I båda delar presenteras både orsaker till varför gender bias uppstår i AI-system och idéer och tankar kring hur problemen bör behandlas. Delar av resultaten diskuteras i denna del men största delen av analysen finns att läsa under avsnittet “Diskussion”.

5.1 Insikter från litteraturstudien

Inledningsvis redogörs huvudpunkterna av sambanden mellan AI-system och könsdiskriminering från teoridelen och därefter redovisas de lösningsförslag som studier, forskare och andra experter föreslår. De huvudsakliga faktorerna till varför gender bias uppstår i AI-system identifieras som redan existerande snedvridningar, snedvriden data, brist på mångfald i näringsliv och forskning samt bristande förståelse mellan teknikutveckling och genusanalys.

5.1.1 Hur uppstår gender bias i AI-system?

Litteraturstudien i föreliggande rapport visar att gender bias i AI-system kan uppstå på grund av en mängd olika faktorer, inklusive men inte begränsat till utformningen av algoritmen eller beslut relaterade till hur data samlas in, kodas, eller används för att träna algoritmer. På policynivå har saker som riktlinjer, ramverk och praxis betydelse för fenomenet. När det gäller vilken typ av forskning som beställs och vilka typer av problem som anses viktiga kan det finnas en obalans som verkar till för- eller nackdel för något av könen. Trots försök att gruppera in viktiga påverkansfaktorer på olika sätt, exempelvis faktorer som framför allt är tekniska respektive faktorer som framför allt är historiska (se t.ex. Friedman och Nissenbaum 1996 eller Mehrabi et al. 2019), så visar litteraturstudien att alla dessa faktorer i olika grad både påverkar och påverkas av varandra vilket gör fenomenet svåröverskådligt. Detta är också något som implicit bekräftas av att allt fler studier efterfrågar ett systematiskt angreppssätt snarare än att hantera problemen med gender bias i AI från isolerade platser i systemet (West et al. 2019 m.fl.). Det räcker till exempel inte med ett tillägg i algoritmen eller att ta bort en variabel (även om det ibland kan lösa vissa omedelbara bias-problem), eller att enbart tillsätta fler kvinnor i utvecklingsteamen (även om det ger bättre förutsättningar för jämställdhet mellan könen).

Gender bias uppstår och finns på många olika nivåer i AI-system men även det sociala och ekonomiska system AI-systemet uppkommit i har betydelse, och litteraturstudien visar att problemet borde hanteras därefter. Ett antal faktorer har dock identifierats vara

av central betydelse vad gäller uppkomsten av gender bias i AI-system: redan existerande snedvridning (praxis och attityder i samhälle, hos institutioner eller hos individ), snedvriden data, brist på mångfald i arbetskraften samt bristande förståelse av sambanden mellan teknik och genusanalys.

Redan existerande snedvridningar

Med redan existerande snedvridningar menas sådana som har rötter i sociala institutioner, praxis och attityder och dessa kan delas upp i historiska respektive individuella snedvridningar. De historiska snedvridningarna är ofta djupt inbäddade i våra samhällen. Exempel som fått uppmärksamhet i samband med användning av AI-algoritmer är att afroamerikaner diskrimineras i USA:s rättsväsende (Lee et al. 2019) och att Google visar högt betalda jobbannonser till manliga användare framför kvinnliga (Ford och Wajcman 2017; Howcroft och Rubery 2019). De individuella snedvridningarna grundar sig i den kognitiva bias som människor har vilka i sin tur kan påverka till exempel de regler och den logik som programmeras in i algoritmer (Weizenbaum 1976). Som både UNESCO och Europeiska Kommissionen påpekar så återspeglar teknik ofta utvecklarens värderingar (UNESCO 2019).

Snedvriden data

Snedvriden data är vad som tycks vara orsaken till snedvridna resultat i majoriteten av de mest uppmärksammade fallen med gender bias i AI-system. Det finns många exempel på hur data kan vara snedvriden: den kan vara ofullständig eller icke-representativ (det vill säga mer representativ för vissa grupper människor än andra), det kan finnas över- eller underrepresentation av subgrupper i dataset, eller det kan vara så att urvalet inte är slumpartat (Mehrabi et al. 2019 m.fl.). Det kan också vara så att data är korrekt i termer av att den representerar verkligheten som den ser ut, men likväl snedvriden eftersom verkligheten på många håll är snedvriden, exempelvis för att samhället eller platsen data kommer ifrån är könsdiskriminerande. Oberoende vilken typ av data som används visar litteraturstudien att det finns många faktorer som ger upphov till snedvridna dataset och att det krävs mycket kunskap om hur, varför och av vem data är insamlad för att ha möjlighet att säkerhetsställa att den inte är snedvriden.

Diskriminering i text, bild och ljud

Snedvridningar i AI-system tar sig varierande uttryck i olika typer av applikationer. Till viss del beror det på vilken typ av data som används. Diskriminering eller snedvridning mot kvinnor är i många fall inneslutet i vårt språk och hur det används, vilket kan reflekteras och cementeras när text används i AI-algoritmer. Saker som att män oftare nämns än kvinnor i text (Leavy 2018), att maskulina ord när norm vid översättning av könsneutrala ord (UNESCO 2019; Prates et al. 2019) eller att de adjektiv som används för att beskriva män respektive kvinnor divergerar (män beskrivs oftare i termer av sitt beteende medan kvinnor oftare beskrivs i termer av sitt utseende och sin sexualitet) (Caldas-Coulthard och Moon 2010) kommer rimligen påverka det resultat som fås när en enorm mängd text processas i en AI-algoritm.

Litteraturstudien visar att oberoende av typ av medium så är gender bias närvarande i AI-system. Algoritmer som används i bildigenkänning baserar sina beslut på stereotyper om vad som är manligt och kvinnligt (Zhao et al. 2017). Googles röstigenkänning är 70 % mer trolig att korrekt känna igen manliga röster jämfört med kvinnliga röster (Tatman 2016; UNESCO 2019). Röstassistenter inte bara speglar utan cementerar och förstärker könsdiskriminering och obalans mellan (UNESCO 2019).

Brist på mångfald i näringsliv och forskning

Det råder brist på kvinnor inom AI-området inom allt från akademin till tech-företag till startup-världen. Globalt sett är endast 22 % av yrkesverksamma inom AI-området kvinnor. UNESCO (2019) betonar att den begränsade representationen av kvinnor i team som utvecklar teknologin hotar både att upprätthålla befintliga och införa nya typer av ojämlikheter mellan könen. I princip alla tech-företag som är med och driver AI-utvecklingen framåt uppvisar olika typer av diskriminering mot icke-män (se t.ex. Mayer 2018; Melendez 2018; Microsoft Gender Case 2019; Kolhatkar 2017). Flertalet studier betonar relationen mellan diskriminering i arbetskraften och diskriminering i systembyggnad.

Bristande förståelse kring samband mellan teknik och genusanalys

Forskare lyfter också en problematik med bristande förståelse som forskare och andra experter tenderar att ha för områden som inte är deras egna (Salminen-Karlsson 2011). Studier har till exempel visat att förståelsen mellan forskare från tekniska områden och forskare från genusområden inom akademin ofta haltar och att exempelvis genusforskningens sätt att använda språk i argumentation och reflektion kan vara främmande för dataingenjörskulturen. Bland många akademiker ses forskning av genus (gender research) som en extrem form av vetenskap som är både politiskt och ideologiskt laddad. (Salminen-Karlsson 2011).

5.1.2 Hur bör problemen med gender bias i AI-system hanteras?

Den genomförda litteraturstudien resulterade utöver teman vad gäller orsaker till gender bias och könsdiskriminering inom AI-system även teman vad gäller hantering av problemet. De sätt att hantera gender bias som identifierats är: data, mångfald i teamen, utvärdering av system, medvetande (om problemet och var det kommer ifrån), inkludering av genusanalys i läroplanen för ingenjörer och datavetare samt att tillämpa ett systematiskt angreppssätt.

Data

För att förstå snedvridning (och för att kunna hantera den) krävs en grundlig förståelse för den sociala kontext från vilken data är producerad och inhämtad (West et al. 2019). Att data är snedvridet tycks, som tidigare diskuterat, vara en av de främsta orsakerna till den gender bias och könsdiskriminering som syns i resultaten AI-algoritmerna levererar. Lösningarna på problemet ligger dock utanför själva datasetet. Att exempelvis ett

dataset över Sveriges chefer innehåller en enorm övervikt av män beror inte på datasetet i sig, och i det fallet inte heller hur informationen samlats in, utan på hur samhället idag ser ut och vilka strukturer och fördomar som har format det.

Mångfald i teamen

Det verkar finnas en relation mellan diskriminerande praxis på arbetsplatser och de diskriminerande AI-verktyg som arbetsplatsen producerar. Detta blir som en återkopplingsslinga som formar AI-industrin. West et al. (2019) diskuterar det systematiska förhållandet mellan mönster av exkludering inom AI-området och industrin som driver produktionen framåt å ena sidan, och snedvridningar och diskriminering som tar sig uttryck i logistiken och resonemangen i AI-tekniker å andra sidan. Utredningsomfånget bör utvidgas till att inte bara beakta hur AI-verktyg kan vara snedvridna och diskriminerande rent tekniskt, utan också hur de formas av de miljöer där de är byggda och de människor som bygger dem (West et al. 2019).

Hittills har problem med mångfald inom AI-området och frågeställningar kring snedvridning i de AI-system som utvecklas tenderat att utvärderas separat. Detta är dock enligt många två sidor av samma problem: diskriminering i arbetsstyrka och utvecklingen av diskriminerande system är djupt sammanflätade (se t.ex. West et al. 2019). Europeiska Kommissionen släppte i april i år (2019) en rapport med riktlinjer för etisk utveckling av AI. I rapporten betonas vikten av att teamen som designar, utvecklar, testar, underhåller, distribuerar och använder dessa system reflekterar mångfalden i samhället i stort. Idealt är teamen inte bara diversifierade i termer av kön, kultur och ålder utan även professionell bakgrund och kompetens (Europeiska kommissionen 2019). Flertalet studier och experter betonar vikten av diversifierade team, inte minst för att ens upptäcka snedvridningen överhuvudtaget (se t.ex. Daly 2019).

Utvärdering av systemen

Det är viktigt att kunna upptäcka om en algoritm eventuellt är biased, och även om det finns försök från olika studier på nya ramverk för att utvärdera AI-system har det inte utvecklats något vedertaget ramverk för att utvärdera AI-system. Det tycks vara upp till varje utvecklare att hitta lämplig metod eller ramverk.

Medvetenhet om problemet och var det kan komma ifrån

Forskare poängterar vikten av att förstå varifrån biasen kommer och vad vi kan göra för att förebygga den för att kunna adressera den gender bias som AI-system påverkas av och leder till (Mehrabi et al. 2019). Forskare är också överens om att AI, precis som tidigare teknologier, kommer att spegla värderingarna från de som utvecklar dem (MacKenzie och Wajcman 1999). Olika typer av kognitiva snedvridningar kan, medvetet eller omedvetet, byggas in i AI-system. Att vara medveten om och uppmärksam på sin bias identifieras därför som en viktig pusselbit i lösningen av problemet.

För att kunna skapa rättvisa system krävs kunskap och förståelse kring hur bias i institutionella infrastrukturer och sociala maktrelationer existerar och kan ta sig uttryck. Det är lätt att förbise strukturell snedvridning ifall medvetande om att den skulle kunna finnas saknas.

I en artikel från 2014 diskuterar Schiebinger hur vetenskaplig forskning misslyckats med att ta hänsyn till kön och hävdar att fenomenet manlig standard i ny teknik visar på denna asymmetri. Schiebinger menar att detta inte är någon aktiv diskriminering: snedvridningen är i stort sätt omedveten. Det maskulina är standard för Google Translate eftersom "han" oftare finns på webben än "hon". (Detta är dock något som är under förändring: en analys av amerikansk-engelska texter i Google Books visar att förhållandet mellan maskulina och feminina pronomen har sjunkit till cirka 2:1, från en topp på 4:1 på 1960-talet). Schiebinger tar upp ett exempel där hon talat med utvecklare på Google och belyst problemet med den maskulina normen på google translate. Hon beskriver hur experterna lyssnade på henne och efter 20 minuter hittade en lösning på problemet. Poängen Schiebinger vill framföra är dock att ett bättre sätt att lösa denna snedvridning på är att inkludera alla kön i alla relevanta forskningsfaser: när man sätter prioriteringar, samlar in och analyserar data, utvärderar resultat, utvecklar patent och slutligen överför idéer till marknader (Schiebinger 2014).

Ändra läroplanen för ingenjörer och datavetare – inkludera genusstudier

Som precis diskuterat så kan vissa problem med könsdiskriminering mildras eller pareras i efterhand när de upptäckts genom att till exempel ändra hur algoritmen tolkar data. En mer djupgående lösning är att integrera medvetenhet om genus och sociala frågor i den grundläggande läroplanen för ingenjörer och datavetare. Om de som utvecklar systemen är medvetna om och kan grunderna i genusanalys skulle risken att de omedvetet utvecklar diskriminerande AI-system minska. Design bör vara könsinkluderande från början så att företag inte måste efter- och omkonstruera teknologi för det exkluderade könet, oavsett om det är kvinnor, män, eller icke-binära personer. Datavetare bör ta examen med både en grundläggande kunskap i genusanalys och etnicitet, samt med ett bredare perspektiv på sociala effekterna av deras arbete (Schiebinger et al. 2011-2018).

