UPPSALA
UNIVERSITET

UPTEC STS 19015
Examensarbete 30 hp
Juni 2019

Ramverk för att motverka algoritmisk snedvridning

Clara Engman
Linnea Skärdin

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

Framework for Preventing Algorithmic Bias

Clara Engman & Linnea Skärdin

In the use of the third generation Artificial Intelligence (AI) for the development of products and services, there are many hidden risks that may be difficult to detect at an early stage. One of the risks with the use of machine learning algorithms is algorithmic bias which, in simplified terms, means that implicit prejudices and values are comprised in the implementation of AI. A well-known case is Google Photo's image recognition, which identified black people as gorillas. The purpose of this master thesis is to create a framework with the aim to minimise the risk of algorithmic bias in AI development projects. To succeed with this task, the project has been divided into three parts. The first part is a literature study of the phenomenon bias, both from a human perspective as well as from an algorithmic bias perspective. The second part is an investigation of existing frameworks and recommendations published by Facebook, Google, AI Sustainability Center and the EU. The third part consists in an empirical contribution in the form of a qualitative interview study which has been used to create and adapt an initial general framework. Ultimately, the framework has been revised according to the interview results.

The framework was created using an iterative methodology where two whole iterations were performed. The first version of the framework was created using insights from the literature studies as well as from existing recommendations. To validate the first version, the framework was presented for one of Cybercom's customers in the private sector, who also got the possibility to ask questions and give feedback regarding the framework. The second version of the framework was created using results from the qualitative interview studies with machine learning experts at Cybercom. As a validation of the applicability of the framework on real projects and customers, a second qualitative interview study was performed together with Sida - one of Cybercom's customers in the public sector. Since the framework was formed in a circular process, the second version of the framework should not be treated as constant or complete. The interview study at Sida is considered the beginning of a third iteration, which in future studies could be further developed.

Handledare: Victoria Jonsson
Ämnesgranskare: Anders Arweström Jansson
Examinator: Elísabet Andrésdóttir
ISSN: 1650-8319, UPTEC STS 19015

Sammanfattning

Användningen av artificiell intelligens (AI) har tredubblats på ett år och anses av vissa vara det viktigaste paradigmskiftet i teknikhistorien. AI återfinns i affärsplaner och i olika former av verksamhetsutveckling hos många olika aktörer samt i många olika tillämpningar. Företag kapplöper om att implementera tekniken samt undersöka dess potential inom nya affärsområden. Även offentlig sektor hakar på tåget då Sveriges strategi för digitalisering antyder en ökad satsning på utbildning, forskning och innovation på området.

Det finns dock en baksida med den lovordade tekniken. Den rådande AI-kapplöpningen riskerar att underminera frågor om etik och hållbarhet, vilket kan ge förödande konsekvenser - inte minst för offentlig sektor som lyder under grundläggande principer som likabehandling och transparens. Artificiell intelligens har i flera fall visat sig avbilda, och till och med förstärka, befintliga snedvridningar i samhället i form av fördomar och värderingar. Detta fenomen kallas algoritmisk snedvridning (algorithmic bias). Ett välkänt exempel inträffade 2015 då Googles bildigenkänningsalgoritm klassificerade mörkhyade personer som gorillor. Detta fall, tillsammans med många andra, har höjt röster inom ämnesområdet och har även bidragit till en efterfrågan på ett universellt ramverk för att undvika denna typ av händelser - vilket har gett gensvar. Inom loppet av ett år har EU, Facebook, Google och ett svenskt initiativ kallat AI Sustainability Center publicerat varsitt ramverk med riktlinjer för hållbar eller pålitlig AI. Alla dessa ramverk är förhållandevis generella och ger inget konkret tillvägagångssätt för att minimera risken för algoritmisk snedvridning. Ramverken anses dessutom vara relativt bristfälliga gällande tillämpningsbarhet för mindre till medelstora företag.

Cybercom är ett medelstort IT-konsultbolag med en tydlig hållbarhetsprofil och erbjuder bland annat rådgivning inom hållbar digitalisering. Som ett steg i detta lanseras under våren 2019 en metod de kallar Sustainability by Design. Metoden ska hitta lösningar för att minimera risken att tekniken katalyserar negativa effekter. Om projektet visar sig involvera AI-utveckling behöver Cybercom därför en konkret metod för att undvika algoritmisk snedvridning. Denna studie syftar till att formulera en metod för att minimera risken att algoritmisk snedvridning uppstår och att anpassa den efter ett medelstort konsultbolag. Metoden, eller ramverket, formulerades i en iterativ process där första iterationen baserades på redan existerande ramverk och fallstudier i ämnet. En workshop genomfördes därefter för att utvärdera ramverket efterföljt av intervjuer med maskininlärningsexperter på Cybercom. Den andra iterationen fokuserade på anpassning till ett medelstort företag som Cybercom och utformades i enlighet med resultaten från intervjuerna. En presentation genomfördes senare för en av Cybercoms kunder för att utvärdera ramverket. En tredje och sista iteration påbörjades med analys

av ramverkets tillämpbarhet på Cybercoms kunder inom offentlig sektor, vilken utfördes med hjälp av intervjuer.

Den kvalitativa fallstudien och studien av EU:s, AI Sustainability Centers, Facebooks och Googles ramverk för hållbar/pålitlig AI visade på sju huvudsakliga faktorer som bidrar till algoritmisk snedvridning. Dessa är: underliggande snedvridningar i samhället; problem med målgruppens användande av tekniken; bildigenkänning; obalanserade träningsdata; icke-transparent träningsdata; användning av riskfyllda entiteter (parametrar så som kön, etnicitet och religion) och användarstyrda indata. Dessa studier utrönande även ett antal metoder för att undvika algoritmisk snedvridning: Granskningsmekanismer, teknisk robusthet och säkerhet, integritet och datastyrning, transparens, tydlig definition av Fairness eller rättvisa, förklaringsmetoder, diversifierade team, diversifierade träningsdata, korrekt användarstyrda indata, eliminering av riskfyllda parametrar, utfallsvalidering för subgrupper samt löpande testning.

I enlighet med resultat från intervjuer med maskininlärningsexperter på Cybercom implementerades alla dessa punkter i ramverket bortsett från förklaringsmetoder och eliminering av riskfyllda parametrar samt med revideringar för diversifierade team och utfallsvalidering för subgrupper. Förklaringsmetoder ansågs vara för teknik-specifikt för att implementeras i ramverket och dessutom inte applicerbart på alla typer av AI. Eliminering av riskfyllda parametrar i ett tidigt stadium visade sig i intervjuer med maskininlärningsexperter på Cybercom inte vara förenlig med deras arbetssätt. Diversifierade team föreföll sig vara svårt för Cybercom att uppnå, delvis på grund av en ovilja att kvotera in mångfald och delvis på grund en homogenitet bland de som väljer att utbilda sig inom ämnet. Istället implementerades en dedikerad mångfaldsgrupp kallad *Fairness Implementation Specialists* i ramverket. Denna grupp har som uppgift att genomföra inledande undersökningar med avseende på de identifierade riskfaktorerna och bör innehålla kompetenser från olika specialistområden, som samhällskunskap, beteendevetenskap och humaniora. Utfallsvalidering för alla subgrupper har modifierats till att enbart inkluderas då träningsdatan inte går att undersöka eller i de fall kvaliteten och tillgången på träningsdata inte är tillräcklig och inte går att syntetisera. Detta då det i intervjuer med Cybercom visade sig vara en tidskrävande och kostsam metod.

Ramverkets tillämpbarhet analyserades slutligen på två av Cybercoms kunder - en inom privat sektor och en inom offentlig sektor. Både inom privat och offentlig sektor förutspås detta ramverk ha god potential att tillämpas, då intervjuresultaten tyder på en efterfrågan för denna typ av metoder. Både privat och offentlig sektor utstrålar dessutom en vilja att utveckla artificiell intelligens inom en snar framtid. Förutsättningen för automatisering och implementering av AI ser dock väldigt olika ut mellan olika aktörer.

Förord

Som det sista momentet på civilingenjörsprogrammet i System i teknik och samhälle vid Uppsala universitet har vi utfört detta examensarbete i samarbete med Cybercom Group. Vi vill tacka alla involverade på Cybercom för ett stort engagemang och för en enorm hjälpsamhet. Ett stort tack vill vi rikta till vår handledare Victoria Jonsson för värdefulla tips, idéer och kontakter till viktiga intervjupersoner samt alla andra på Cybercom Advisory som har varit delaktiga i vår process och bidragit med stöttning och idéer.

Vi vill också rikta ett särskilt stort tack till vår fantastiska ämnesgranskare Anders Arweström Jansson på Uppsala universitet för ovärderlig vägledning, uppmuntran, tips och idéer. Sist men inte minst vill vi tacka alla intervjudeltagare som har givit oss en del av sin värdefulla tid och bidragit stort till detta examensarbete.

Trevlig läsning!

Clara Engman och Linnea Skärdin

Begreppsordlista

Algoritm - En process eller en uppsättning regler som ska följas i beräkningar eller andra problemlösningsprocedurer.