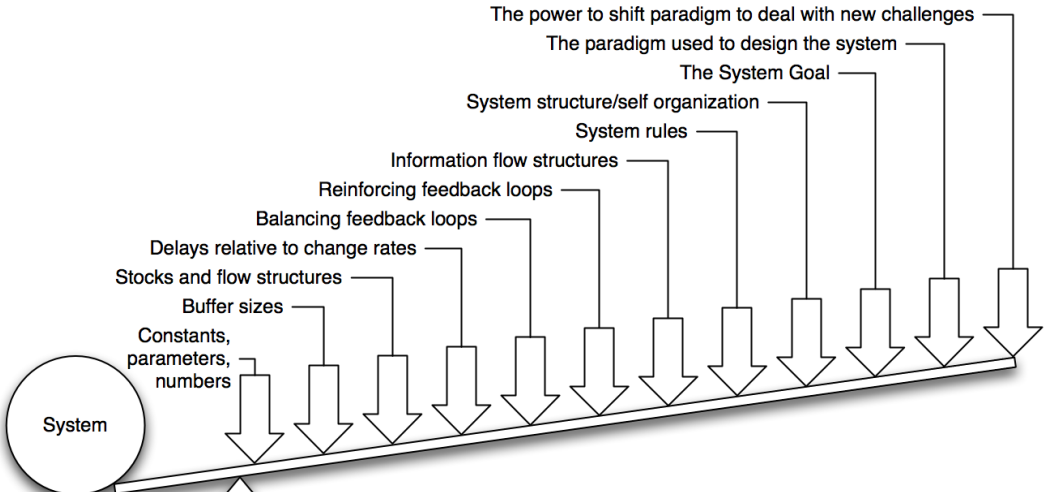
Systematiskt angreppssätt

Vissa av problemen med gender bias kan pareras med olika typer av tillägg i algoritmerna. Till exempel kan den latenta snedvridning som finns i data på grund av att män oftare nämns än kvinnor pareras genom att lägga in en kvot i träningsdata (Leavy 2018) och det finns tekniker som kan minska stereotypiska ordinbäddningar så att till exempel ordet “barnvakt” förknippas lika starkt med “morfar” som med “mormor” (Schiebinger et al. 2011-2018). Många hoppas också på att ökad mångfald i arbetskraften kommer göra stor skillnad. Men i många fall räcker det inte. En majoritet av de studier som undersöker bias fokuserar på en definition av bias som lätt kan mätas eller uttryckas rent tekniskt och algoritmer behandlas ofta som rent tekniska utmaningar

snarare än de socio-tekniska problem som de kommit att bli. En "lösning" på felaktigheter som uppstår i det tekniska systemet behöver inte nödvändigtvis lösa de övergripande systematiska problemen, och i många fall är det inte ens möjligt (Powles 2018). Flertalet studier betonar behovet av att på ett mer systematiskt sätt diskutera bias och hur bias påverkar designen och resultaten av AI-system (Howard och Borenstein 2017). Att ha en kontextualiserad adressering kan möjliggöra en mer omfattande redogörelse för uppkomsten av, och lösningen på, gender bias.

Det finns också en förväntan på att fler kvinnor i teamen kommer att "lösa problemet" med könsdiskriminering enligt idén om "just add women and stir". Denna förhoppning är inte helt oproblematisk. Om det bara tillsätts kvinnor utan att systemet samtidigt medvetet ändras kommer troligen inte problemen att lösas (Harquail 2012). Studier har visat att att bara addera fler kvinnor till system som är strukturerade efter män och där kvinnor har exkluderats från design- och management-fasen inte kommer ta kvinnor framåt i de systemen (Gender Working Group United Nations Commission on Science and Technology for Development 1995).

Det är vanligt att försöka hitta faktorer som kan göra stor skillnad ett system. Människor som gör systemanalyser tenderar att ha en övertro till att hitta så kallade hävstångseffekter (Meadows 1999). Dessa är platser i ett komplext system (ett företag, en ekonomi, en levande kropp, en stad, ett ekosystem) där en liten förändring i en sak kan ge stora förändringar i hela systemet. Meadows, forskare, lärare och författare, diskuterar systemanalys och hävstångseffekter i artikeln "Leverage Points: Places to Intervene in a System" (1999), och listar utifrån effektivitet i faktisk förändring olika insatser som kan göras för att få verkan i system. Meadows tar i artikeln upp tolv handlingsområden, se Figur 7 nedan, från de med lägre relativ effektivitet för förändring: som att ändra konstanter, parametrar och nummer (såsom skatter och substitutioner) till de med högsta relativ effektivitet för förändring: som att ändra målet med systemet eller det mindset eller paradigm ur vilket systemet – dess mål, struktur, regler, fördröjningar, parametrar – uppstår (Meadows 1999).



Figur 7. Meadows (1999) definierar tolv typer av åtgärder som kan tas för att ingripa i ett system. Åtgärderna är ordnade efter ökad ordning av effektivitet, där modifiering av konstanter, parametrar och nummer är den relativt minst effektiva åtgärden, medan ett paradigmbyte är det relativt mest effektiva sätt att ingripa i ett system.

Litteraturstudien visar att även om det finns en tendens att ha stor tilltro till hävstångseffekter så börjar allt fler allt mer efterfråga ett helhetsgrepp för att hantera problemen med just gender bias. Gender bias i AI-system specifikt kan bero på många olika faktorer i samhället, allt från vilka människor som är involverade (till exempel mångfald i utbildning, forskning och företag), till vilken data som algoritmer matas med, till vilka parametrar det är som utvärderas i ett system. Precis som flertalet forskare och andra experter poängterat så bör problemen med bias hanteras på flera olika nivåer och i alla olika steg i utvecklingen av AI-system (Schiebinger 2014; West et al. 2019 m.fl.).

5.2 Resultat från djupintervjuer

Intervjupersonernas svar har anonymiserats genom att varje person har kodats med en bokstav (A, B, C, D och E). Informationen har återgetts så nära som möjligt det exakta uttalandet från respektive person, men de citat som varit på engelska har översatts till svenska för att värna om anonymitet. De teman som identifierades från djupintervjuerna är ”data”, ”medvetande, motivation och kunskap”, ”gruppsammansättning”, ”utvärdering”, ”tekniska lösningar”, ”projektperspektiv och fokus på leverans”, ”systemperspektiv”, samt ”AI som katalysator”. Först redovisas en mycket övergripande bild av var KPMG befinner sig i sin AI-resa följt av några talande exempel gällande inställningen till bias och AI bland de intervjuade. Därefter presenteras de ovan nämnda identifierade temana. Slutligen förs diskussion kring ansvar och möjligheten AI ger att pådriva förändring.

5.2.1 AI på KPMG

KPMG Sverige är i startfasen av sin AI-resa medan KPMG Danmark kommit betydligt längre då de bland annat leder den globala inriktningen för KPMG:s AI-utveckling (Person E). I Danmark arbetar KPMG framför allt med "GOFAI", "Good Old Fashioned AI" (Person E), vilket är symbolisk AI som bygger på symboliska representationer av problem och logik (Lewis 2000). AI-arbetet består bland annat av att hjälpa företag komma igång med sitt AI-arbete, till exempel genom att ta fram handlingsplaner och långsiktiga visioner för AI-utveckling men också genom att utbilda om AI och exempelvis sitta med ledningsgrupper och gå igenom med dem vad AI är och hur det fungerar (Person D).

Etik och bias i AI-utveckling är något som är aktuellt och som det arbetas aktivt med på KPMG (Person D). Flera av intervjupersonerna framhäver den förändring som tekniken innebär och de möjligheter den medför. Teknik generellt och AI specifikt ger nya förutsättningar vad gäller maktförhållanden. Detta både förutsätter och medför även nya strukturella förhållanden (Person B). Person B menar att en ny maktbalans håller på att skapas i och med denna teknikutveckling. Kunskap är makt på ett sätt som enligt Person B skiljer sig från hur det varit tidigare, vilket också kräver nya strukturer och nya sätt att värdera kunskap. "*I dessa tider av gigantisk förändring kan vi inte hålla fast vid gamla strukturer. (...) Det kan vara helt andra människor som skapar helt andra förutsättningar som personer som tidigare hade auktoritet blir beroende av. Den förändringen är en jättestor förändring som jag märkt"* (Person B).

Samtidigt menar Person E att den bias och de snedvridna resultat vi allt oftare ser öppnar möjligheter att diskutera problem och få nya perspektiv och insikter. Ofta slängs en modell bort när det blir tydligt att ett dataset eller ett resultat är snedvridet åt något håll, och istället för att diskutera varför resultatet blev som det blev tar arbetet ofta slut där (Person E). Person E menar att diskussionen krävs för att förstå varför snedvridningen finns och huruvida den är befogad eller inte. *"Vi måste undvika bias, men vi måste också förstå den"* (Person E).

För att konkretisera vikten av att förstå bias lyfter Person E ett exempel när en algoritm visade att kvinnor generellt sett nöjer sig med en lägre lön än män på motsvarande nivå (med nivå menas här en viss typ av kompetens och erfarenhet). Person E berättar att diskussionen kring det resultatet aldrig togs, *"Och jag tycker att det är något som borde ha diskuterats"* (Person E). Person E fortsätter sedan att berätta att anledningen till att kvinnor generellt sett "nöjer sig" med en lägre lön troligen är att kvinnor generellt sett är mer riskaversiva än män (vilket bekräftas av flertalet studier, se t.ex. Jianakoplos och Bernasek 1998; Borghans et al. 2009; Charness och Gneezy 2012), och att det är något som kräver reflektion i sig. Person E lyfter även ett annat, relaterat, exempel: i Danmark är det förbjudet att använda kön som variabel när det kommer till försäkringspremier. Eftersom kvinnor är mindre benägna att ta risker än män, borde dock rimligen kvinnors försäkringspremie vara lägre än mäns eftersom de är mindre sannolika att hamna i en olycka än män (Person E). Men eftersom det genom lag är förbjudet att väga in kön i

riskbedömningen görs ingen skillnad mellan könen i försäkringspremie, trots att det verkar finnas en faktisk skillnad som i detta fall är av betydelse. Lagen att inte inkludera kön i bedömningen verkar alltså i detta fall till nackdel för kvinnor. Dessa två exempel som Person E tar upp visar hur viktigt det är att ha en djupare förståelse för vad bias är, varifrån snedvridningarna kommer samt kunskap och möjlighet att bedöma skillnad på när bias är skälig, och därmed bör inkluderas, och när den är orättvis, och därmed bör pareras. Vikten av att ha en djup förståelse för bias är något som litteraturstudien också betonar (se t.ex. Mehrabi et al. 2019). De exempel Person E lyfter exemplifierar inte bara komplexiteten i problemet med gender bias i AI-system, utan även vidden av förståelse som Person E har i ämnet. *"Så det här blir problematiskt när man försöker undvika bias. När du vill undvika bias är det enklaste men också mest ogenomtänkta sättet att göra det att säga 'Okej, så vi ser är det finns bias här, men nu bestämmer jag att det inte är någon bias' istället för att säga 'Okej jag kan se att det finns bias, låt mig försöka förstå varifrån den här biasen egentligen kommer? Är det något som är skäligt? Det kanske är okej att det finns en bias här? Kanske är kvinnor faktiskt bättre, och borde därför ha en lägre premie?'"* (Person E).

5.2.2 Data

Snedvridningar kommer ofta in i AI-algoritmer från de dataset som används, att förstå den data som används är således en viktig del av arbetet, och det är det även på KPMG (Person A, C, E och D). Den data som används på KPMG kan vara allt från intern data till extern kunddata (A, C) och när KPMG kommer in i bilden är deras jobb att identifiera vilka datapunkter som behövs för ett specifikt projekt eller uppdrag (Person A). Person D poängterar att data är avgörande när det kommer till AI-system som leder till gender bias. En betydande mängd arbete läggs på att förstå vilken typ av data som finns tillgänglig, vad i den som kan användas, om det behövs någon kompletterande data (Person D och E) men också att försöka förstå hur datagrupper korrelerar med varandra (Person D). Utvecklare understryker att data pre processing är en stor del av arbetet och att det är det viktigaste sättet som bias i AI-system kan undvikas på (Person C). Person D betonar att om datahanteringen inte görs noggrant i ett tidigt skede så kan det komma igen senare, vilket ofta leder till problem som är både dyrare och svårare att lösa i efterhand (Person D). Detta är ytterligare anledningar till varför datahantering är ett viktigt steg i tidigt led i utvecklingsprocessen.

Bias oftare från data än från algoritmerna

Fyra av fem intervjupersoner menar att det i många fall finns gender bias i data (Person A, B, E och D). Person E menar att många organisationer inser att bias i många fall inte skapas av AI utan är något som finns i indata och snarare tydliggörs av AI. Det finns olika alternativ på hur organisationer hanterar bias i indata: vissa bestämmer sig för att ignorera den bias som finns i data eftersom det inte är AI:n som skapat snedvridningen. Andra ser det istället som en chans att lyfta diskussionen om varför denna bias finns till att börja med (Person E). På KPMG Danmark premieras det senare, och vikten av att förstå bias och varifrån den kommer lyfts som viktigt och eftersträvansvärt. Person E

menar att både bias generellt och gender bias specifikt ofta finns och syns i dataset, i allt från rekrytering till försäkring till lön. Person C har däremot aldrig sett någon gender bias i data som används på KPMG och menar att även om det alltid finns viss bias och varians i data så har hen svårt att relatera till frågan om gender bias i sitt eget arbete eftersom hen ”*inte har arbetat med data som handlar om personer per se*” (Person C). Givetvis beror också hanteringen av data på vilket projekt, vilken typ av data, och vilken typ av modell som ska användas (Person B).

På frågan hur man bestämmer om data är tillräckligt bra så svarar Person D att mängden är avgörande eftersom det behövs så stora mängder data för att träna, testa och köra AI-modeller. Kvaliteten är givetvis också avgörande eftersom man ju inte kan använda data om det saknas datapunkter. Och självklart undersöks även vilken typ av data det är, vilka grupper som finns, för att vara säkra på att det är rätt typ av data.

5.2.3 Medvetenhet, motivation, kunskap

Av intervjuerna att döma så råder det viss diskrepans när det kommer till medvetande om sambanden mellan gender bias och AI-system vad gäller KPMG Danmark respektive KPMG Sverige. Utifrån de intervjuer som hållits verkar det danska kontoret ha en lite högre medvetandegrad och kunskapsnivå vad gäller olika typer av källor till gender bias och komplexiteten i problemet. Båda kontoren uppvisar dock en stor motivation att förstå och lära sig mer. Det råder också delade meningar kring huruvida det talas och diskuteras tillräckligt om gender bias på KPMG, men konsensus är att det borde göras mer. Person A tycker inte att det talas tillräckligt mycket om bias generellt på KPMG Stockholm och berättar att det ännu inte är en självklar del av processen. Eftersom processen är under konstruktion då KPMG Stockholm är så pass nytt “*Vi försöker sätta alla strukturer just nu*” (Person A), finns det dock goda möjligheter att införa diskussioner kring bias generellt och gender bias specifikt i processen. Person A säger flera gånger under intervjun att “*detta är något vi borde införa*” när exempelvis granskningskontroll, ramverk eller utbildning kommer på tal. Person A menar också att vi (människor generellt men i detta fall de som arbetar på KPMG) måste bli mer medvetna om vad vi själva har för bias. Person A tror inte att det går att ta bort bias helt och hållet, utan menar snarare att mycket handlar om att göra folk medvetna och få in det medvetandet i utvecklingsprocessen. Person A belyser att även en människa är beslutsfattare och ställer sig frågan hur man kan se till att den bias den människan har inte påverkar den individen i arbetet. Person A har gått en kurs i AI och etik men menar att det var en liten del av medvetandet som berördes och att hen behöver och vill ha mer kunskap om ämnet. Person D anser att det diskuteras en hel del och berättar att de t.ex. tidigare hade en kvinna på kontoret som arbetade med just etik och gender bias i förhållande till AI, men tycker samtidigt att det finns mycket mer att göra, och att den bias vi ser i dataset och resultat från AI-algoritmer är ett väldigt bra sätt att påminna oss om att det finns mer att göra. *“Det finns alltid mer vi kan göra. Vi måste pusha gränserna lite mer”* (Person D). Person D är också med i ett europaråd för etik och håller bland annat på att undersöka hur de diskussionerna som fortgår i rådet kan

användas på ett väldigt hands on och konkret sätt på KPMG. Person E visar på en motivation att förstå vad som ligger bakom den bias som finns och ser uppkomsten av bias i AI-system som ett tillfälle att diskutera varför det ser ut som det gör vad gäller obalans och diskriminering mellan könen i samhället. Både Person D och E lyfter att bias är något som kan vara både bra och dåligt, allt beror på vad det är en är ute efter med sina algoritmer.

Person A och B visar på motivation till att lära sig mer om bias och etik i förhållande till AI-utveckling och en vilja att lära sig och följa den praxis och den kunskap som finns. Person A och B uttrycker också önskan om att lära sig av andra och att få så mycket input och nya perspektiv som möjligt för att kunna utveckla sin process och sätta sina strukturer på ett sätt som är inkluderande, ifrågasättande och etiskt ansvarstagande från början. KPMG är generellt sett väl medvetna om riskanalys och har ett risk- och säkerhetstänk i allt de gör (Person A och B). Person A menar att på samma sätt som risk-aspekten kommer naturligt så skulle även etik och hållbarhet kunna bli en självklar del av alla processer. Det skulle också göra att etisk analys inte är något som kan prioriteras bort även om det råder stor tidspress (vilket ofta är fallet när man arbetar projektbaserat) (Person A). För att detta ska fungera behövs det dock ett annat, nytt mindset och att reflektion kring etik och diskriminering kommer in som en naturlig del hos alla medarbetare (Person A). Person B menar att AI redan idag ses som en risk på KPMG, men då framför allt i relation till det revisionsarbete som görs eftersom användning av AI kan påverka ett företags värde.

Person C berättar att hen fått kunskap om gender bias och etik under sin utbildning vilket var givande och något hen är tacksam för eftersom hen upplever att många globalt sett saknar den kunskap som behövs inom ämnet och att många gör saker "bara för att" utan att tänka på att data kan vara biased och utan ha tänka på etik alls. Att det är så många som inte tänker på exempelvis etik har också fått Person C att i ännu större utsträckning än tidigare inse vikten av diversifierade team. Person C är inte ensam om att tycka att fler behöver kunskap i ämnet: Person A tycker att det behövs kurser och efterfrågar kunskap kring gender bias "*inte bara inom applikationsområde utan lite mer allmänt – allt från rekrytering till alla delar av systemeutvecklingsportföljen*".

5.2.4 Hantering av gender bias i AI-utveckling

Alla intervjupersoner är mer eller mindre medvetna om att gender bias kan uppstå i AI-system. Det finns dock olika uppfattningar om varför det finns gender bias i AI-system och hur det bör hanteras i praktiken. Person C, som är utvecklare, menar att om det finns bias så kommer den från den data som används. *"Det* [problemet] *ska inte vara i algoritmen. I alla fall majoriteten av problemet kommer inte från algoritmen. Det kommer från data. För algoritmen är enkel matte, det är bara logik. Den kan inte påverka data"* (Person C). Samtidigt som Person C menar att algoritmen inte är snedvriden eftersom den bara är enkel matte belyser litteraturstudien hur programmerare själva kan påverka och bidra till snedvridning genom att programmera in sin egen bias i algoritmer (se t.ex. Friedman och Nissenbaum 1996, Harrison 2016 eller Weizenbaum

1976), till exempel när mål och attribut som påverkar om och hur målen nås bestäms (Barocas et al. 2019; Hao 2019). Person A menar att ett sätt att undvika gender bias är att ta bort köns-variabeln helt. Tanken är att försöka ta bort starka variabler som inte är relevanta för det som eftersöks (Person A). Person A är dock medveten om att den variabeln kan påverka ändå, eftersom andra variabler kan korrelera starkt med den, vilket orsakar en typ av implicit bias. I brist på bättre alternativ är det ändå så problemet med gender bias i AI hanteras just nu (Person A).

Vad man anser vara orsaken till ett problem påverkar rimligen vad man anser vara den bästa lösningen för detsamma, vilket också blev tydligt utifrån hur intervjupersonerna svarade på frågan hur bias undviks eller pareras i efterhand. Flera av de intervjuade lyfter olika exempel där implicit bias inträffat för att exemplifiera hur komplext problemet är och hur svårt det är att adressera. Person D och E lyfter olika exempel på implicit bias och menar att till exempel bara ta bort variabeln "kön" inte nödvändigtvis behöver lösa problemet med gender bias eftersom andra typer av kopplingar då kan hittas, som ger samma typ av obalanserade resultat. Försöken att undvika implicit bias handlar framför allt om att medvetet gå in och titta på vilka kopplingar en modell gör och vilka kopplingar som finns i den data som används (Person D). Person D betonar dock att detta är något som är svårt att göra. "*Även om du är medveten är det svårt att hitta alla dessa kopplingar som görs"* (Person D). Detta är något som även Person E belyser *"Dessa saker är inte så lätta att bara plocka bort, du kan ha en typ av andra ordningens eller till och med tredje ordningens bias som inte är så tydliga när du tittar på det"*. Person E lyfter ett exempel med ett projekt (inte ett av KPMG:s egna projekt) där de som utvecklade systemet trodde att de hade tagit bort bias men resultatet som kom ut från algoritmen visade att en av de viktiga bestämmande variablerna för resultatet var ett medlemskap i en idrottsklubb. Man ställde sig då frågan vad detta kan ha att göra med vem som blir anställd, och fann att det på vissa sportklubbar i princip bara fanns män (Person E). *"Vad du också kan se är att – och det är här det blir lite speciellt – även när du försöker ta bort bias så kommer det ibland tillbaka ändå!"* (Person E). Trots att könsvariabeln togs bort hittade alltså algoritmen kopplingar som visade att det var män som premierades. Både att upptäcka och eliminera det problemet kräver kunskap, engagemang och uppmärksamhet.

Person C menar att om det finns bias så kommer den från den data som algoritmerna matas med. På följdfrågan hur man i så fall kan tackla snedvridning i data svarar hen: *"Det finns flera sätt att tackla det. Det grundläggande sättet är att normalisera och reglera"* (Person C). På frågan om vad som görs för att upptäcka bias svarar Person A att fokus ligger på att hitta bias innan den uppstått, men att om något skulle hittas så får man titta på modellen och försöka förstå vad som kan ha blivit fel. Person B hoppas att man i framtiden kommer att kunna använda AI för att upptäcka den bias som finns och Person C talar om hur viktigt det är att hålla sig up-to-date med tekniken och det som görs inom branschen och hoppas på så sätt snabbt lära sig de senaste och bästa modellerna. Person A hoppas att när teamen är mer diversifierade vad gäller kön så kommer fler snedvridna resultat att upptäckas.

Diversitet i teamen

De flesta intervjupersonerna betonar vikten av ett väl sammansatt team, och ser värde i att ha diversifierade team vad gäller allt från kön till ålder till etnicitet (Person A, B, C och D). Person D lyfter diversitet i teamen som en av de viktigaste faktorerna för att undvika bias generellt och poängterar att just teamsammansättningen är något som hen faktiskt kan påverka, till skillnad från exempelvis dataset. Person B framhöll också den inneboende fördel som finns med att vara ett globalt storbolag eftersom företagets kultur innebär att bolagsmässiga värderingar måste fungera och vara samma över alla länder. Det finns då inget utrymme för rasism eller liknande menade hen. Person B och D betonade särskilt att de eftersträvar att rekrytera diversifierade team, och även om Person D också sa att det inte alltid fungerade bra så är det ett mål i sig att ha diversifierade team. Det finns en stark tilltro till och förhoppning på att ökad mångfald (i termer av kön) kommer minska bias generellt och gender bias specifikt. "*Min naiva förväntning är att när vi har fler kvinnor så kan de också se fler "biaskällor" som en man av någon anledning kanske missar. Kvinnor är mycket mer alerta. (...) Nu är det så många män i vårt team som försöker ta bort bias men det kanske blir lättare när kvinnor också tar lead och kan leda oss lite"* (Person A).

Accepterande och prestigelös arbetsmiljö

Mångfald ger upphov till många olika typer av diskussioner och både Person B och D lyfter särskilt vikten av en miljö där diskussioner främjas och utmanande av idéer uppmuntras samtidigt som man ska kunna erkänna att man har fel oavsett vilken nivå man befinner sig på. *"Vi vill ha en miljö där vi kan diskutera saker och jag tror att det är i diskussionen man lär sig mycket om sin egen bias också och kanske den bias man inte trodde man hade"* (Person D). "*Även fast X är director och Y är associate så blir de jämlika*" (Person B). KPMG har generellt sett en hierarkisk företagsstruktur med tydliga nivåer och arbetsuppgifter, men enligt Person B och D är det viktigt att diskussioner förs jämlikt mellan alla, oavsett om du är partner eller associate. *"Det ska inte spela någon roll om man är partner eller director eller konsult, alla har sin åsikt och alla hörs och det är den diskussionen vi försöker hålla vid liv och faktiskt uppmuntra folk att ha också"* (Person D).

Tekniska lösningar

Precis som nämnts tidigare så kan problemet med gender bias brytas ned på olika plan, och vissa delar av problemet kan i vissa fall pareras med rent tekniska lösningar. Enligt Person C finns det många tekniker med vilka man kan adressera bias i data. *"Det finns många normaliseringstekniker och många pre processing-tekniker. Det finns många sätt att hantera bias i data"*. Person E beskriver också mer konkret att när man utvecklar algoritmer med hjälp av maskininlärning kan man tillhandahålla en lista med funktioner som dyker upp som mest avgörande för resultaten. Man tittar då på den listan och kan ibland direkt se variabler som ålder, kön, etc. som man redan där kan bestämma att inte använda. I de fall man bestämmer sig för att ändå använda de variablerna kan

man sedan ibland se att de kommer upp som viktiga determinanter för utfallet. Om så är fallet tas de bort då. Men även om ingen av dem är närvarande så måste man undersöka den andra ordningen och försöka att ta reda på om det finns någon relation mellan resultatet och exempelvis kön, etnicitet eller ålder i de parametrar som visade sig vara avgörande faktorer för utfallet. Det här kan vara väldigt svårt eftersom om andra eller tredje ordningen visar att människor som har gjort någon viss sak är män (exempelvis haft som hobby att vara en gamer eller har en bakgrund inom en fintech-startup) kan det vara ett problem med implicit bias ändå eftersom exempelvis mer än 90 % av fintech-startups är män (Person E). Det är av bland annat dessa skäl som det är så viktigt att utvärdera resultaten noga och ur flera olika synvinklar och perspektiv.