Felfrekvens - Antalet fel som uppstått i förhållande till den totala populationen

Klustring - Ihopsättning av datapunkter med liknande egenskaper.

Metrik - Måttsystem eller standard för mätning.

Modell - En matematisk representation av en verklig process.

Open source - Mjukvara vars källkod är tillgänglig för alla att använda, modifiera och distribuera.

PoC - Proof of Concept. En prototyp som genomförs innan beslut tas om ett riktigt projekt ska initieras eller ej. PoC:en ämnar ge information om konceptets genomförbarhet och potential för implementation.

Program - En samling instruktioner som utför en specifik uppgift när den exekveras av en dator.

Uppkoppling – En översättning av engelskans connectivity, som innebär att mjuk- eller hårdvara kan kommunicera med andra enheter.

Snedvridning - En översättning av engelskans bias och innebär en avvikelse från det “sanna” värdet på grund av olika kognitiva eller statistiska faktorer.

Innehållsförteckning

1. Inledning

Efter att ha genomgått ett antal toppar och dalar i intresse för området har artificiell intelligens (AI) återigen sett dagens ljus. Teknikens användning har tredubblats på ett år och anses av vissa vara det viktigaste paradigmskiftet i teknikhistorien (MMC Venture, 2019). Bara ett av tio företag i Europa är villiga att vänta på att en AI-lösning ska implementeras i deras föredragna mjukvaruprodukter (MMC Venture, 2019). Även offentlig sektor deltar i denna AI-kapplöpning. Regeringens målsättning är att Sverige ska vara ledande i världen på att implementera digital teknik i den offentliga sektorn, där digitaliseringen ska användas som medel för att uppnå "effektivitet, konkurrenskraft, full sysselsättning samt ekonomiskt, socialt och miljömässigt hållbar utveckling" (Regeringskansliet, 2017). Då artificiell intelligens är en av de snabbast växande teknikerna i dagsläget har den en central roll i digitaliseringsprocessen. I Sveriges strategi för digitalisering skriver Näringsdepartementet (2018) därför att artificiell intelligens bör ses som ett hjälpmedel att utnyttja i största möjliga mån. De poängterar även att utbildning, forskning, innovation och inte minst infrastruktur för AI bör satsas på för att kunna ligga i framkant i utvecklingen.

Det finns många möjligheter med artificiell intelligens men det har också visat sig finnas många överhängande risker. Kapplöpningen riskerar att underminera frågor om etik och hållbarhet, vilket kan ge förödande konsekvenser - inte minst för offentlig sektor som lyder under grundläggande principer om likabehandling och transparens. Artificiell intelligens kan innebära stora risker även för privat sektor. 44% av de europeiska företagen föredrar att köpa AI istället för att utveckla den själv (MMC Venture, 2019), vilket kan innebära en minskad transparens bakom de beslut som den artificiella intelligensen tar. Uppkomsten av så kallad algoritmisk snedvridning (algorithmic bias) är en sådan risk som innebär att värderingar och fördomar implementeras i tekniken. Begreppet är något missvisande då algoritmen i sig inte nödvändigtvis är snedvriden ur ett statistiskt perspektiv, men blir snedvriden ur ett socialt perspektiv, då den reflekterar existerande social snedvridning hos individen eller samhället. Ett välkänt exempel på algoritmisk snedvridning inträffade 2015 då Googles bildigenkänningsalgoritm klassificerade mörkhyade personer som gorillor (Alciné, 2015). Detta fall, tillsammans med många andra, har bidragit till en efterfrågan på ett universellt ramverk för att undvika denna typ av händelser. Som gensvar har EU, Facebook, Google och ett svenskt initiativ kallat AI Sustainability Center publicerat varsitt ramverk med riktlinjer för hållbar eller pålitlig AI inom loppet av 24 månader. Alla dessa ramverk är förhållandevis generella och ger inget konkret tillvägagångssätt för att minimera risken för att algoritmisk snedvridning uppstår. Ramverken anses dessutom vara relativt bristfälliga gällande tillämpningsbarhet för mindre till medelstora företag.

Cybercom är ett medelstort IT-konsultbolag med huvudsakligt marknadsområde inom telekom, industri och offentlig sektor i Norden. Företaget har en tydlig hållbarhetsprofil och ger bland annat rådgivning inom hållbar digitalisering. Som ett steg i detta lanserar Cybercom en ny metod kallad Sustainability by Design. Metoden ska hjälpa kunden att besvara frågan om deras planerade IT-projekt kommer att katalysera negativa effekter och hitta lösningar för att minimera riskerna för dem. Om projektet visar sig involvera AI-utveckling behöver Cybercom använda sig av en konkret metod för att minimera risken för algoritmisk snedvridning. I denna studie formuleras en sådan metod, vilken sedan anpassas efter ett medelstort företag som Cybercom. Metoden, eller ramverket, formuleras i en iterativ process där första iterationen baseras på redan existerande ramverk och fallstudier i ämnet och den andra iterationen på intervjuer med Cybercom och deras kunder.

Rapportens första del ger relevant bakgrundsinformation om artificiell intelligens gällande dess historia, nutid och framtid samt information om Cybercom som företag. Denna del efterföljs av en genomgång av studiens teoretiska ramverk, vilken är uppdelad i mänskliga och algoritmiska snedvridningar. Denna del redogör även för fallstudier inom området och befintliga riktlinjer från EU, AI Sustainability Center, Facebook och Google för hållbar/pålitlig AI. Metoden med vilken studien är genomförd presenteras sedan efterföljt av en presentation av studiens resultat. I resultatdelen sammanfattas bakomliggande orsaker för algoritmisk snedvridning från fallstudierna, de befintliga riktlinjerna för hållbar/pålitlig AI, intervjudata samt ramverkets olika iterationer. Slutligen förs en diskussion kring ramverket och dess tillämpning efterföljt av slutsatser. Samtliga delar av examensarbetet har likafullt bidragits till och slutförts av bägge författarna.

1.1 Syfte och frågeställning

Denna studie har två syften. Studiens första syfte är att skapa ett ramverk för att minimera riskerna för algoritmisk snedvridning i AI-utveckling. Studiens andra syfte är att anpassa ramverket efter ett medelstort konsultbolag som Cybercom.

Studiens första syfte besvaras med hjälp av följande två frågeställningar:

- Vilka huvudsakliga faktorer orsakar algoritmisk snedvridning?
- Vilka metoder finns för att undvika algoritmisk snedvridning?

Studiens andra syfte besvaras med hjälp av ytterligare två frågeställningar:

- Hur arbetar Cybercom och deras kunder med AI-utveckling?
- Vilka förutsättningar för automatisering finns hos Cybercoms kunder?

1.2 Avgränsningar

Denna studie avgränsas till att formulera ett ramverk utefter välkända fallstudier och relevanta befintliga ramverk enligt information som finns tillgänglig idag. Då teknikutvecklingens framfart är explosionsartad, särskilt inom AI-utveckling, publiceras ständigt nytt material - vilket kan komma att ändra utgångspunkterna från vilka ramverket är utformat. Ramverket har dock utvecklats med en iterativ metod som tillåter ytterligare revideringar via framtida iterationer. Ramverket ska således inte betraktas som konstant eller färdigt, utan snarare som en grund för ett ramverk i förändring.

Studien avgränsas vidare till att undersöka två av Cybercoms kunder - en aktör i privat sektor och en aktör i offentlig sektor. Dessa två aktörer anses i dagsläget vara representativa för Cybercoms kunder då AI-utveckling generellt sett sker på en närmast experimentell nivå för kunderna och företaget självt. Det anses dock vara av stor vikt att inkludera både privat och offentlig sektor på grund av skillnader i förutsättningar dem emellan.