Utvärdering

För att upptäcka om en algoritm är snedvriden och ger snedvridna resultat måste den utvärderas. Beroende på vad det är man är ute efter är det givetvis olika saker som utvärderas, så det är inte bara den faktiska utvärderingen som är viktig utan även hur den går till och vilka aspekter det är som granskas. KPMG har ofta väldigt klara specifikationer från kund om vad det är de vill åt, till exempel "*lösningen ska fungera*", "*modellen ska svara inom två sekunder*" eller "*noggrannheten ska ligga inom intervallet X och Y*" (Person D). KPMG kan naturligtvis också komma med idéer om vad som borde ingå i en bra lösning, men det är kraven som kunden ställer som är det som det fokuseras på (Person D). När en algoritm sedan utvärderas så tittar man framför allt hur bra den går (Person B), vilket enligt Person B är att den skulle kunna gå av sig själv när människorna sover, alltså att ingen behöver övervaka den. Det finns enligt Person C många tekniker för att utvärdera systemen. Korsvalidering är en vanlig teknik men vilken teknik som används avgörs helt av vilken algoritm som utvärderas (Person C). Enligt Person A finns det idag ingen direkt metod för att utvärdera systemen men det finns verktyg och processer för hur man ska arbeta när fel upptäcks. Man kan till exempel titta på modellen och se om någon av variablerna har övervikt, och om så är fallet kan man antingen ta bort data eller ändra algoritmen så den inte inkluderar vissa fält (Person A).

I litteraturstudien framkom att allt fler företag efterfrågar ramverk och praxis just eftersom det inte finns några tydliga sådana, och det är något som även KPMG gör. Person A och B efterfrågar fler lager av kontroll för just etisk AI-utveckling. KPMG Sverige försöker sätta sina strukturer just nu (Person A) och både Person A och B poängterar vikten av att hela tiden påminna sig om att det finns fördomar men också vikten av att kunna testa sig själv angående sina egna fördomar (Person B), något ett ramverk kanske skulle kunna göra lättare. På KPMG Danmark finns ett ramverk med en uppsättning regler och krav på vad som bör ingå i AI-utvecklingen för att uppnå så kallad "trusted AI" och för att AI-utvecklingen sker på ett hållbart sätt (Person E). En av dessa sju regler är att man måste se till att man undviker bias i sättet man utvecklar AI-system på (Person E). Det pågår just nu en diskussion huruvida dessa regler kan användas i något slags ramverk så att "*när du utvecklar är det* [att utveckla etiskt] *något*

som kommer naturligt som en sak som du bara måste göra. Så verktyget får dig att göra det" (Person E). Person A vill i sin roll också införa någon slags granskningsmall som kan ganska uppifrån.

Projektperspektiv och fokus på leverans

Det finns på KPMG en stor motivation till att bli branschledande inom AI för sin egen bransch [revision m.m.], "*vi arbetar bara inom områden vi är starka*" (Person B). Men det finns också en viss stress över att det måste gå fort framåt *"Det är hastighet som gäller. Vi ska inte vara där om tio år. Vi ska vara där om tio månader"* (Person B). Att KPMG ofta utför arbete åt någon annan och att mycket av arbetet är projektbaserat och leveransfokuserat är något som tas upp flertalet gånger i flera intervjuer som något inte är helt oproblematiskt utifrån ett etiskt perspektiv. Vilken data som används avgörs till viss del av omfattningen på projektet (Person A). KPMG är generellt väldigt leveransorienterat, och majoriteten av tid och fokus ligger av förklarliga skäl på just leveransen (Person A och D). Etik är något som ibland kan glömmas bort eller nedprioriteras om man är under tidspress (Person A) och att minska pressen på leveranstid ser Person A som en av två viktigaste saker för att omställningen till mer etisk utveckling av AI ska lyckas på KPMG. Den andra saken Person A ser som viktigast för omställningen till mer etisk utveckling av AI är att ändra mindset kring AI och få människor att tycka biasanalys i projekt är en lika central analys att göra som annan riskanalys är.

5.2.5 Vem är ansvarig?

Då konsekvenserna av bias inom AI kan vara skadliga samtidigt som ramverken kring etisk utveckling och prioriteringar vad gäller tidsåtgång respektive etiskt fokus inte är helt klara kan man fråga sig vem det är som är ansvarig när det kommer till att säkerhetsställa att systemen är bra ur etiskt perspektiv. Utifrån empiri förstås att det just nu är öppet fält, det finns inga riktiga ramverk och ingen som tvingar utvecklare att tänka på allt man kanske borde (Person E). "*Just nu handlar det mer om välvilja skulle jag säga*" (Person E). Idag är det teamet eller personen som utvecklar som måste vara alert och som har ansvar för att utvärdera sin egen programvara (Person A och D). Det är en data scientists ansvar att granska modellens förmåga att mappa till facit, men en paraplyprocess i datasciencedelen är också att se "okej, finns det bias här?" (Person A). Person A menar att det vore bra med en kontrollfunktion av något slag, men vill helst inte att det ska vara en person som kontrollerar utan i så fall snarare en person som lär ut hur man kan kontrollera sig själv. *"Det räcker inte att en eller två personer har rätt mindset, alla måste ha det"* (Person A).

Utvecklingsteamet har väldigt stort ansvar i hela processen eftersom de både bestämmer vilken data de ska använda (i samråd med kund), utvecklar själva algoritmen och utvärderar systemen när de är klara (Person C och D). *"De får komma på en algoritm som de tror är bäst (...) det ligger hos teamet för det är de som är experter när det kommer till det"* (Person D). Detta är ytterligare anledning till varför diversitet i teamet

är så viktigt, för att fler perspektiv ska påverka utvärderingen (Person D). (Person C) talar också om eget ansvar, både när det kommer till att ha kunskap och kännedom generellt *"Det är otroligt viktigt att personen som kodar vet vad han eller hon gör"* (Person C). Det är också viktigt, och nödvändigt, att hålla sig uppdaterad om exempelvis nya, bättre modeller. *"Det är absolut nödvändigt att en AI-ingenjör alltid håller sig uppdaterad med marknaden. Även om jag inte har läst någonting så blir jag automatiskt uppdaterad när jag pratar med mina kollegor eller går på konferenser eller liknande. Om du är nyfiken så blir du uppdaterad angående nya modeller eftersom du när du hör om något som du inte känner till så vill du läsa på om det!"* (Person C).

5.2.6 AI som katalysator för diskussion och förändring

Flera av de intervjuade påpekar att problemen med gender bias i AI-system har möjlighet att ge oss kunskap om och insikter kring problemen med könsdiskriminering eftersom de visar tydliga bevis på problem som finns. Nedan är några citat från olika intervjuer som går i linje med tesen att den gender bias som speglas i AI-system har möjlighet att leda till diskussioner om könsdiskriminering i samhället som annars kanske inte tagits.

> *"Det finns ju en väldigt klar gender bias i samhället som vi väldigt aktivt måste arbeta för att få bort, och det här visar ju från ett annat perspektiv att gender bias finns där. För om vi tar fram en algoritm som vi absolut inte tar fram för att visa på bias som identifierar att det finns bias på den data vi har, det visar ju igen att vi inte bara känner att det finns bias utan att vi med data också kan visa att den finns. Och det är ett bra sätt att trycka på att det måste göras mer"* (Person D).

> *"De flesta organisationer förnekar på vissa sätt bias och sen förstår de att bias inte är något som skapas av AI, det är snarare tvärtom – det avslöjas av AI! Och när de förstår att det är så det ligger till så väljer de i de flesta fall att inte ta hänsyn till det. Från mitt perspektiv är det också lite problematiskt. För när det finns olika snedvridningar kan det vara en god sak att ta upp dem och öppna upp diskussion kring dem. Både säga 'okej men varför är det här ett problem!?' och gör det tydligt att den här snedvridningen existerar"* (Person E).

> *"Det ger liksom en annan infallsvinkel på mycket av de diskussionerna som pågår. T.ex. kan det vara så att man inte tycker det är något problem men jo om vi tittar på data och hur algoritmerna fungerar så ser vi ju att det finns en gender bias. Och det har vi ju sett också när vi tittar på företag och ledningsgrupper – hur många kvinnor sitter i dem? Det är ju väldigt begränsat idag"* (Person D).

6. Diskussion

I denna del diskuteras och analyseras de resultat som redovisas i resultatdelen. Inledningsvis redogörs mycket övergripande för KPMG:s förutsättningar och ramar för AI-utveckling. Därefter diskuteras de faktorer och de teman som framkommit från litteraturstudie och djupintervjuer samt likheter och skillnader dem emellan. Diskussion förs utifrån de resultat som identifierats i förhållande till de frågeställningar som presenteras i början av rapporten: ”Hur kan gender bias eller könsdiskriminering uppstå i ett AI-system?”, ”Hur ser relationen mellan gender bias och AI-utveckling ut?” samt ” Hur resonerar en teknisk avdelning på ett konsultföretag som arbetar eller snart ska inleda sitt arbete med utveckling av AI kring relationen mellan gender bias och AI?” vilka legat till grund för studien.

Från litteraturstudien och de intervjuer som genomförts i denna studie förstås att bias generellt och gender bias specifikt kan smyga in i systemen medvetet eller omedvetet på grund av många faktorer som kan härstamma från alla faser i ett AI-systems livscykel. Litteraturstudie och djupintervjuer visar också att några faktorer är mer kritiska än andra samt att några faktorer är lättare att korrigera för än andra. Eftersom de djupintervjuer som gjorts baseras på identifierade teman i litteraturstudie stämmer de teman som hittats i intervjuerna av förklarliga skäl till viss del, men inte helt, överens med de i litteraturstudien. Intervjuerna visar också att praktik skiljer sig från teori i vad som är viktigt och hur problemen hanteras.

6.1 KPMG:s betydelse för AI-utveckling generellt

AI-arbetet på KPMG består bland annat av att hjälpa företag komma igång med sin AI-verksamhet, utbilda om AI genom att exempelvis gå igenom med ledningsgrupper vad AI är och hur det fungerar, samt göra AI-implementationer. KPMG kan därför i sitt arbete påverka många företag – inte bara explicit genom designen av AI-implementationer utan även genom att påverka hur företag tänker kring AI och vad som anses viktigt när det kommer till utveckling och användning av AI. De värderingar och den kunskap KPMG har kring AI ger således ringar på vattnet vilket understryker KPMG:s roll i AI-utvecklingen även utanför sin egen verksamhet.

6.2 Kundens betydelse

KPMG är leverans- och kunddrivet. Ofta har KPMG väldigt klara specifikationer från kund om vad som ska åstadkommas och det är de kraven som sätter ramarna för respektive projekt. Omfattning på projekt i termer av tid och budget är avgörande för vad som kan hinnas med och vad som kan prioriteras när ett projekt planeras, genomförs och utvärderas. När arbetet utvärderas så utvärderas det rimligen utifrån de krav som kunden initialt ställt. Det är således framför allt kundens krav på en lösning som avgör vad lösningen innefattar. Det gör att om inte kunden är medveten om etiska aspekter av AI-utveckling i termer av exempelvis könsdiskriminering, förslagsvis

genom att ha frånvaro av könsdiskriminering som ett mål eller delmål, finns det risk att det inte är något KPMG väljer att prioritera och fakturera för. Det kan vara svårt att motiveras att arbeta etiskt om det inte är något som varken efterfrågas eller utvärderas. Kundperspektivet och den makt kunden har i ett projekt kan således göra att genus och etik inte utvärderas. Det stora fokus på leverans och den tidspress leveransaspekten ofta medför tas av flera intervjupersoner upp som en annan avgörande faktor till att arbete för och utvärdering av etiska aspekter ibland undermineras och bortprioriteras. En förändring skulle kräva dels att kunden är medveten om problemen men också att kunden är både beredd och kapabel att betala för den eventuellt längre process ett tillägg av etisk analys skulle medföra.

6.3 Teknikens begränsning

Det uppstår alltid vissa problem när en modellering av en verklighet ska göras. Eftersom en modell per definition är en förenkling måste oundvikligen mer eller mindre komplexa förhållanden förenklas. En AI-algoritm kan ses som en modell av den verklighet algoritmen försöker visualisera. Generellt sett krävs det att målen som ska användas i en AI-algoritm i någon mån är mät- och beräkningsbara – både för att de ska kunna översättas till en algoritm och för att det ska gå att avgöra om de har uppnåtts. En förutsättning för att en algoritm ska kunna nå ett mål är alltså att målet kan översättas till något som kan beräknas och uttryckas matematiskt logiskt, vilket givetvis är olika svårt beroende på vilket mål det handlar om. Att mäta etisk korrekthet generellt och könsdiskriminering specifikt är av uppenbara och mångbottnade skäl svårt, även utan kriteriet att de ska kunna uttryckas matematiskt. En majoritet av de studier som undersöker bias fokuserar, förmodligen av ovan nämnda skäl, på en definition av bias som lätt kan mätas eller uttryckas rent tekniskt, men det är många aspekter av diskriminering som inte ryms i rent tekniska uttryck. Vilka attribut som beslutar om ett mål uppnås beror dessutom rimligen av vilken data som finns till hands, och sannolikt väljs attribut efter den data som finns framför den data som varit önskvärd. Den tekniska begränsningen i hur en algoritm är utformad och anpassad efter den data som finns kan således påverka vilka mål och attribut som väljs, vilket i flera fall riskerar att bli på bekostnad av etisk korrekthet.