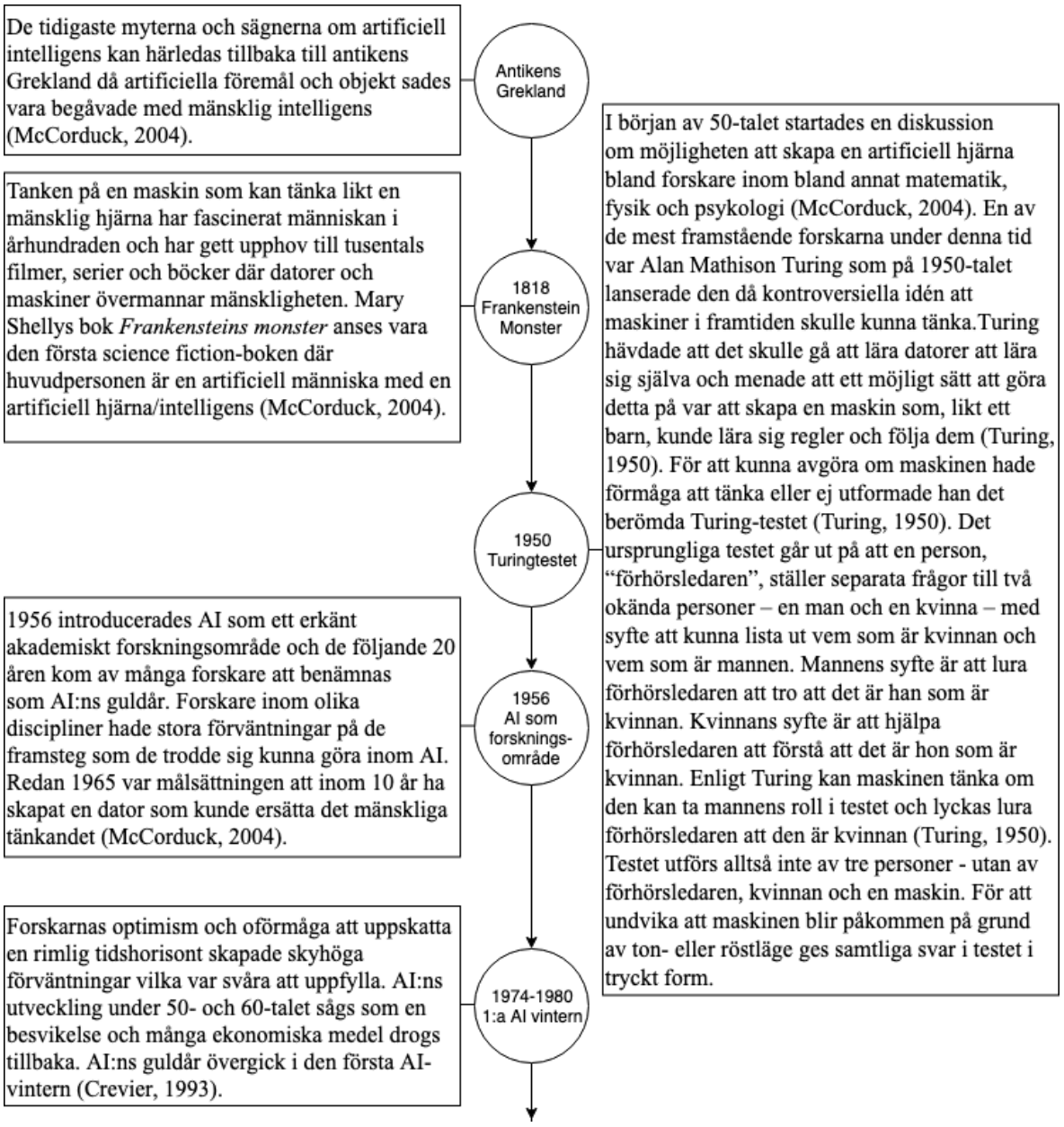
2. Bakgrund

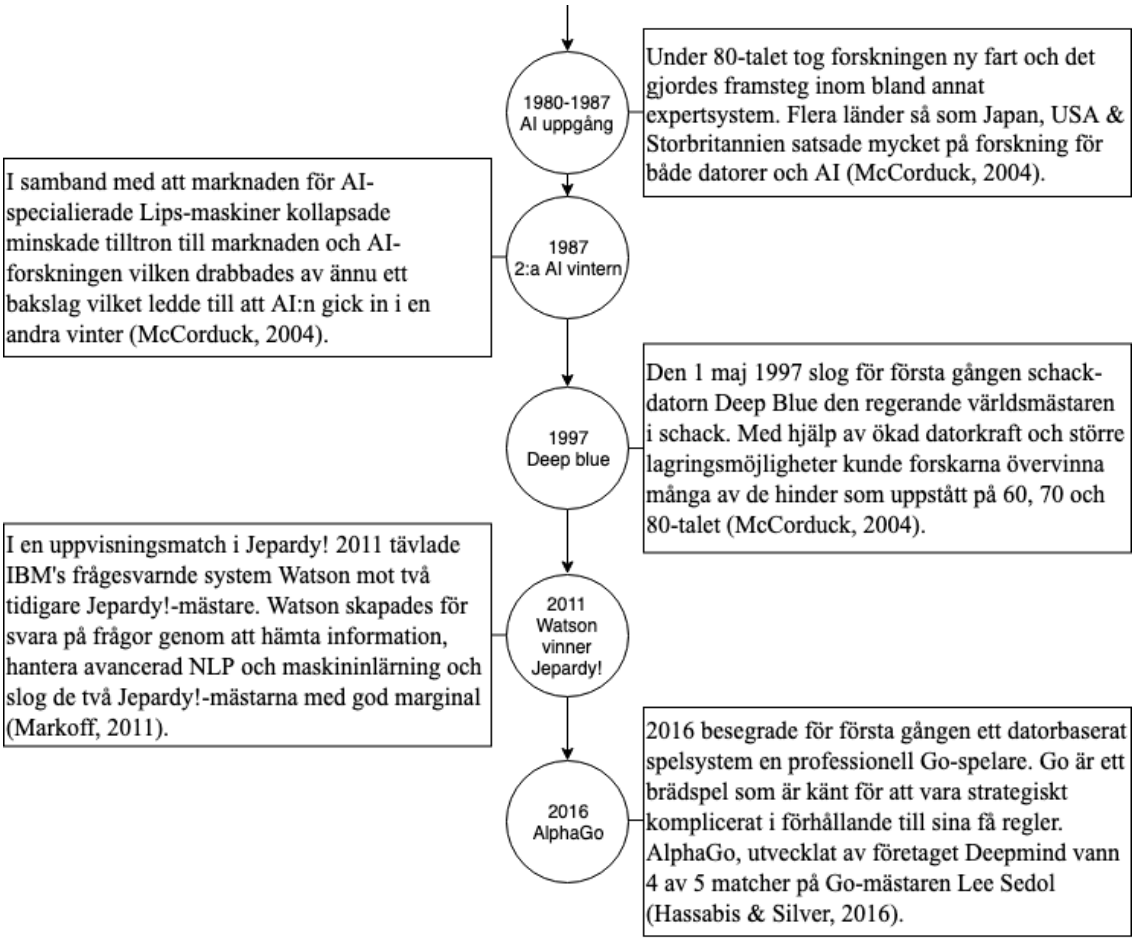
Detta avsnitt presenterar relevant bakgrundsinformation gällande begreppet artificiell intelligens (AI) och Cybercom som företag. För att få en djupare förståelse för artificiell intelligens ges inledningsvis en övergripande bild av begreppets historia. Denna efterföljs av en mer teknisk redogörelse för AI idag med en genomgång av metoderna bakom tekniken. Som avslutning till bakgrunden för artificiell intelligens presenteras även spekulationer kring teknikens framtid.

Därefter introduceras Cybercom som företag med en genomgång av deras affärsidé, deras kunder och den specifika metoden till vilken detta ramverk ska bidra. En djupdykning görs i offentlig sektor, då denna är en stor del av Cybercoms kunder och antas särskilja sig från privat sektor eftersom den lyder under ett antal principer. Under detta avsnitt presenteras även Sveriges digitaliseringsstrategi. Då digitalisering är ett brett begrepp med en tudelad innebörd redogörs detta slutligen för.

2.1 AI:ns historia

Nedan i figur 1 följer en tidslinje över de mest avgörande händelserna i AI:ns historia.


De tidigaste myterna och sägnerna om artificiell intelligens kan härledas tillbaka till antikens Grekland då artificiella föremål och objekt sades vara begåvade med mänsklig intelligens (McCorduck, 2004).
Antikens Grekland
Tanken på en maskin som kan tänka likt en mänsklig hjärna har fascinerat människan i århundraden och har gett upphov till tusentals filmer, serier och böcker där datorer och maskiner övermannar mänskligheten. Mary Shellys bok *Frankensteins monster* anses vara den första science fiction-boken där huvudpersonen är en artificiell människa med en artificiell hjärna/intelligens (McCorduck, 2004).
1818 Frankenstein Monster
I början av 50-talet startades en diskussion om möjligheten att skapa en artificiell hjärna bland forskare inom bland annat matematik, fysik och psykologi (McCorduck, 2004). En av de mest framstående forskarna under denna tid var Alan Mathison Turing som på 1950-talet lanserade den då kontroversiella idén att maskiner i framtiden skulle kunna tänka.Turing hävdade att det skulle gå att lära datorer att lära sig själva och menade att ett möjligt sätt att göra detta på var att skapa en maskin som, likt ett barn, kunde lära sig regler och följa dem (Turing, 1950). För att kunna avgöra om maskinen hade förmåga att tänka eller ej utformade han det berömda Turing-testet (Turing, 1950). Det ursprungliga testet går ut på att en person, "förhörsledaren", ställer separata frågor till två okända personer – en man och en kvinna – med syfte att kunna lista ut vem som är kvinnan och vem som är mannen. Mannens syfte är att lura förhörsledaren att tro att det är han som är kvinnan. Kvinnans syfte är att hjälpa förhörsledaren att förstå att det är hon som är kvinnan. Enligt Turing kan maskinen tänka om den kan ta mannens roll i testet och lyckas lura förhörsledaren att den är kvinnan (Turing, 1950). Testet utförs alltså inte av tre personer - utan av förhörsledaren, kvinnan och en maskin. För att undvika att maskinen blir påkommen på grund av ton- eller röstläge ges samtliga svar i testet i tryckt form.
1950 Turingtestet
1956 introducerades AI som ett erkänt akademiskt forskningsområde och de följande 20 åren kom av många forskare att benämnas som AI:ns guldår. Forskare inom olika discipliner hade stora förväntningar på de framsteg som de trodde sig kunna göra inom AI. Redan 1965 var målsättningen att inom 10 år ha skapat en dator som kunde ersätta det mänskliga tänkandet (McCorduck, 2004).
1956 AI som forsknings-område
Forskarnas optimism och oförmåga att uppskatta en rimlig tidshorisont skapade skyhöga förväntningar vilka var svåra att uppfylla. AI:ns utveckling under 50- och 60-talet sågs som en besvikelse och många ekonomiska medel drogs tillbaka. AI:ns guldår övergick i den första AI-vintern (Crevier, 1993).
1974-1980 1:a AI vintern



Figur 1. Tidslinje över AI:ns historia

2.2 AI idag

Idag är begrepp som artificiell intelligens (AI) mycket eftertraktade att använda i beskrivningen av allt från existerande simpla lösningar till komplexa robotar och självkörande bilar. Riskkapitalbolaget MMC Venture har tillsammans med banken Barclays undersökt användningen av begreppet och har slagit fast att 40% av startup-företagen i Europa som påstås vara AI-bolag i själva verket inte ägnar sig åt tekniken på ett sådant sätt att den kan sägas tillhöra deras kärnverksamhet (MMC Venture, 2019). Rapportens författare poängterar att det i de flesta fallen inte är företagen själva som har framställt sig som AI-bolag men att de aktivt har låtit bli att korrigera uppfattningen, eftersom AI uppfattas något prestigefyllt att ägna sig åt. I sin bok *AI: Its Nature and Future* (2016) skriver Margaret A. Boden (2016) att artificiell intelligens har två huvudsakliga mål - ett tekniskt och ett vetenskapligt. Det tekniska tar form i att ge datorer användbara uppgifter. Det vetenskapliga fokuserar på att finna svar angående mänskligheten och andra levande varelser.

Det finns tyvärr ingen universellt accepterad definition av vad artificiell intelligens är. Definitionen finns i många olika exemplar och är viktig att tydliggöra. MMC Venture (2019) definierar AI som en generell term för all hård- eller mjukvara som visar

beteende som uppfattas som intelligent. Enligt MMC Ventures rapport är de mest populära AI-tillämpningarna chattbotar, processautomation och analys av bedrägerier.

AI delas vanligen grovt in i två evolutionära steg: svag (eller begränsad) AI och stark (eller obegränsad) AI (Wirth, 2018). Benämningarna syftar till hur stark tekniken är i förhållande till den mänskliga intelligensen. Stark AI är datorprogram som inte är begränsade till specifika uppgifter och som i princip kan tänka som människor. Tolkningen bygger på antagandet att all medvetenhet är framkallad från beräkningar som utförs i hjärnan och därmed även kan utföras av en dator (Colman, 2015). Denna typ av AI existerar inte i dagsläget (Wirth, 2018). Svag AI, däremot, är begränsad till specifika uppgifter och kan inte användas för andra åtaganden utan att modifieras eller tränas om (Wirth, 2018). Begreppet är något missvisande då svag AI inte på något sätt är enkel att utveckla och inte heller nödvändigtvis svag inom sitt specialistområde. DeepBlue (McCorduck, 2004), Watson (Markoff, 2011) och DeepMind (Hassabis & Silver, 2016) är tydliga exempel på detta. Den snabba utvecklingen inom AI har dock krävt en mindre grov indelning än dessa två extremer. Program som tillämpar flera svaga AI-tekniker omnämns därför ibland som hybrida (Wirth, 2018). Eftersom stark AI inte existerar i dagsläget refereras artificiell intelligens till svag eller hybrid AI när det benämns i denna rapport.

Det finns en mängd olika subgrupper till AI, där de flesta inkluderar både tekniska och vetenskapliga mål. Några av dessa kan utläsas i tabell 1.

Tabell 1. Subgrupper till AI och dess tillämpningar (utvecklad från Ashok m.fl. (2016))

Subgrupp till AI	Tillämpning
Neural networks (Neuronnät)	Bokstavsigenkänning (från handskrivet), bildkomprimering, börsprediktioner
Evolutionary and genetic computing (Evolutionär och genetisk beräkning)	Fordonsdesign, datorspel, kryptering, optimerad placering av telefonmaster
Vision recognition (Synfältsigenkänning)	Bildigenkänning, navigering, medicinsk bildanalys, människa-dator-interaktion
Robotics (Robotik)	Uppladdningsbar mobilitet, tillgänglighet på områden med extrema omständigheter
Expert systems (Expertsystem)	Bakterieanalys, stavningskontroll, militär radaranalys
Speech recognition (Röstigenkänning)	Automatisk identifiering, röststyrning (t.ex. Apples Siri)
Natural Language Processing, NLP (Naturlig språkbehandling)	Förstärkta sökmotorer, transkribering, ordvalsförslag, uppläsningsfunktion
Machine Learning (Maskininlärning)	Röstigenkänning, bildigenkänning, spam-filter, genetik, ansiktsigenkänning

Det begrepp som kanske mest förekommande associeras med AI är maskininlärning, vilket är en av teknikerna med vilken det går att skapa artificiell intelligens. Maskininlärning berör hur datorer kan förbättra sig och lära sig från data (Han m.fl., 2011). Figur 2 förtydligar skillnader mellan hur AI och maskininlärning tillämpas. All maskininlärning är artificiell intelligens, men all artificiell intelligens är inte maskininlärning (MMC Venture, 2019).

	Fall 1	Fall 2	Fall 3
Maskininlärning	Känner igen ansikten	Förutspår utrustningskollaps	Förutspår sannolikheten för sjukdom
Artificiell intelligens	Förstår att du är ledsen	Schemalägger reparation	Hittar en ny behandlingsmetod
	Fall 4	**Fall 5**	**Fall 6**
Maskininlärning	Klustrar dokument efter likhet	Lär sig att känna igen synliga defekter	Förutspår ett systems prestanda
Artificiell intelligens	Sorterar dokument efter prioritering	Hittar dolda defekter	Designar ett system som möter prestandamålen

Figur 2. Exempel på hur maskininlärning och artificiell intelligens tillämpas för olika fall. (utvecklad från Overton (2018))

Algoritmer som tillämpar maskininlärning är mer eller mindre komplexa men har åtminstone en gemensam nämnare: tillgången till, och kvaliteten på, data är avgörande för att kunna göra tillförlitliga prediktioner. Detta eftersom tekniken tillåter algoritmen att lära sig genom träning istället för att följa uppsatta regler. För att programmet ska kunna lära sig hur något fungerar drar det slutsatser från så kallad *träningsdata* och anpassar algoritmen efter det. Träningsdatan behöver vara etiketterade för att algoritmen ska kunna åstadkomma detta. Ett enkelt exempel på ett två-klassificeringsproblem är en algoritm som ska förutspå om du kommer att tycka om en ny låt eller inte. Träningsdatan skulle i detta fall bestå av ett dataset med ett (relativt stort) antal låtar som du redan har etiketterat att du *tycker om* eller *inte tycker om* för att algoritmen ska kunna dra samband mellan låtarnas karaktärsdrag och sannolikheten att du skulle tycka om dem. Denna etikettering genomförs av en domänexpert då det krävs bakomliggande förståelse för datan för att kunna genomföra det korrekt.

Det finns tre typer av maskininlärning (Han m.fl., 2011):

- Oövervakad - Involverar typiska klassificeringsproblem, exempelvis då en modell ska gissa vilka låtar från en databas som en person kommer att tycka om. Inlärningsprocessen är oövervakad eftersom indatan inte är etiketterad, det vill säga eftersom användaren inte redan har gett information om att hen tycker om låten eller inte. Klustring av datapunkterna i träningsdatan används vanligen för att hitta mönster för klassificeringen.
- Semi-övervakad - Använder både etiketterade och icke-etiketterade data i träningen. Den etiketterade datan används för att definiera klustergränser och icke-etiketterade data används för att förfina modellen.

- Aktiv inlärning - Låter användaren spela en aktiv roll i inlärningen. Användaren (domänexperten) ombeds etikettera data - som kan vara både riktiga, icke-etiketterade, data eller syntetiserade (skapade). Målet med den aktiva inlärningen är att optimera modellens prestationsförmåga.

Efter att modellen har tränats valideras den vanligen på ett "osett" dataset, där dess prediktioner jämförs med faktiska data. I låtexemplet skulle vi ha sparat en andel av de redan etiketterade låtarna till valideringssteget. Sedan skulle vi ha jämfört hur många av låtarna algoritmen förutspår att du tycker om med dina faktiska etiketter av *tycker om* och *tycker inte om* för respektive låt. Detta innebär att även valideringsdatan behöver vara etiketterade i förväg. Metrikerna för valideringen varierar beroende på problemets karaktär.