6.4 Rättvisa

Det finns ett uttalat problem i att definition av rättvisa skiljer sig åt beroende på social kontext och personliga åsikter. Både inramning av målbild, val av för målet relevanta attribut, och definition av vad som anses rättvist påverkar hur en algoritm konstrueras – och i betydande utsträckning om och hur snedvriden den blir. Även om både KPMG och deras kunder är medvetna om etik och snedvridningar i förhållande till AI-system kommer sättet ”rättvisa” definieras på att ha betydelse för både hur rättvisa förväntas uppnås och hur det utvärderas om så gjordes. Precis som Selbst et al. (2019) säger så måste rättvisa förstås i förhållande till den sociala kontexten. Det är till exempel

problematiskt om rättvisa definieras på samma sätt i straffrättsliga processer som i rekryteringsprocesser.

Frågan om rättvisa är komplex på flera plan. Vikten av att ha en djup förståelse för gender bias för att kunna hantera problemet är något som litteraturstudien betonar och exempel från både intervjuer och litteraturstudie visar att den förståelsen också krävs för att kunna utvärdera och ifrågasätta vad rättvisa är och hur det står i relation till AI-system. Exemplet som en av respondenterna lyfte där könsvariabeln togs bort i beslut av försäkringspremier illustrerar komplexiteten när det gäller rättvisa och gender bias. Även om man förstår varifrån bias kommer och varför den finns, så måste beslut tas huruvida den ska användas eller förkastas för att nå så rättvist resultat som möjligt. Detta kan då mynna ut i en fråga på ett större, mer filosofiskt, plan – är rättvisa att alla behandlas lika oavsett kontext? Eller är rättvisa att individer behandlas utifrån egenskaper som har betydelse för utfall? Vilka egenskaper är det i så fall som ska inkluderas och vem ska ha makt att bestämma vilka de är? Det är inga lättvindiga frågeställningar och att förvänta sig att deras lösningar är mindre komplicerade är problematiskt.

6.5 Data

Snedvridningar i dataset är en av de viktigaste anledningarna till att algoritmers beslut och resultat kan bli snedvridna och diskriminera olika grupper. Problemen med obalanserade dataset är både komplexa och utbredda, och hade det funnits en enkel lösning på dem hade de troligen inte funnits. Två olika typer av resonemang kring hur problemen ska hanteras kan dock urskiljas i de intervjuer som genomförs. Det finns dels en inställning att göra det bästa utifrån förutsättningarna, det vill säga att även om en medvetenhet om att det existerar bias i dataset finns så används dataseten, men man applicerar olika tekniker för att parera biasen så gott det går. Det andra resonemanget handlar snarare om att försöka förstå och lyfta diskussioner kring varför det ser ut som det gör till att börja med. Båda dessa två inställningar framkommer också i litteraturstudien, och det bör poängteras att det ena inte behöver ske på bekostnad av det andra. En tydlig skillnad mellan litteraturstudie och praktik är dock att en stor mängd studier efterfrågar förändring på många olika plan i samhället, medan ett företag som arbetar utifrån givna förutsättningar får finna sig i de ramar som ges just nu.

I strävan efter att förbättra rättvisan i data behövs det dock återigen reflekteras över vad rättvisa är, vilket flertalet studier belyser (se t.ex. Schiebinger et al. 2011-2018). Ska data representera världen så som den ser ut eller så som vi önskar att den ser ut? Vem ska ta dessa beslut? Ingenjörer och datavetare? Etiska kommittéer? Regeringar eller företag? Det är stora frågor och svårt att se hur det ska gå till i praktiken.

6.6 Strukturellt angreppssätt

Både litteratur och empiri visar och efterfrågar i olika utsträckning att ett strukturellt och grundligt angreppssätt på problemen med gender bias inom AI-system behövs.

Intervjuerna visar dock att praktik skiljer sig från teori gällande var fokus ligger och hur problemen hanteras. Det är av flera skäl naturligt att omfånget av vad de som forskar på AI och gender bias eftersöker respektive vad de som arbetar med AI i praktiken gör, skiljer sig åt. I litteraturstudien har ett stort antal studier och fallstudier sammanställts och ett stort antal perspektiv har därför utvärderats i syfte att få en heltäckande bild. I praktiken har ett företag inte alltid möjlighet att påverka alla delar av systemet som påverkar resultatet och ofta prioriteras att göra det bästa av de resurser och de verktyg som finns. Att diskutera hur data samlas in och vilka typer av problem som premieras på ett högre plan är av förklarliga skäl inte KPMG:s huvudprioritet, medan den diskussion som förs av akademiker och i populärkultur både kan ha högre och andra ambitioner och rikta sig mot en större målgrupp.

Problemen med implicit bias i data är svåra att lösa och av informationen som erhållits i både litteraturstudie och intervjuer att döma verkar ingen ha utvecklat metoder som helt och hållet kan hantera den omedvetna snedvridningen. Just nu handlar arbetet framförallt om att vara medveten om att implicit bias finns, uppmärksam nog att hitta snedvridningar och oönskade kopplingar i data samt motiverad att avsätta tid i syfte att hantera bias.

Förändring av mindset är något som återkommer gång på gång som svar på hur förändringen av att börja arbeta mer etiskt skulle kunna ske rent praktiskt. Utifrån Meadows (1999) är just förändring av mindset en nyckel till faktiskt förändring. En av intervjupersonerna lyfter exemplet med att riskanalys har kommit att bli en naturlig del av alla processer på KPMG och på motsvarande sätt skulle etik (och genusanalys) kunna bli det. Det finns dock en väsentlig skillnad mellan riskanalys och etiskt ansvar som inte diskuterades under intervjuerna – nämligen den ekonomiska aspekten. Risk i IT-verksamhet medför en stor ekonomisk risk. Könsdiskriminering ses inte som en risk på samma eller motsvarande sätt. Hade den gjort det, eller hade det funnits ekonomiska incitament att inte könsdiskriminera, så hade historien och en stor del av de snedvridna dataset som representerar historien troligen sett helt annorlunda ut idag. Och kanske är det därför det inte är något som automatiskt görs idag heller. *"(...) det finns inget som egentligen tvingar dem att tänka på detta. Det är mer goodwill skulle jag säga"*(citat från en av intervjuerna). Kanske kommer det etiska ansvaret snart backas av ekonomiska incitament, men om det blir så och vilka konsekvenser det skulle få är ännu oklart.

Att det danska kontoret har en lite högre medvetandegrad och kunskapsnivå vad gäller olika typer av källor till gender bias och komplexiteten i problemet än den svenska är till viss del naturligt med tanke på hur mycket längre KPMG Danmark arbetat med AI-utveckling än KPMG Sverige. Troligen kommer skillnaderna utjämnas med tiden då det är samma företag och det hos alla inblandade finns stor motivation att både lära sig, ifrågasätta, lyfta diskussioner och ta problemen på allvar. Den explicita önskan om att lära sig mer om gender bias i alla delar av verksamheten, inte bara ur ett AI-perspektiv, utan mer allmänt i allt från rekrytering till alla delar av systemeutvecklingsportföljen, är

hoppfullt med tanke på det strukturella och genomgående angreppssätt som litteraturstudien efterfrågar.

6.7 Mångfald och arbetsklimat

Att exempelvis det tidiga arbetet med utvecklingen av Cortona till stor del handlade om röstassistentens sexliv tyder på att problemen med sexuella trakasserier finns redan i teamen som utvecklar. Parallellt visar studier att kvinnor är minst lika kapabla som män att bidra till tekniska lösningar och tenderar att bidra mer än män när det kommer till samhälls- och etiska aspekter (Nesta 2019). Detta stöder tanken på att kognitiv mångfald påverkar den producerade forskningen och antyder att ökad inkludering av kvinnor leder till ökat engagemang i sociala frågor (World Economic Forum 2019b). Större mångfald i teamet som utvecklade Cortona hade kanske lett till andra resultat. Problematiken med röstassistenterna visar också på okunskap vad gäller vidden av genomslagskraft en allmänteknologi kan ha, vilket är anmärkningsvärt. Precis som Schiebinger et al. (2011-2018) skriver så bör datavetare ta examen med både en grundläggande kunskap i genusanalys och med ett bredare perspektiv på de sociala effekterna av deras arbete – något som kanske är viktigare än någonsin.

Betydelsen av mångfald på arbetsplatser är något som både litteraturstudie och empiri lyfter genomgående. Flertalet studier understryker vikten av att teamen som kravställer, designar, utvecklar, testar, underhåller, distribuerar och använder AI-system reflekterar mångfalden i samhället i stort. Både litteraturstudie och intervjupersoner betonar att diversifierade team reducerar risken för att snedvridningar förbises och ökar chanserna att de upptäcks om de skulle smyga in i systemen. En relaterad aspekt som tog stort utrymme i empiri men inte litteraturstudie var vikten av en god arbetsmiljö och betydelsen av att ha ett bra arbetsklimat i termer av främjande av diskussioner och acceptans samt uppmuntran till olika typer av idéer och ifrågasättande av status quo. Vikten av att allas ord ska vara lika mycket värda lyftes också, något som kanske är av extra betydelse på ett företag som generellt sett är väldigt hierarkiskt. De fördelar som diversifierade team medför, exempelvis mycket diskussioner tack vare gruppmedlemmars olika perspektiv, förutsätter ju att alla kan komma till tals och alla blir lyssnade på, men det är ändå en variabel som lätt kan glömmas bort. Det kan också poängteras att kan vara något som ibland är lätt att säga i teorin men svårt att leva efter i praktiken. Det blev under intervjuerna också tydligt att det i praktiken är viktigt att låta alla komma till tals eftersom det idag till skillnad från hur det var tidigare inte är lika självklart vem som har den relevanta kunskapen. En av de intervjuade talade om en ny maktbalans som skapas i och med den nya tekniken och hur olika typer av kunskap värderas på nya sätt, vilket både förutsätter och medför nya strukturella förhållanden. Det här är en skillnad från hur det såg ut tidigare enligt personen i fråga, och den skillnaden kommer troligen accelerera i symbios med den tekniska utvecklingen.

Betydelsen av ett sunt och vänligt arbetsklimat är ett perspektiv som inte fick stort utrymme i litteraturstudien, möjligtvis för att det är svårt att som utomstående få inblick

i hur beslut fattas och hur arbetsklimat ser ut på företag generellt. Det är svårt att utifrån få insyn i företag vad gäller mångfald och beslutsfattande eftersom bara angivandet av antal män och antal kvinnor inte ger en heltäckande bild av situationen. Precis som mångfald är viktigt ur ett horisontellt perspektiv, alltså att det finns mångfald på alla avdelningar i ett företag, är det högst centralt även ur ett vertikalt perspektiv, alltså både bland de som tar beslut och bland de som verkställer dem. Flertalet exempel och studier som redovisas i litteraturstudien tyder också på att det finns en relation mellan diskriminering i arbetskraft och diskriminering i systembyggnad. Att KPMG aktivt arbetar för ett gott arbetsklimat och fria diskussioner är således viktigt ur många perspektiv.

6.8 Betydelsen av utbildning

Många forskare anser att läroplanen för ingenjörer och datavetare bör ändras till att innehålla bland annat genuskritiskt tänkande. Den av intervjupersonerna som senast tog examen från universitetet lyfter själv upp hur den utbildning i etik som hen fått i sin ingenjörsutbildning påverkar det kritiska tänkandet och att det har gett ett väldigt värdefullt perspektiv som hen upplever att de som inte fått utbildning i detta saknar. Schiebinger et al. (2011-2018) skriver att datavetare bör ta examen med både en grundläggande kunskap i genusanalys och etnicitet, och med ett bredare perspektiv på sociala effekterna av sitt arbete. Kanske hade de utvecklare som programmerade in lekfulla svar på sexuella trakasserier tänkt annorlunda om de haft grundläggande kunskap i genusanalys samt en uppfattning om och intresse i de sociala effekterna deras arbete kunde komma att få i en verklig kontext. Den bristande förståelse som verkar finnas mellan tekniskt kunniga och kunniga inom genusanalys skulle troligen minska och synen på att gender research är en extrem form av vetenskap som är både politiskt och ideologiskt laddad skulle kanske förändras.

De som är skickliga inom AI-utveckling är ofta nyfikna och snabblärda personer. I och med att området är så pass nytt så måste man vara nyfiken och anpassningsbar för att följa med i det snabba tempot. Den utvecklare som intervjuats bekräftar detta genom att diskutera vikten av att vara lyhörd när det kommer till den senaste tekniken och de senaste modellerna. En utvecklares strävan efter att alltid hålla sig uppdaterad på vad som händer inom området är något som AI-branschen borde ta tillvara på. Om ämnen som gender bias inom AI lyfts på konferenser och i andra sammanhang där den senaste tekniken, kunskap och trender presenteras finns det stor chans att ämnet blir viktigt för fler utvecklare.