I dessa typer av uppgifter är det önskvärt att datasetet är balanserat - det vill säga att det innehåller ungefär lika många observationer från varje klass. I exemplet ovan innebär det att det borde finnas ungefär lika många låtar som du tycker om som de du inte tycker om i träningsdatan. Om majoriteten av låtarna i träningsdatan är låtar som du inte tycker om kommer modellen inte att få tillräckligt med information gällande låtarna du faktiskt tycker om. Det beror på att traditionell klassificering fokuserar på att göra en så pass korrekt klassificering som möjligt och då spelar den mindre klassen också mindre roll för den totala träffsäkerheten. Ett mer livsavgörande exempel är cancerdiagnostisering. De allra flesta fall i en patientdatabas har inte diagnosen cancer, så om en modell ska göra en träffsäker och tillförlitlig klassificering kan den välja att alltid markera att det inte är cancer. Men det skulle ge en förödande effekt på de som faktiskt har cancer. Därför är det viktigt att titta på sanna positiva (true positives), falska positiva (false positives), sanna negativa (true negatives) och falska negativa (false negatives) värden illustrerade i figur 3.

	I verkligheten: **Cancer**	I verkligheten: **Inte cancer**
Markerad som: **Cancer**	Sanna positiva	Falska positiva
Markerad som: **Inte cancer**	Falska negativa	Sanna negativa

Figur 3. Sammanblandningsmatris (confusion matrix) för cancer-diagnostisering

Sanna positiva värden är fall som modellen har markerat som cancer och som är cancer i verkligheten. Falska positiva värden är fall som modellen har markerat som cancer men som inte är cancer i verkligheten. Sanna negativa värden är fall som inte har markerats som cancer och som inte är cancer i verkligheten. Falska negativa värden är fall som är cancer i verkligheten men som inte har markerats som cancer av modellen. En vanlig valideringsmetrik att använda sig av är träffsäkerhet, som helt enkelt mäter hur stor del av klassificeringen som har blivit rätt, se ekvation 1.

$$Träffsäkerhet = \frac{Sanna\ positiva + Sanna\ negativa}{Sanna\ positiva + Falska\ positiva + Sanna\ negativa + Falska\ negativa} \quad (1)$$

Eftersom denna metrik antar att datasetet är helt och hållet balanserat och att falska positiva värden och falska negativa värden är lika dåliga i realiteten bör i vissa fall, till exempel gällande cancerdiagnostisering, andra metriker användas. Några vanliga sådana är precision, sensitivitet och F1, se ekvation 2, 3 och 4. I exemplet med diagnostiseringen av cancer mäter precisionen hur många av de cancer-markerade fallen som faktiskt är cancer i verkligheten. Låg precision innebär att det förekommer många falska negativa värden och hög precision innebär att det förekommer många sanna positiva värden. Sensitiviteten mäter hur många cancerfall som modellen missar. Ett högt värde på sensitiviteten innebär att modellen inte missar så många cancerfall och ett lågt värde på sensitiviteten innebär att sannolikheten att missa cancerfallen är hög. Om utvecklingsteamet enbart skulle fokusera på att få ett högt värde på sensitiviteten skulle modellen justeras till att inkludera så många falska positiva värden som möjligt - vilket skulle minska värdet på precisionen och försämra modellens träffsäkerhet. Därför används vanligen en metrik kallad F1 - som är ett medelvärde av precision och sensitivitet och som ofta anses vara en mer relevant metrik då den balanserar dessa.

$$Precision = \frac{Sanna\ positiva}{Sanna\ positiva + Falska\ positiva} \tag{2}$$

$$Sensitivitet = \frac{Sanna\ positiva}{Sanna\ positiva + Falska\ negativa} \tag{3}$$

$$F1 = 2 * \frac{Precision * Sensitivitet}{Precision + Sensitivitet} \tag{4}$$

Det finns olika metoder för att uppnå ett balanserat dataset, varav några vanliga är översampling (oversampling), undersampling (undersampling), tröskelförflyttning (threshold moving) och ensemble-tekniker (ensemble techniques) (Han m.fl., 2011). Över- och undersampling ändrar distributionen av observationer i träningsdatan. Det finns olika varianter för att genomföra en översampling. Det går till exempel att samla in nya data eller syntetisera data. Vid syntetisering av data skapas nya datapunkter och det finns ett flertal färdiga verktyg för att göra detta inom olika maskininlärningsområden. Inom bildigenkänning är det till exempel möjligt att syntetisera data genom att duplicera och spegelvända bilderna.

En tröskelförflyttning ändrar gränsen för modellens beslut i klassificeringen så att den baseras på en kontinuerlig variabel i utfallet (Han m.fl., 2011). De värden som befinner sig över utfallsvariabeln betraktas som positiva och de som befinner sig under värdet betraktas som negativa. Denna manipulering kan även ske med hjälp av viktning av variabler. Tröskelförflyttning är enkel och fungerar bra på två-klassificeringsdata men är inte lika populär som över- och undersampling (Han m.fl., 2011). Ensemble-tekniker är ett samlingsnamn för olika kombinerade metoder och ändrar alltså modellen i sin helhet, vilket de andra tre alternativen inte gör.

Enligt MMC Venture (2019) finns det fler än 15 olika maskininlärningstekniker. En maskininlärningsteknik som har nått stora framsteg i dagens AI-utveckling är deep learning, som försöker efterlikna hur djur lär sig utföra skarpsinniga uppgifter. Tekniken

består av nätverk av artificiella neuroner, neuronnät (neural networks), som analyserar indata för att sedan plocka ut viktiga variabler och optimera dem i förhållande till problemet. Det är deep learning som tillåter självkörande bilar att känna igen föremål omkring dem genom synfältsigenkänning (computer vision) och som tillåter röstigenkänningsverktyg att förstå när vi pratar med dem genom NLP (natural language processing). Andra populära maskininlärningsalgoritmer är decision trees, bayesian networks och support vector machines (MMC Venture 2019). Det finns för- och nackdelar med samtliga metoder och dataingenjörer arbetar ständigt med att välja ut vilken metod som ger bäst resultat i förhållande till uppgiften.

2.3 AI i framtiden

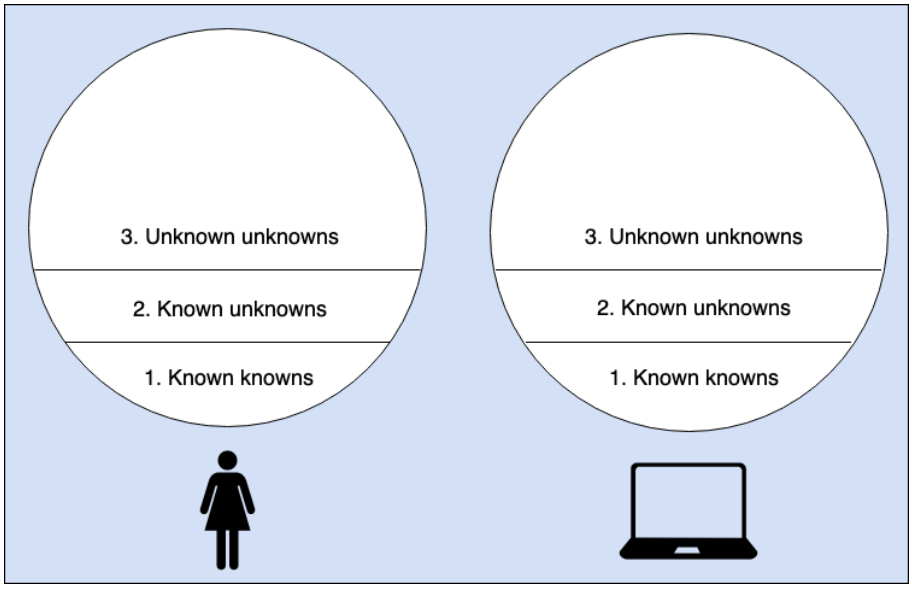
En framträdande forskare inom artificiell intelligens och dess framtid är den svenske fysikern Max Tegmark (2017) som i sin bok *Liv 3.0 att vara människa i den artificiella intelligensens tid* (2017) resonerar hur vi i konstruktion av AI bör resonera för att den i framtiden ska hjälpa och inte stjälpa den mänskliga utvecklingen. Tegmark beskriver de olika stadierna av mänskligt liv som tre nivåer: Liv 1.0, Liv 2.0 och Liv 3.0. Det tidigaste livet på jorden kallas för Liv 1.0 och inkluderar organismer och bakterier vilka inte kan lära sig något eller ta in någon kunskap. Liv 2.0 motsvarar människan som den är idag vilken kan lära sig, ta in ny kunskap och "installera ny mjukvara i hjärnan" (Tegmark, 2017). Människans intelligens har möjliggjort inlärning av allt från språk och simpel matematik till design av avancerad teknologi. Liv 3.0 existerar inte ännu men motsvarar den teknologi som i framtiden förutspås kunna designa och utveckla sig själv utan mänsklig interaktion. Detta beskrivs i tidigare avsnitt som stark AI och skulle i framtiden kunna leda till att AI utvecklar egen AI vilket Tegmark argumenterar för skulle kunna resultera i en sorts superintelligens.