6.9 Olika angreppssätt vad gäller bästa lösning

Vad man anser vara orsaken till ett problem påverkar rimligen vad man anser vara den bästa lösningen på detsamma. Men även ens individuella möjlighet att påverka problemet verkar ha en viss betydelse för vilka verktyg man vill ta till. Rimligen utgår en person som möter ett problem i ett projekt med en eventuellt snäv deadline från vad

som är möjligt att göra med det personen har till hands. De utvecklare som arbetar på KPMG har av naturliga skäl ofta inte möjlighet att samla in egen, mer balanserad, data om de upptäcker att den data de har är snedvriden. Följaktligen söker de lösa problemen med de verktyg de har. På motsvarande sätt ser chefer möjligheter i att lösa problemen utifrån de verktyg de har – t.ex. att påverka mångfalden i teamen genom vilka de anställer. De chefer som intervjuats lyfter mångfald som en extremt viktig faktor för att hantera gender bias i AI-utveckling, och flera intervjupersoner ser det till och med som den viktigaste faktorn.

6.9.1 Tekniska lösningar på sociotekniska problem

Många studier fokuserar bara på de tekniska systemen och placerar alla problem inom dem utan att ta hänsyn till kontexten de är implementerade i kombinerat med incitamenten som driver dess utveckling. Detta är också något som analysen av intervjuerna till viss del bekräftar, då den intervjuperson som i dagsläget arbetar mest tekniskt också ser framför allt tekniska lösningar på de problem som syns. Samtidigt går det inte att säga att personerna i fråga inte skulle vilja lösa problemen på andra sätt om möjlighet fanns; det är lätt att i teorin ha stora planer på hur saker och ting ska lösas men arbetar man med något praktiskt så är det ibland tvunget att inse att det finns en diskrepans mellan teori och praktik av en eller flera anledningar.

6.9.2 Det finns ingen quick fix

Människor som gör systemanalyser tenderar att ha en övertro till att hitta så kallade hävstångseffekter. Försöken att hitta hävstångseffekter kan liknas vid försök att hitta enkla lösningar på komplexa problem. Det finns till exempel förhoppningar om att tillägg i kod eller nitisk "cleaning" av data ska lösa problem med gender bias i AI-system. Det finns också tilltro till att fler kvinnor på plats i teamen kommer att "lösa problemet" enligt principen "*just add women and stir*". Men sannolikt behövs större, mer omfattande, förändringar än så. Studier har visat att förändringar av parametrar bör ske i samband med medvetna systemförändringar (se t.ex. Gender Working Group United Nations Commission on Science and Technology for Development 1995). Att män tenderar att bedömas för sin potential medan kvinnor bedöms för sina prestationer (Kanze et al. 2018) grundar sig troligen i både kognitiv bias hos riskkapitalister och de historiska snedvridningar, fördomar och diskrimineringar som finns i samhället som stort. Att bara få in fler kvinnor i investeringsrundor för AI-startups kommer inte ta fler kvinnor framåt inom AI så länge det är så att kvinnor systematiskt får färre och mindre investeringar just för att de är kvinnor.

Den manliga normen i teknikutveckling är inte något nytt, och att det maskulina fortsätter vara norm även i AI-system beror på flera anledningar. När Google blev uppmärksammade på att Google Translate som standard översatte könsneutrala ord till maskulina åtgärdades det relativt snabbt och enkelt. Men forskare poängterar att ett bättre sätt att lösa snedvridningen är att inkludera kön i alla relevanta forskningsfaser:

när prioriteringar sätts, när data samlas in och analyseras, när resultat utvärderas, när patent utvecklas och slutligen när idéer överförs till marknader (Schiebinger 2014).

Att lösa problemen innan de uppstår är också vad majoriteten av intervjupersonerna strävar efter, men samtidigt är det lättare sagt än gjort. Det finns inga enkla lösningar på diskriminering inom AI-system. Det är en pågående process, precis som diskriminering i alla andra aspekter av samhället (Hao 2019). Men även om de snabba och enkla lösningarna ofta har visat sig vara för bra för att vara sanna är det ändå de som ett företag i praktiken har att tillgå. Både intervjupersoner och litteraturstudien ger flera exempel där snabba lösningar som till exempel att ta bort könsrelaterade variabler inte löser problemen med snedvridning utan bara förflyttar dem och ibland gör dem svårare att upptäcka. KPMG försöker lägga fokus på att lösa problemen innan de uppstår, vilket görs genom att använda sig av olika tekniker, lägga mycket tid på datahantering och försöka tillsätta diversifierade team. Att arbeta med att "städa" data är ett väldigt viktigt steg i utvecklingsprocessen, men även om man "gör man det man kan" så är det uppenbarligen inte tillräckligt. Enligt alla intervjuade på KPMG, oberoende av nivå av medvetenhet kring problem med gender bias i AI-system, finns det alltid mer att göra.

6.10 Vem är ansvarig?

De personer som utvecklar har stort eget ansvar i utvecklingsprocessen. Att AI är en teknik som är så pass ny och okänd för en stor del människor, även högt uppsatta, påverkar troligen varför en stor del av ansvaret delegeras längre ner i kedjorna. Både litteraturstudie och empiri är överens om att problemet med gender bias är väldigt utbrett och härstammar från många olika källor i utvecklingsprocessen. Att gender bias uppstår och kommer från många olika delar i systemet gör inte bara att det är svårt att hantera, utan också att det är svårt att lägga ansvar på en eller ett fåtal specifika källor. Att alla i viss mån är ansvariga medför risk att ingen faktiskt tar på sig ansvaret. Tanken bekräftas då en av respondenterna berättar om organisationer som när de inser att snedvridningen kommer från data och inte från det som de själva skapat inte längre anser sig vara ansvariga för de snedvridna resultaten. Men vem är i så fall ansvarig? Vem ansvarar för att data ser ut som det gör? Och ännu viktigare – vad kan göras för att nå förändring? För att nå dit krävs omställning på många plan – och en förutsättning för den är att erkänna, lyfta och diskutera att problemet finns till att börja med.

Ytterligare en dimension av ansvarsutkrävande läggs till den speciella struktur en konsultfirma har. Eftersom kunden är den som sätter ramarna för ett projekt är det i viss mån upp till kunden att bedöma könsdiskriminering som ett viktigt perspektiv och central variabel ha med i processen. Det finns en risk att etiska aspekter undermineras eftersom ansvar för att ta hänsyn till dem ur båda aktörers synvinkel ligger på den andra parten.

Att det inte finns en uppenbar ansvarig till att det ser ut som det gör kan å andra sidan göra problemet lättare att ta i, eftersom det kan uppenbara sig även när syftet varit något helt annat och ofta utan att en person blir direkt ansvarig. Den könsdiskriminering som

uppenbarligen alltid funnits i det moderna samhället – som att kvinnor nämns mer sällan i texter eller representeras annorlunda än män – blir konkretiserad och på sätt och vis lättare att ta på i och med användningen av AI. Det går inte längre att ignorera den diskriminering som finns eftersom resultaten är så tydliga, vilket är något som flera intervjupersoner ser som en möjlighet till förståelse, förändring och förbättring.

6.11 AI som katalysator för debatt kring gender bias

De problem med gender bias som AI-system uppdagar är inte nya – snarare tvärtom. I de fall AI-algoritmer visat på könsdiskriminering bekräftar resultaten i många fall bara helt enkelt hur det ser ut i samhället idag. AI-algoritmen föreslår vita män till chefspositioner framför kvinnor – för att det är så det ser ut idag. Bildigenkänning placerar kvinnor i köket och män på tennisplanen – för att det är stereotyper som många i vårt samhälle har idag. Röstassistenter ger avböjande, otillräckliga eller ursäktande svar på verbala sexuell trakasserier eftersom sexuella trakasserier mot kvinnor idag är normaliserat – både bland förövare och offer. Men som flertalet studier påpekar och varnar för finns det risk att den könsdiskriminering som finns idag inte bara speglas, utan även förstärks i och med utveckling och användning av AI.

Problemen med gender bias har på sätt och vis alltid varit närvarande, men kanske har det blivit än tydligare nu att könsdiskriminering existerar och påverkar människor i vardagen. Obalansen syns i både dataset och resultat och det går inte längre att blunda för fenomenet. Människor har alltid till varit mer eller mindre biased och fördomar har alltid i viss utsträckning påverkat beslutfattande. Men frågan är om, i och med spridningen och den utbredda användningen av den i raden senaste allmänteknologin artificiell intelligens, fler kommer att bli medvetna om bias generellt och gender bias specifikt eftersom det speglas av en teknik som används varje dag? Flera av personerna som intervjuats menar att detta är en chans att uppmärksamma problemen som finns och en möjlighet att diskutera varför det ser ut som det gör.

Produkterna från AI-industrin påverkar redan miljontals liv och finns i ett stort spann av sektorer. Att ta itu med mångfaldsfrågor är därför inte bara i teknikindustrins intresse utan något centralt för alla vars liv påverkas av AI-verktyg och tjänster. Kan AI i så fall verka som en katalysator för genusdebatten? Eftersom både AI och gender bias inom AI är nya områden är det svårt att göra en jämförelse huruvida just AI-fältet blivit mer medvetet sedan det visade sig att både dataset och resultat till stor del är diskriminerande mot något av könen. Av både litteraturstudie och djupintervjuer att döma kan dock ett mönster av ökad medvetenhet kring frågor som rör gender bias förstås, och kanske beror det på den nya infallsvinkel som gender bias i AI-system ger på debatten. Som diskuteras i bakgrunden så har teknikutveckling till stor del och under lång tid utgått från en manlig norm. Att AI fått stor uppmärksamhet för att vara diskriminerande mot vissa grupper (oftast icke-vita och icke-män) kan kanske leda till förändring på ett större plan. Mer data och fler fall där det gått fel ger upphov till fler diskussioner och med dem förhoppningsvis fler insikter och mer kunskap. AI har inget

filter i termer av att den har en uppfattning om vad som är politiskt korrekt eller inte. Den kan förstärka och visa på snedvridningar hos individer som individen inte ens visste att hen hade, och kanske är det den insikten som behövs för förändring.

6.12 Metodologiskt angreppssätt

I denna studie har en litteraturstudie och en empirisk undersökning i form av djupintervjuer gjorts. Initialt fanns incitament att i den empiriska delen jämföra uppfattningen och behandlingen av gender bias i AI-utveckling mellan män och kvinnor samt mellan chefer och utvecklare. Tanken att undersöka fenomenet från dessa olika aspekter fanns från start och dess relevans bekräftades av litteraturstudien som genomfördes. Ett större antal intervjuer och en större spridning vad gäller befattning och kön är således något som skulle kunna utvecklas. Då det på det undersökta företaget inom området för AI-utveckling inte arbetade varken många kvinnor eller många utvecklare visade sig detta vara svårt att genomföra. Då detta är så som verkligheten ser ut och ett syfte med den empiriska undersökningen också var att få en bild av nuläget i realiteten ses detta dock egentligen inte som ett problem.

Vidare fanns det initialt en förhoppning att undersöka hur AI-utveckling går till i praktiken, men inte heller det var möjligt eftersom arbetet med AI-utveckling är relativt nytt på KPMG Sverige. KPMG Danmark har kommit betydligt längre i sitt AI-arbete och av den anledningen intervjuades personer som är involverade i AI-utvecklingen på KPMG Danmark, men det var inte möjligt att undersöka djupare än så. Det hade därför möjligen varit att föredra att genomföra studien på ett företag som är mer specialiserat på AI-utveckling. Det hade dock blivit en annan typ av frågeställning eftersom en del av syftet med att undersöka just en konsultfirma likt KPMG är dess förankring i generell samhällsutveckling. Att få inblick i hur KPMG Danmark arbetar blev en indikation på hur KPMG Sverige kan komma att arbeta inom en inte alltför avlägsen framtid, vilket är en intressant, om än inte på förhand planerad, aspekt på frågeställningarna.

Under intervjuerna framhölls den viktiga roll kunden spelar i utformning av ramar för projekt, och det hade därför varit intressant att intervjua en eller flera av KPMG:s kunder.

7. Slutsatser

Denna studie visar att artificiell intelligens i flera fall speglar och förstärker befintliga snedvridningar mellan kön i samhället. Alla steg i designen av AI-algoritmer – från vem som beställer dem, till vem som utvecklar dem till vilken data som används för att utveckla och träna dem påverkas av och kan leda till könsdiskriminering.

Litteraturstudien visar på ett antal huvudsakliga faktorer som kan leda till könsdiskriminering i AI-system generellt. Dessa är: redan existerande snedvridning (hos institutioner och attityder i samhälle eller hos individ), snedvriden data, brist på mångfald i arbetskraft samt bristande förståelse av samband mellan teknik och genusanalys. Empiri lyfter ytterligare faktorer som är specifika för det undersökta företaget, nämligen kravställning från kunden samt tidspress. De lösningar som föreslagits i både litteraturstudien och under djupintervjuerna handlar dels om att hantera respektive identifierad orsaksfaktor men även att se problemet från ett helhetsperspektiv. Det är centralt att se sambanden mellan gender bias i AI-system och gender bias i samhället, samt att reflektera kring hur respektive orsaksfaktor beror av och korrelerar med andra faktorer. Essensen av resultaten är att det inte räcker att ändra någon av parametrarna om inte systemets struktur samtidigt ändras.