Trots att vi ännu inte står inför artificiell superintelligens är det enligt Tegmark (2017) viktigt att vi från början är med och styr utvecklingen av AI. För att minimera risken att hamna i ett läge där AI används för massförstörelsevapen, mördarrobotar och social kartläggning måste vi i ett tidigt skede vara med och kontrollera dess styrka, styra dess riktning samt ge AI:n en tydlig destination. En del av denna styrning är att minimera risken för algoritmisk snedvridning då detta, vilket redovisas i senare avsnitt, har visat sig vara ett återkommande problem inom AI-utveckling.

En fråga som uppstår i debatter om superintelligens är om den verkligen kommer att kunna bli så superintelligent. Illustrationen i figur 4 beskriver den mänskliga respektive en dators kunskapsrymd och kan ses som kritik till superintelligensens möjligheter. Kunskapsrymden illustreras i form av en cirkel som fylls på ju mer människan eller datorn lär sig. När människan föds är cirkeln tom och ju äldre hon blir och ju mer hon erfar desto mer fylls cirkeln i. Kunskap kan komma från direkta lärdomar och erfarenheter från det egna livet eller från indirekta lärdomar genom studier och kunskap som läses in. Hur mycket människan än lär sig kommer hon dock aldrig uppnå

fullständig kunskap om allt och kunskapsrymden kommer således aldrig att bli helt fylld. Donald H. Rumsfeld (US Department of Defence, 2002) förklarar det som:

"There are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns -- the ones we don't know we don't know."



Figur 4. Människans respektive datorns kunskapsrymd

Vad gäller maskiner och datorer är problematiken densamma. Datorer tränar på dataset som någon eller något har gett den. Dessa dataset kan motsvara steg 1, 2 och 3 i cirkeln. Precis som människan kommer datorn aldrig kunna nå fullständig kunskap då den tränar på data som är begränsade av "known knowns" och " known unknowns". Utan kunskap om den totala populationen av uppgifter som en dator bör klara av för att anses kunna absolut allt, kommer den aldrig att uppnå detta stadium. Även om en dator har betydligt större processorkraft och minne än en människa, och på så vis kan absorbera mer kunskap finns det idag ingen möjlighet för en dator att veta när den hanterar "unknown unknowns" vilket en är begränsning till superintelligensens utveckling.

2.4 Skillnaden mellan "digitization" och "digitalization"

Digitalisering har under de senaste decennierna kommit att bli ett av de mest centrala begreppen i samhället och i stort sett alla, såväl företag som myndigheter, kommuner och landsting hävdar att de arbetar med digitalisering på ett eller annat sätt. Med ett så stort tillämpningsområde är det dock lätt att ordet används slarvigt och utan nyans. För att få en förståelse för vad som faktiskt menas med digitalisering är det därför viktigt att känna till variationerna i ordets innebörd. I det engelska språket har begreppet delats upp i två ord, digitization och digitalization vilka på svenska skulle kunna översättas till "digitisering" och digitalisering".

“Digitisering” kommer från engelskans ord för siffra (digit) och syftar till transformationen från analogt till digitalt (International Data Group, 2017). Filer så som papperskopior, fotografier och ljudfiler kan med hjälp av siffror (ettor och nollor) konverteras till digitala versioner av sig själv. De digitala versionerna kan tas bort och modifieras utan att den analoga versionen påverkas (I-scoop, 2016). Till exempel innebär en digital bild av en byggnad inte att byggnaden har digitiserats (och att bilden är en digital version av byggnaden) utan enbart att det finns en digital representation av den.

“Digitisering” är i mångt och mycket en förutsättning för det näst intill identiska ordet digitalisering. Digitalisering skiljer sig dock något från “digitisering” och syftar till en digital transformation av kommunikationsmedel, interaktionskanaler och affärsmodeller. Digitaliseringen kan leda till såväl ändrade arbetsmetoder och organisationsprocesser som nya affärsmodeller och verksamhetsstrukturer och kan tillämpas på alla olika sorters marknader och länder (I-scoop, 2016). Ett typexempel på digitalisering är transformationen inom musikbranschens som genom “digitisering” av LP-och CD-skivor övergick till digitala filer på mp3-spelare och slutligen till streamingtjänster. Omvandlingen påverkade hela musikbranschen och förde med sig helt nya affärsmodeller, arbetssätt och företag.

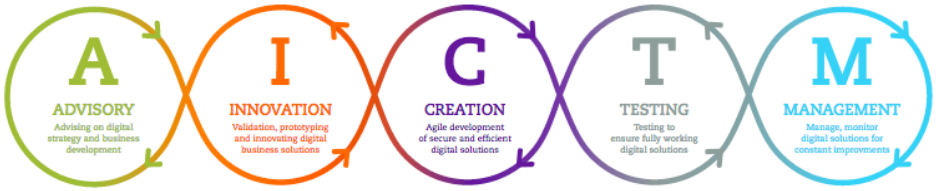
2.5 Cybercom

Cybercoms huvudsakliga marknadsområde är telekom, industri och offentlig sektor i Norden. Några av Cybercoms största kunder inom privat sektor är: Ericsson, Husqvarna, IKEA, MTV Finland och Volvo Cars (Cybercom, 2019c). Offentlig sektor står för ca 35% av Cybercoms omsättning och inkluderar kunder som Skatteverket, FMV (Försvarets materielverk), Kronofogdemyndigheten, Arbetsförmedlingen, Försäkringskassan och Sida (Styrelsen för internationellt utvecklingssamarbete). Cybercom beskriver på sin hemsida att de hjälper sina kunder att “ta tillvara på den uppkopplade världens möjligheter för att stärka konkurrenskraften eller göra effektivitetsvinster.” (Cybercom, 2019c). För att uppnå detta ägnar sig Cybercom åt strategisk rådgivning, innovation, systemutveckling, testning, kvalitetssäkring och drift ur ett livscykelperspektiv. Företagets expertis-områden är IT- och kommunikationsteknik och deras största verksamhet är utveckling, vilken representeras av Creation i figur 5. Exempel på produkter som Cybercom har varit med och utvecklat är Filmstadens nya app och BankID. Företagets erbjudande inkluderar fyra delområden: Connected Industry, Connected Consumer, Connected City och Connected Citizen (Cybercom, 2019a). På dessa delområden erbjuder företaget stöd på tre olika nivåer:

1) **Specifika verktyg:** implementering av väldefinierade lösningar för att till exempel göra produkter uppkopplade
2) **Marknadsområde:** bredare digital strategi inom ett specifikt område som kan dra fördel av digitalisering och uppkoppling

3) **Affärsmodell:** generell support för en organisation att i sin helhet dra fördel av den uppkopplade världen. (Cybercom, 2019a)

För var och en av dessa nivåer erbjuder Cybercom allt från optimering och påbyggnad av redan existerande produkter till total transformation.



Figur 5. Cybercoms delar i livscykeln. (Källa: Cybercom, 2019a)

2.5.1 Nettopositivitet och Sustainability by Design

Hållbarhet är ett viktigt begrepp för Cybercom, som kopplar alla sina projekt till FN:s 17 globala hållbarhetsmål för 2030 och arbetar utifrån ett nettopositivt förhållningssätt. Detta innebär en målsättning om att lösningarna för deras kunder ska bidra till en totalt större positiv än negativ inverkan och att på så sätt använda digitaliseringen för att driva på cirkulära istället för linjära processer. Förhållningssättet implementeras även av företag som IKEA, Ericsson, HP och Fujitsu. Detta förhållningssätt skiljer sig från traditionella sådana eftersom ytterligare ett fokuslager adderas till effekter utanför kundens egen verksamhet. Det förflyttar även fokus från att använda digitaliseringen till att göra verksamheter mindre dåliga till att hitta nya lösningar som gör dem bättre. Nettopositivitet är en naturlig strävan inom offentlig sektor men det finns även ett intresse för hållbara lösningar inom privat sektor då dessa i slutändan inte sällan även ger positiva resultat i form av ökad avkastning. Effekter som vanligen bidrar till nettopositivitet i Cybercoms olika projekt är energibesparing, dematerialisering, tjänstefiering, minskad naturresursanvändning, säkerhet eller annan produkt- eller serviceutveckling (Cybercom, 2019a).

För att uppnå nettopositivitet i sina lösningar lanserar Cybercom under våren 2019 en metod de kallar “Sustainability by Design”. Metoden fokuserar på specifika IT-/digitaliseringsprojekt som redan har beställts, planerats eller påbörjats och ska ingå i Cybercoms leveranser. Sustainability by Design ska hjälpa kunden att besvara frågan om deras planerade IT-projekt kommer att accelerera något bra eller något dåligt och hitta lösningar för att minimera riskerna för det sistnämnda. Metoden inkluderar även en första bedömning av om kunden riskerar att låsa in sig i situationer som är svåra att rätta till. Om projektet visar sig inkludera AI med risk för algoritmisk snedvridning ska en separat metod användas för att minimera risken. Det är denna metod som studiens ramverk ämnar att illustrera.