Både likheter och skillnader identifierades mellan teori och praktik. En stor mängd studier efterfrågar förändring på flera olika plan i samhället. Från djupintervjuer framkom att KPMG parallellt som de för diskussioner om etiska ramverk med mera ändå behöver anpassa sig efter de begränsningar som finns just nu. KPMG fokuserar på att hantera problem med könsdiskriminering i AI idag via de medel som de konkret kan påverka, exempelvis mångfald i teamen och datahantering. Då kravställning från kunden lyfts som central anledning till eventuell utebliven etisk aspekt förstås påverkan av denna som en viktig del i hanteringen av problemet.

Slutligen kan en generell slutsats om AI:s betydelse för gender bias i samhället dras. Förändring av förhållningssätt är något som återkommer gång på gång som svar på hur omställning till att börja arbeta mer aktivt med etisk utveckling av AI skulle kunna ske rent praktiskt. I många, men inte alla, fall skapas inte snedvridningen av AI per se utan är något som finns i indata och snarare tydliggörs av AI. Empiri lyfter explicit att den gender bias som kan ses i AI-system både ger bevis för könsdiskriminering som existerar men även möjlighet att öppna upp för diskussioner kring ämnet för att försöka hantera det. Att en majoritet av källorna i denna studie kommer från de två senaste åren visar också att ämnet kommit fram i ljuset på senare tid. Viljan att förstå varifrån snedvridningar och obalanser kommer istället för att bara försöka parera bort dem tyder på en vilja att förändra på ett strukturellt plan. Att istället för att ändra parametrar i systemet försöka ändra det mindset eller paradigm ur vilket systemet – dess mål, struktur, regler och parametrar – uppstår, är enligt Meadows (1999) nyckeln till faktiskt förändring och något som både litteraturstudie, djupintervjuer och analys av dessa visar behövs.

8. Fortsatta studier

Framtida arbete bör undersöka tillvägagångssättet för själva systemdesignen och studera AI-system i verkligheten. Undersökning av varför ett system designas på ett visst sätt, hur det är konstruerat, och vilkas intressen som bestämde de mätvärden i vilka dess framgång eller misslyckande bedöms är av intresse för att öka kunskapen inom området. Istället för att endast fokusera på att förbättra existerande dataset eller specifika algoritmer kan framtida arbete också undersöka hur diskriminering i samhället syns i dataseten, undersöka processen i vilken dataset är konstruerat och betänka hur kulturella normer och stereotyper var numrerade och representerade den tid då data skapades (West et al. 2019; Gebru et al. 2018).

Vidare är det av intresse att utvidga undersökningen av kön till att inte endast inkludera binära utan även icke-binära personer, samt att analysera frågan med bias generellt och gender bias specifikt utifrån ett intersektionellt perspektiv.

Slutligen vore det intressant att istället för att undersöka AI som paraplybegrepp analysera respektive undergrupp av AI och utreda olika undergruppers påverkan på och av gender bias och könsdiskriminering.

Referenser

Tryckta källor:

Alvesson, M. och Sköldberg, K., 2008, "Tolkning och Reflektion: vetenskapsteori och kvalitativ metod" 2. upplagan. Lund: Studentlitteratur

Bryman, A. och Bell, E., 2013, "Företagsekonomiska forskningsmetoder" 2. upplagan. Stockholm: Liber AB.

Caliskan, A., Bryson, J.J. och Narayanan, A., 2017, "Semantics derived automatically from language corpora contain human-like biases", *Science (New York, N.Y.),* vol. 356, no. 6334, pp. 183-186

Graziano, A.M. och Raulin, M.L., 2013, "Research methods: a process of inquiry" 8. upplagan. Boston: Pearson.

MacKenzie, D.A. och Wajcman, J, 1999, "The social shaping of technology", 2. upplagan, Open University Press, Maidenhead;Philadelphia, Pa;.

Van Otterlo, M., 2013, "A machine learning view on profiling." I: Hildebrandt M and de Vries K (eds) Privacy, Due Process and the Computational Turn-Philosophers of Law Meet Philosophers of Technology. Abingdon: Routledge

Weizenbaum, J., 1976," Computer power and human reason: from judgment to calculation", Freeman: San Francisco.

Yin, R., 2011, "Kvalitativ forskning från start till mål" Studentlitteratur AB: Lund.

Källor online:

Ashcraft, C., McLain, B., Eger, E., 2016, "Women in Tech: the facts", NCWIT's Workforce Alliance. Tillgänglig online: https://www.ncwit.org/sites/default/files/resources/womenintech_facts_fullreport_05132016.pdf (2019-10-13)

Bass, D. och Huet, E., 2017, "Researchers Combat Gender and Racial Bias in Artificial Intelligence" Bloomberg.com. Tillgänglig online: https://www.bloomberg.com/news/articles/2017-12-04/researchers-combat-gender-and-racial-bias-in-artificial-intelligence (2019-10-02)

Barocas, S., Hardt, M., Narayanan, A., 2019, "Fairness and Machine Learning", Fairmlbook.org, Tillgänglig online: http://www.fairmlbook.org (2019-10-30)

Baeza-Yates, R. 2018, "Bias on the web", Communications of the ACM, vol. 61, no. 6, pp. 54-61. (2019-11-02)

Bhattacherjee, A., 2012, "Social Science Research: Principles, Methods, and Practices" University of South Florida. Tillgänglig online: https://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=1002&context=oa_textbooks (2019-10-10)

Bogost, I., 2017, "'Artificial Intelligence' has become meaningless." The Atlantic, Tillgänglig online: https://www.theatlantic.com/technology/archive/2017/03/what-is-artificial-intelligence/518547/ (2019-10-20)

Borghans, L., Golsteyn, B., Heckman, J., Meijers, H., 2009. "Gender Differences in Risk Aversion and Ambiguity Aversion," Journal of the European Economic Association, MIT Press, vol. 7(2-3), pages 649-658, 04-05. (2019-10-25)

Brockman, J., 2019, "Ready for Robots? How to Think About the Future of AI Possible Minds: Twenty-five Ways of Looking at AI" Penguin Press. (2019-10-20)

Buolamwini J., och Gebru, T, 2018, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Conference on Fairness, Accountability and Transparency". Tillgänglig online: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf (2019-09-28)

Byrd, J.S., 2006, "Confirmation Bias, Ethics, and Mistakes in Forensics", Journal of Forensic Identification, vol. 56, no. 4, pp. 511-525. (2019-09-28)

Caldas-Coulthard, C. R. och Moon, R., 2010, "'Curvy, hunky, kinky': Using corpora as tools for critical analysis", University of Birmingham, UK. Tillgänglig online: https://journals-sagepub-com.ezproxy.its.uu.se/doi/pdf/10.1177/0957926509353843 (2019-10-02)

Charness, G., och Gneezy, U., 2012, "Strong Evidence for Gender Differences in Risk Taking", Journal of Economic Behavior and Organization, vol. 83, no. 1, pp. 50-58. (2019-10-27)

CHM (Computer History Museum), 2019, "John McCarthy", Tillgänglig online: https://www.computerhistory.org/fellowawards/hall/john-mccarthy/ (2019-10-03)

City of Boston, 2017, "Street Bump". Tillgänglig online: https://www.boston.gov/departments/new-urban-mechanics/street-bump (2019-10-04)

Costanza-Chock, S., 2018, "Design Justice, A.I., and Escape from the Matrix of Domination", Journal of Design and Science. Tillgänglig online: https://jods.mitpress.mit.edu/pub/costanza-chock (2019-10-01)

Daly, C., "AI Bias Isn't A Data Issue – It's A Diversity Issue". Tillgänglig online: https://aibusiness.com/ai-bias-diversity-payal-jain/ (2019-10-22)

de Vries K, 2010, "Identity, profiling algorithms and a world of ambient intelligence", Ethics and Information Technology 12(1): 71–85

Dhande, M., 2017, "What is the difference between AI, machine learning and deep learning", Tillgänglig online: https://www.geospatialworld.net/blogs/difference-between-ai%EF%BB%BF-machine-learning-and-deep-learning/ (2019-11-15)

Eynon, R. 2018, "Feminist perspectives on learning, media and technology: recognition and future contributions", Learning, Media and Technology, vol. 43, no. 1, pp. 1-2 (2019-10-20)

Europeiska kommissionen, 2019, “Ethics guideline for thrustworthy AI”. Tillgänglig online: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (2019-09-20)

Fessler, L., 2017, “We tested bots like Siri and Alexa to see who would stand up to sexual harassment”, Quartz. Tillgänglig online: https://qz.com/911681/we-tested-apples-siri-amazon-echos-alexa-microsofts-cortana-and-googles-google-home-to-see-which-personal-assistant-bots-stand-up-for-themselves-in-the-face-of-sexual-harassment/ (2019-10-02)

Finansministeriet og Erhvervsministeriet, 2019, “National strategi kor kunstig intelligens”. Tillgänglig online: https://www.regeringen.dk/media/6537/ai-strategi_web.pdf (2019-10-30)

Ford, H., och J. Wajcman. 2017. “‘Anyone Can Edit’, Not Everyone Does: Wikipedia’s Infrastructure and the Gender Gap.” Social Studies of Science 47 (4): 511–527. (2019-10-03)

Friedman, B. och Nissenbaum, H., 1996, “Bias in Computer Systems”, ACM Transactions on Information Systems (TOIS), vol. 14, nr. 3, p. 330–347. (2019-10-05)

Garbade, M. J., 2018, “A Simple Introduction to Natural Language Processing”, Medium. Tillgängling online: https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32 (2019-11-01)

Gender Working Group, United Nations Commission on Science and Technology for Development, 1995, “Missing Links - Gender equity in Science and Technology for Development”. Tillgänglig online: https://idl-bnc-idrc.dspacedirect.org/bitstream/handle/10625/17228/IDL-17228.pdf?sequence=1 (2019-10-30)

Genpact, 2019, “AI 360: insights from the next frontier of business”. Tillgänglig online: https://www.genpact.com/downloadable-content/AI-360-Research.pdf (2019-09-30)

Gjengegal, K., 2019, “Cars are still designed for men” GenderResearch.no. Tillgänglig online: http://kjonnsforskning.no/en/2019/06/cars-are-still-designed-men (2019-10-10)

Google, 2018, “Google 2018 Diversity Annual Report”. Tillgänglig online: https://static.googleusercontent.com/media/diversity.google/sv//static/pdf/Google_Diversity_annual_report_2018.pdf (2019-10-15)

Hajian, S., Bonchi, F. och Castillo, C., 2016, "Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining", ACM, , pp. 2125. (2019-10-15)

Hao, K., 2019, “This is how AI bias really happens - and why it’s so hard to fix”, MIT Technology Review. Tillgänglig online: https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/ (2019-10-30)

Harrison, D., 2016, “Deborah Harrison, Editorial Writer, Cortana - RE•WORK Virtual Assistant Summit”, YouTube. Tillgänglig online: https://www.youtube.com/watch?v=-WcC9PNMuL0(2019-10-20)

Harquail, 2012, “Add Women and Stir” Won’t Keep Women In Tech”. Tillgänglig online: http://authenticorganizations.com/harquail/2012/05/16/add-women-and-stir-wont-keep-women-in-tech/#sthash.kX7ZLf3b.R5JxM8N8.dpbs (2019-10-30)

Hendricks, L. A, Burns, K., Saenko K., Darrell1 T., Rohrbach, A., 2018, “Women Also Snowboard: Overcoming Bias in Captioning Models”, UC Berkeley, Boston University. Tillgänglig online: http://openaccess.thecvf.com/content_ECCV_2018/papers/Lisa_Anne_Hendricks_Women_also_Snowboard_ECCV_2018_paper.pdf (2019-09-23)

Howard A., Borenstein, J., 2017, “The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity”, Springer Science+Business Media B.V. (2019-10-02)

Howcroft, D. och Rubery, J., 2019, “‘Bias in, Bias out’: gender equality and the future of work debate”, Labour & Industry: a journal of the social and economic relations of work, 29:2, 213-227. (2019-10-02)

Hunt, E., 2016, “Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter”, The Guardian. Tillgänglig online: https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter?CMP=twt_a-technology_b-gdntech (2019-10-02)

Jain, S., 2006, “Injury: The Politics of Product Design and Safety Law in the United States”. Princeton: Princeton University Press. (2019-10-02)

Jianakoplos, N. A. och Bernasek, A., 1998, "Are women more risk averse?", Economic Inquiry, vol. 36, no. 4, pp. 620-630. (2019-10-27)

Jovanovic, B. och Rousseau, P., 2005, “General Purpose Technologies”, finns i Handbook of Economic Growth, Volume 1B. Tillgänglig online: 10.1016/S1574-0684(05)01018-X Tillgänglig online: https://www.nyu.edu/econ/user/jovanovi/JovRousseauGPT.pdf (2019-10-10)

Jönköpings-Posten, 2018, “Minister vill ha kvinnlig krockdocka - oftast testas bilar bara på män”. Tillgänglig online: https://www.jp.se/article/minister-vill-ha-kvinnlig-krockdocka-oftast-testas-bilar-bara-pa-man/ (2019-09-26)