2.6 Offentlig sektor

Cybercom arbetar som tidigare nämnts med stora aktörer inom den offentliga sektorn och har många viktiga ramavtal med dessa. Ramverket ska vara tillämpbart på kunder inom både offentlig och privat sektor. Vid framtagandet av ett ramverk finns således anledning att förstå skillnaden mellan privata och offentliga aktörer. Den svenska offentliga förvaltningen omfattar allt från skola och sjukvård till stora statliga myndigheter som Skatteverket, Arbetsförmedlingen och Sida. Den offentliga sektorns arbete kan ses som statens förlängda arm och finansieras med hjälp av skattemedel. De styrs av sittande politiker på såväl kommunal som statlig nivå och dess verksamhet ska verka för allmänheten och samhällets nytta. Till skillnad från privata aktörer lyder offentlig sektor under speciella förvaltningsrättsliga lagar, som exempelvis likhetsprincipen och offentlighetsprincipen (Bengtsson, 2012). Likhetsprincipen är fastställd i regeringsformen (SFS 1974:152) och innebär att ”Domstolar samt förvaltningsmyndigheter och andra som fullgör uppgifter inom den offentliga förvaltningen skall i sin verksamhet beakta allas likhet inför lagen samt iakttaga saklighet och opartiskhet”. Utifrån ett medborgarperspektiv är det således viktigt att aktörer inom offentlig sektor i sitt arbete med såväl digitalisering som med all annan verksamhet beaktar medborgarnas bästa och agerar opartiskt och icke-diskriminerande.

I och med ständigt växande krav på effektivisering och kostnadsbesparingar inom den offentliga sektorn har digitalisering fått en central plats i arbetet. Många myndigheter, kommuner och landsting har redan kommit en god bit på vägen genom att erbjuda olika e-tjänster och digitala verktyg åt sina medborgare. För att samordna arbetet startades 2018 *Myndigheten för digital förvaltning* och året innan kom regeringen ut med en gemensam digitaliseringsstrategi för den offentliga sektorn (Myndigheten för digital förvaltning, 2018). Regeringens målsättning är att Sverige ska vara ledande i världen på att implementera digital teknik i den offentliga sektorn, där digitalisering ska användas som medel för att uppnå “effektivitet, konkurrenskraft, full sysselsättning samt ekonomiskt, socialt och miljömässigt hållbar utveckling” (Regeringskansliet, 2017).

Som ett led i Sveriges strategi för digitalisering tog Näringsdepartementet (2018) fram riktlinjer för hur offentlig sektor bör arbeta med artificiell intelligens. Enligt dessa riktlinjer ska AI ses som ett hjälpmedel som bör utnyttjas i största möjliga mån. Näringsdepartementet poängterar dock även att det krävs goda förutsättningar för utbildning, forskning, innovation och inte minst infrastruktur för att kunna ligga i framkant i utvecklingen. Näringsdepartementet konstaterar samtidigt att det finns stora risker med artificiell intelligens kopplade till snedvridna och manipulerade data samt snedvridna algoritmer. Oaktsamhet vid implementering av AI kan leda till diskriminering, minskad tillit till den offentliga sektorn och i värsta fall påverka hela demokratins funktionalitet. För att minimera dessa risker trycker därför Näringsdepartementet på vikten av att i ett tidigt skede identifiera och aktualisera de risker som finns med artificiell intelligens.

3. Teoretiskt ramverk

Detta avsnitt sammanför relevant teori angående mänskliga och algoritmiska snedvridningar. För att underlätta förståelsen för algoritmiska snedvridningar presenteras inledningsvis teorin bakom kognitiva snedvridningar hos människan. Avsnittet är baserat på arbeten från de mest omtalade forskarna inom området. Därefter ges en genomgång av begreppet algoritmisk snedvridning. Under detta avsnitt presenteras även fallstudier inom ämnet. Slutligen redogörs för existerande ramverk och riktlinjer hos EU, AI Sustainability Center, Google och Facebook.

3.1 Mänskliga snedvridningar

Snedvridning, eller bias, berör systematiska missuppfattningar och/eller okunskap grundad i otillräcklig förståelse. Det är ett utbrett fenomen som uppstår i allt från vardagliga händelser till rent statistiska beräkningar. Begreppet innebär en avvikelse orsakad av olika kognitiva eller statistiska faktorer och är ett omtalat problem i de flesta vetenskapliga studier som behandlar statistik. Avvikelsen orsakar systematiska fel som kan vara svåra att kontrollera och kan även ge upphov till att människor tar irrationella beslut (Kahneman, 2013). Det ligger således i många aktörers intresse att förstå sig på och kunna styra, men inte nödvändigtvis eliminera, både statistiska och kognitiva snedvridningar.

Snedvridning i allmänhet och kognitiv snedvridning i synnerhet är ett område som har fått ta stor plats i forskarvärlden och det finns många identifierade typer av kognitiv snedvridning. Gemensamt för alla dessa är att de förklaras av människans natur. Två av de mest framstående forskarna på området är Daniel Kahneman och Amos Tversky som under slutet av 1900-talet bedrev forskning om kognitiva snedvridningar och mänskliga beslutsmekanismer. Kahneman tilldelades 2002 Sveriges Riksbanks pris i ekonomisk vetenskap till Alfred Nobels minne (nobelpriset i ekonomi) för sin forskning inom beteendeekonomin och är också författare till en av 2000-talets mest kända böcker om mänskligt beteende, *Tänka snabbt och långsamt* (2003).

1982 lanserade Kahneman och Tversky boken *Judgment Under Uncertainty: Heuristics and Biases* vilket är en sammanställning av deras forskning om mänsklig heuristik och mänskliga snedvridningar. Boken är en redogörelse för många av de experiment som de utfört i såväl laboratorier som i större sociala, medicinska och politiska sammanhang. Ett antal av de snedvridningar som lyfts i boken och som även återfinns i *Tänka snabbt och långsamt* (2013) har en stark koppling till algoritmisk snedvridning och är således centrala för att få en djupare förståelse kring problematiken med snedvridningar - såväl mänskliga som algoritmiska.

3.1.1 Konfirmeringssnedvridning - Confirmation Bias

Konfirmeringssnedvridning är den snedvridning som uppstår på grund av människans tendens att tolka information på så vis att det bekräftar hennes redan existerande

övertygelser och ignorera eller lägga liten vikt vid motsatta fakta. Övertygelserna kan även inkludera förhoppningar och förutsägelser om framtiden. Snedvridningen tenderar att uppkomma i särskilt stor utsträckning rörande, för individen, viktiga ämnen (Britannica Academic, 2016). Människor är bättre på att ge en objektiv bedömning när de behandlar ämnen som inte ligger lika nära dem emotionellt.

En teori kring varför konfirmeringssnedvridning uppstår är att den ger människor snabb vägvisning i akuta valsituationer där mycket information behöver tas in och övervägas (Britannica Academic, 2016). Det är mycket ineffektivt, och kanske rent av omöjligt, för en människa att ta in och noggrant värdera all information hon exponeras för utan att ta hänsyn till tidigare erfarenheter av ämnet i fråga. Trots detta existerar situationer då människan borde göra just detta. Genom objektivitetsprincipen är det till och med grundlagsstadgat att domstolar och offentlig förvaltning ska iaktta opartiskhet och allas likhet inför lagen (SFS 1974:152).

En nyligen publicerad doktorsavhandling vid Uppsala universitet författad av Moa Lidén (2018) visar att konfirmeringssnedvridning existerar inom det svenska rättsväsendet. Lidén genomförde experimentella studier på poliser, åklagare och domare, där deltagarna fick läsa ett tiotal scenarier och sedan antingen själva avgöra om den misstänkte skulle gripas/häktas eller blev informerade om detta av en kollega/åklagare. De yrkesverksamma grupperna jämfördes med en kontrollgrupp jurist- och psykologstudenter som fick samma uppgift.

Efter att ha läst scenarierna fick deltagarna i polisexperimentet förbereda förhörsfrågor till de misstänkta genom att fritt skriva ner dem, alternativt välja från en lista. Studien visade att de studenter eller poliser som hade blivit informerade om att den misstänkte var anhållen i större utsträckning ställde frågor som indikerade på att denne var skyldig. I experimentet med åklagarna fick deltagarna värdera hur trovärdig den misstänktes utsago var och hur väl bevisningen indikerade på att personen var skyldig. De fick även ta beslut om åtal och beslut om någon ytterligare utredning var nödvändig.