Kanze, D., Huang, L., Conley, M.A. och Higgins, E.T., 2018, "We Ask Men to Win and Women Not to Lose: Closing the Gender Gap in Startup Funding", Academy of Management Journal, vol. 61, no. 2, pp. 586-614. (2019-09-26)

Kolhatkar, S., 2017, “Discrimination problem”, The New Yorker. Tillgänglig online: https://www.newyorker.com/magazine/2017/11/20/the-tech-industrys-gender-discrimination-problem (2019-09-20)

KPMG, 2019, “Navigating bias and supremacy in artificial intelligence (AI)”. Tillgänglig online: https://home.kpmg/au/en/home/insights/2019/01/navigating-bias-supremacy-artificial-intelligence.html (2019-10-15)

Kuczmarski, J., 2018, “Reducing gender bias in Google Translate”. Tillgänglig online: https://www.blog.google/products/translate/reducing-gender-bias-google-translate/ (2019-10-15)

Lai, C. K. och Banaji, M. R., 2019, “The psychology of implicit intergroup bias and the prospect of change.” In D. Allen och R. Somanathan Difference without Domination: Pursuing Justice in Diverse Democracies. Chicago, IL: University of Chicago Press. (2019-10-15)

Leavy, S., 2018, "Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning". (2019-09-20)

Lee, N. T., Resnick, P., Barton, G., 2019, “Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms”, Center for Technology Innovation, Brookings. Tillgänglig online: https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/#footnote-7 (2019-10-01)

Lewis, J., 2000, “GOFAI”, Tillgänglig online: https://www.cs.swarthmore.edu/~eroberts/cs91/projects/ethics-of-ai/sec3_1.html (2019-10-20)

Mahendra, R., 2019, “AI is the new electricity”, Pöyry Management Consulting. Tillgänglig online: https://www.smart-energy.com/industry-sectors/new-technology/ai-is-the-new-electricity/ (2019-11-02)

Mayer, D., 2018, “Amazon Reportedly Killed an AI Recruitment System Because It Couldn’t Stop the Tool from Discriminating Against Women”, Fortune. Tillgänglig online: https://fortune.com/2018/10/10/amazon-ai-recruitment-bias-women-sexist/ (2019-09-27)

McCarthy, J., 2007, “What is artificial intelligence?”, Computer Science Department

Stanford University Stanford, CA 94305. Tillgänglig online: http://www-formal.stanford.edu/jmc/whatisai.pdf (2019-10-20)

McKinsey Podcast, 2019, “The ethics of artificial intelligence”. (2019-09-25)

Meadows, D., 1999, “Leverage Points: Places to Intervene in a System”, The Sustainability Institute. (2019-09-27)

Medium, 2018, “Racial Bias and Gender Bias Examples in AI systems”. Tillgänglig online: https://medium.com/thoughts-and-reflections/racial-bias-and-gender-bias-examples-in-ai-systems-7211e4c166a1 (2019-10-02)

Mehrabi N, Morsatter F, Saxena N, Lerman K, & Galstyan A, 2019, “A Survey on Bias and Fairness in Machine Learning”. Tillgänglig online: https://arxiv.org/pdf/1908.09635.pdf

(2019-09-20)

Melendez, S., 2018, “Uber driver troubles raise concerns about transgender face recognition”, Fast Company. Tillgänglig online: https://www.fastcompany.com/90216258/uber-face-recognition-tool-has-locked-out-some-transgender-drivers (2019-09-26)

Microsoft Gender Case, Senast uppdaterad 2019, “Microsoft Gender Discrimination Class Action Lawsuit”. Tillgänglig online: https://microsoftgendercase.com/ (2019-09-20)

Mittelstadt, B. D., Allo P., Taddeo M., Wachter S., Floridi L., 2016. “The ethics of algorithms: Mapping the debate”, Big Data & Society. Tillgänglig online: https://journals.sagepub.com/doi/pdf/10.1177/2053951716679679 (2019-09-24)

MMC Ventures, 2019, The State of AI: Divergence. Tillgänglig online: https://www.mmcventures.com/wp-content/uploads/2019/02/The-State-of-AI-2019-Divergence.pdf (2019-10-18)

Morgan, G., Sturdy, A., Frenkel, M., ,2019, “The Role of Large Management Consultancy Firms in Global Public Policy”, Finns i: The Oxford Handbook of Global Policy and Transnational Administration. (2019-10-30)

Mullane, M., 2018, “Eliminating bias from algorithms”, International Electrotechnical Commission, Tillgänglig online: https://iecetech.org/Technical-Committees/2018-06/Eliminating-bias-from-algorithms (2019-10-15)

Nationalencyklopedin, “*kvalitativ metod*”. Tillgänglig online: http://www.ne.se.ezproxy.its.uu.se/uppslagsverk/encyklopedi/lång/kvalitativ-metod (2019-10-10)

Nesta, 2019, “Gender Diversity in AI Research”. Tillgänglig online: https://media.nesta.org.uk/documents/Gender_Diversity_in_AI_Research.pdf (2019-10-13)

Nickerson, R., 1998, “Confirmation Bias: A Ubiquitous Phenomenon in Many Guises” in Review of General Psychology. (2019-10-11)

O'Donnell, S., Condell, S. och Begley, C.M., 2004, "‘Add Women & Stir’—The Biomedical Approach to Cardiac Research", European Journal of Cardiovascular Nursing, vol. 3, no. 2, pp. 119-127. (2019-10-15)

Olteanu A., Castillo C., Diaz F., Kıcıman, E., 2019., “Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries”. Tillgänglig online: https://doi.org/10.3389/fdata.2019.00013 (2019-09-20)

Powles, J., 2018, “The Seductive Diversion of ‘Solving’ Bias in Artificial Intelligence”, Medium. Tillgänglig online: https://medium.com/s/story/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53 (2019-10-02)

Prates, M.O.R., Avelar, P.H., Lamb, L.C., 2019 “Assessing gender bias in machine translation: a case study with Google Translate”. (2019-09-26)

Rottingen, J-A., 2018, “Gender dimension should be broadly integrated in research” GenderResearch.no. Tillgänglig online: http://kjonnsforskning.no/en/2018/12/gender-dimension-should-be-broadly-integrated-research (2019-10-10)

Salminen-Karlsson, M, 2011, “The Problem in the Eye of the Beholder: Working with Gender Reforms in Computer Engineering”, International Journal of Gender, Science and Technology, 3 (2) 445-459.

Schiebinger, L., 2014, “Scientific research must take gender into account”. Nature 507(7490), 9 (2019-10-10)

Schiebinger, L., Klinge, I., Sánchez de Madariaga, I., Paik, H. Y., Schraudner, M., Stefanick, M. (Eds.) (2011-2018). Gendered Innovations in Science, Health & Medicine, Engineering and Environment. Tillgänglig online: https://genderedinnovations.stanford.edu/case-studies/nlp.html#tabs-2 (2019-10-01)

Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S., Vertesi, J. 2019, "Fairness and abstraction in sociotechnical systems" (2019-10-30)

Shashkevich, A., 2019, “Stanford researcher examines earliest concepts of artificial intelligence, robots in ancient myths”, Stanford News Service. Tillgänglig online: https://news.stanford.edu/2019/02/28/ancient-myths-reveal-early-fantasies-artificial-life/ (2019-10-02)

Shoham, Y., Perrault, R., Brynjolfsson, E., Clark, J., Manyika, J., Niebles, J. C., Lyons, T., Etchemendy J., Grosz, B, and Bauer, Z., 2018, "The AI Index 2018 Annual Report”, 2018 AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, Stanford, CA. Tillgänglig online: http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf (2019-10-13)

Snow, J., 2018, “‘We’re in a diversity crisis’: cofounder of Black in AI on what’s poisoning algorithms in our lives”, MIT Technology Review, Tillgänglig online: https://www.technologyreview.com/s/610192/were-in-a-diversity-crisis-black-in-ais-founder-on-whats-poisoning-the-algorithms-in-our/?utm_source=facebook.com&utm_medium=social&utm_content=2018-03-11&utm_campaign=Technology+Review&fbclid=IwAR3Dv6jHhiWMdfnN4PorzNnQnfUSMRwrlcKQ7oSAuYlcLYIumBrUa0KGehY (2019-10-04)

Suresh, Harini och V. Guttag, John. 2019. “A Framework for Understanding Unintended Consequences of Machine Learning”. Tillgänglig online: https://arxiv.org/pdf/1901.10002.pdf (2019-09-20)

Tatman, R., 2016., “Google’s speech recognition has a gender bias. Making Noise and Hearing Things”. Tillgänglig online: https://makingnoiseandhearingthings.com/2016/07/12/googles-speech-recognition-has-a-gender-bias/ (2019-09-15)

Temm, T., 2008, “If You Meet the Expectations of Women, You Exceed the Expectations of Men: How Volvo Designed a Car for Women Customers and Made

World Headlines.” I Schiebinger, L. (Ed.), Gendered Innovations in Science and Engineering, pp. 131-149. Stanford: Stanford University Press.

Bolukbasi, T., Chang, K-”., Zou, J., Saligrama, V., Kalai, A., 2016, “Man is to computer programmer as woman is to homemaker? debiasing word embeddings.”. Finns i Advances in Neural Information Processing Systems. (2019-09-16)

UNESCO, 2019, “I’d blush if I could - closing gender divides in digital skills through education”, EQUALS. Tillgänglig online: https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1 (2019-09-05)

West, S.M., Whittaker, M. and Crawford, K., 2019, “Discriminating Systems: Gender, Race and Power in AI. AI Now Institute”. Tillgänglig online: https://ainowinstitute.org/discriminatingsystems.html (2019-09-23)

World Economic Forum, 2018, “Global Gender Gap Report 2018”. Tillgänglig online: http://www3.weforum.org/docs/WEF_GGGR_2018.pdf (2019-10-02)

World Economic Forum, 2019a, “This is why AI has a gender problem”. Tillgänglig online: https://www.weforum.org/agenda/2019/06/this-is-why-ai-has-a-gender-problem/ (2019-10-15)

World Economic Forum, 2019b, “AI is showing signs of serious pro-male bias, study finds”. Tillgänglig online: https://www.weforum.org/agenda/2019/08/ai-is-in-danger-of-becoming-too-male-new-research/ (2019-09-21)

Wikipedia, 2019, Tillgänglig: Wikipedia.org (2019-10-03)

Zhang, L., Wu Y., Wu,X.; 2017. “A Causal Framework for Discovering and Removing Direct and Indirect Discrimination”, In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17. 3929–3935. (2019-09-23)

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K-W., 2017, “Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints”, Cornell University. (2019-10-10)

Appendix A

Intervjumall för semistrukturerade intervjuer på KPMG HT 2019

Generellt

- Vad är din roll?

Data

- Hur mycket reflekteras över de data som tas in och den data som tas ut? Litar ni på data? Vad görs för att ifrågasätta/kontrollera datas korrekthet?
- Hur avgörs vilka subgrupper som behövs tas hänsyn till/finnas representerade?
- Hur ofta vet ni hur träningsdata är insamlat/ kommer ifrån?
- Finns det entiteter/parametrar som ni inte inkluderar?
- Vad kan göras för att parera bias i data?
- Finns det några källor är mer benägna att leda till förstärkning av bias - t.ex. bilder eller anonyma data? Något man bör se upp med?

Gender bias

- Hur mycket pratar ni om gender bias i relation till AI-system?
- Hur definierar du gender bias/könsdiskriminering?
- Upplever du att algoritmer kan vara diskriminerande? Mot vilka? (kön/ras/osv)
- På vilket sätt skulle du säga att AI system kan leda till könsdiskriminering (om du tycker att det kan det)? Vilka faktorer är det som bidrar?
- Vilka delar i implementation/en produktion kan leda till gender bias? Hur tar ni hänsyn till/arbetar med det?
- Vad gör ni för att undvika bias?
- Finns det något mer som borde göras men som inte görs?
- Finns det någon kunskap om gender bias du önskar du hade men som du inte har?
- Har arbetet med/mot diskriminering ändrats på senare tid?
- Vad/vilka faktorer (faktorer = ex diversitet i teamen, ramverk för etisk utveckling, logiska resonemang i algoritmerna, OSV) är viktiga för att undvika gender bias?

Utvärdering

- Hur utvärderas AI-system/implementationer som görs idag?
- Vilka områden är det som utvärderas (effektivitet, resultat, diskriminering, osv)?
- Utvärderas det någonsin om det finns gender bias i applikationen/implementationen?
    - Hur?
    - Finns det några verktyg? Skulle det behövas fler verktyg?

Beslut

- Hur går ett beslut till i ett AI-projekt?
- Hur styrt är ert arbete? Hur mycket frihet har man att ta egna beslut i utvecklingsprocessen?
- Vilka delar av processen äger ni (från vad som ska göras av vem till vilken data som ska användas och hur modellen ska byggas/tränas)? Varifrån kommer besluten ni inte äger?
- Vem bestämmer vad det är som ska utvecklas/göras?
- Hur avgörs vilka användare slutprodukten ska användas på?

Diversifiering av team

- Hur jobbar ni med diversifiering av team?
- Hur tillsätts teamen som arbetar med AI?
- Vad ser ni för värde med att ha ett diversifierat team?

Övrigt

- Är AI-system mer eller mindre biased än status quo (människor)? Kan algoritmer lära bort sin bias? Kan människor?