Resultaten visade att åklagarna var mer benägna att väcka åtal om den misstänkte hade blivit gripen och hade givits en låg trovärdighetsvärdering medan trovärdigheten inte spelade någon roll alls för de som inte hade blivit gripna. Dessutom var åklagarna mindre benägna att åberopa ytterligare utredning om de själva hade väckt åtal mot den misstänkte och ännu mindre benägna om denne dessutom hade gripits. Även i domarexperimentet fick deltagarna värdera trovärdigheten och bevisstyrkan men dessutom också besluta om att fälla eller fria. Konfirmeringssnedvridningen talade även här sitt tydliga språk. Domarna var nästan 3 gånger mer benägna att fälla om de själva hade tagit beslutet om att häkta den misstänkte än om en kollega hade gjort det. Trots att det vetenskapliga sättet att testa hypoteser på är att försöka motbevisa dem tenderar både vanliga människor och forskare att välja data som konfirmerar deras hypotes (Kahneman, 2013).

3.1.2 Attributionssnedvridning - Attribution Bias

Attributionssnedvridning beskriver fel som uppstår på grund av människans tendens att försöka finna orsaker till det egna eller andras beteenden eller handlingar (Heider, 1958). Fenomenet är utbrett och därför finns många olika typer av attributionssnedvridning. En typ av attributionssnedvridning är det fundamentala attributionsfelet. Det fundamentala attributionsfelet innebär att människor tenderar att förklara utfallet av en situation eller en annan persons beteende som en konsekvens av interna faktorer (som individens personlighet) snarare än externa faktorer (som yttre omständigheter i situationen) (Kahneman, 2013). Ett exempel är då orsaken till en olycka kan framstå som att den beror på interna faktorerna hos en individ trots att orsaken i själva verket egentligen var kopplad till de yttre faktorerna i sammanhanget.

Självtjänstgörande snedvridning är en typ av attributionssnedvridning som beskriver människors tendens att anknyta deras framgångar till interna, personliga faktorer och deras motgångar till externa faktorer, exempelvis slumpen (Colman, 2015). Blaufus m.fl. (2015) ställde sig frågan om det som betraktas som moraliskt försvarbart definieras av vad som bäst gynnar individen. De genomförde en undersökning där studenter fick olika, verkliga, möjligheter att undfly skatt på intjänad inkomst från experimentet. Studenterna fick sedan ange hur moraliskt försvarbara olika situationer var, däribland legal skatteflykt. Resultaten visade att de som inte hade fått någon möjlighet att undfly skatten hade angett en betydligt högre moral gentemot skatteflykt, medan de som hade givits möjlighet att fly skatten hade uppgett en mer förlåtande inställning.

3.1.3 Överkonfidens - Overconfidence

Överkonfidens innebär en överdrivet stark tilltro till egna uppfattningar. Människor tenderar att tro att de vet mer än de i verkligheten gör. Överkonfidens bidrar till att människor inte kräver ytterligare bevis för att bli övertygade, även om det skulle vara rationellt att göra det. Det har visat sig att människans tilltro till sina egna uppfattningar inte baseras på informationens kvantitet eller kvalitet. Istället baseras den på kvaliteten av den skildringen de kan koppla till det de ser och skapa logiska mönster till (Kahneman, 2013).

Moore och Schatz (2017) delar in överkonfidens i tre kategorier för att poängtera deras olika underliggande psykologiska konstruktioner och för att minska inkonsekvent användning av begreppet: överestimering, överplacering och överprecisering. Överestimering är att överskatta den egna förmågan, prestationen eller det egna värdet. Överplacering är att överskatta sig själv i förhållande till andra. Överprecisering innebär en överdriven tilltro till vad som är sant och är kanske det som innebär störst risker inom exempelvis läkarvården. Det har nämligen visat sig att överkonfidens är en av de största kognitiva snedvridningar som förekommer i medicinska beslut (Redelmeier m.fl., 2016).

3.1.4 Inramningseffekter - Framing

Det finns en anledning till varför viktiga förfrågningar och försäljningar behöver förberedas noga för att ge positiva utslag. Inramningseffekter uppkommer av att människor påverkas känslomässigt av hur ett påstående formuleras, vilket i sin tur påverkar de beslut de tar (Kahneman, 2013). Kahneman och Tversky (1981) har studerat fenomenet närmare och genomförde ett experiment med ett exempel de kallar *Det asiatiska epidemiproblemet.* 152 studenter fick i experimentet besvara följande fråga:

Föreställ dig att USA förbereder sig för ett utbrott av en ovanlig asiatisk sjukdom som väntas döda 600 människor. Två alternativa program för att motverka sjukdomen har föreslagits. Anta att den vetenskapliga exakta uppskattningen av programmets konsekvenser är:

- Program A – 200 personer räddas
- Program B – 1/3 chans att 600 personer räddas och 2/3 risk att ingen räddas

Vilket program föredrar du?

Majoriteten av studenterna (72%) valde i detta fall Program A, vilket är den riskundvikande strategin av de två. Experimentet genomfördes även på 155 andra studenter som fick samma fråga men med följande svarsalternativ:

- Program C – 400 personer dör
- Program D – 1/3 chans att ingen dör och 2/3 risk att 600 personer dör

I detta fall valde majoriteten av studenterna (78%) Program D, som är den risksökande strategin. Program A och Program C har identiska förväntade utfall liksom Program B och Program D. Att studenterna ändå valde olika alternativ beror på inramningseffekter - människor kan tänka sig att ta en större risk när problem formuleras i förluster snarare än i vinster (Kahneman & Tversky, 1981).

3.1.5 Prevalensfel - Base Rate Fallacy

Prevalensfel uppstår när människor utgår från fel initialvärde vid bedömningar på grund av subjektivt uppskattade betingade sannolikheter. Kahneman (2013) skriver i sin bok *Tänka snabbt och långsamt* att prevalensfel härrör från en överdriven tilltro till representativitet, det vill säga stereotyper, och att relativa frekvenser tenderar att bortses ifrån. Kahneman illustrerar fenomenet med ett exempel han kallar *Tom W:s studieinriktning.* Om någon frågar dig vad Tom W studerar skulle du antagligen svara med hjälp av relativa frekvenser – det vill säga välja den utbildning som du tror är störst i det landet där Tom bor. Men om någon först skulle ge dig en beskrivning av Tom W:s egenskaper och färdigheter och sedan fråga vad han studerar skulle du antagligen använda representativiteten för att göra din gissning. I Kahnemans experiment bad han studenter rangordna en lista på 9 utbildningar. De fick även en beskrivning av Tom W

som var stereotypiskt lik studenter inom datavetenskap, maskinteknik eller biblioteksstudier. Trots att dessa är små utbildningar relativt övriga inom svarsalternativen (till exempel humaniora, samhällsvetenskap och medicin) rangordnade studenterna de små utbildningarna högst och de större utbildningarna lägst. Kahneman poängterar att representativitet är ett effektivt verktyg att använda sig av i många uppskattningar men att det finns risker med att utesluta relativa frekvenser, särskilt då bevisföringen är knapphändig. Prevalensfel kan således bidra till att potentiellt värdefull information ignoreras.

3.1.6 Haloeffekten - Halo Effect

Halo-effekten omnämns av Kahneman (2013) som *"utvidgad känslomässig tillskrivning"* och innebär en generalisering hos människan att antingen tycka positivt eller negativt kring allt eller ingenting hos en person eller ett objekt. Halo-effekten drar således slutsatser kring sådant vi inte ännu vet och förenklar därmed bilden av världen (Kahneman, 2013). Detta medför att ordningen i vilken en persons egenskaper beskrivs kan ha påverkan på hur personen uppfattas.

Halo-effekten återfinns i många vardagliga situationer. En studie genomförd av Berggren m.fl. (2016) visar till exempel att politiker gynnas av stereotypiskt vackra utseenden i "låginformationsval" - val som inte följs av storskaliga media och som inte involverar så många politiker som har synts mycket i media eller har politik som huvudsaklig sysselsättning. Det har även visat sig att halo-effekten snedvrider oss när vi bedömer vilket forskningsmaterial vi vill läsa och även bedömningen av kvaliteten på forskningen. En studie genomförd vid University of Essex och University of Cambridge (Callan m.fl., 2017) har visat att stereotypiskt vackra forskare generellt sett får mindre intresse för sin forskning och bedöms som mindre kompetenta.

3.1.7 Planeringsfelet - Planning Fallacy

Planeringsfelet har namngivits av Amos Tversky och Daniel Kahneman och är orsaken till när prognoser och planer saknar realism, endast skulle kunna slå in under otroliga, ideala omständigheter och skulle kunna förbättras med hjälp av statistik (Kahneman, 2013). Enligt Kahneman och Lovallo (1993) är överoptimism en stor orsak till varför planeringsfel uppstår. De menar att människor tenderar att överskatta risker med projekt på grund av att de av ogrundade anledningar tror att just deras projekt är unikt i relation till tidigare, misslyckade projekt. Ytterligare en konstaterad orsak till planeringsfelet är makt. I en studie genomförd av Mario Weick och Ana Guinote (2010) undersöktes riskunderskattningens korrelation med förekomsten av olika typer av makt. Resultaten visade att högre makt i alla undersökta former gav mer optimistiska och mindre korrekta tidsåtgångsprediktioner av projekt. Forskarnas teori är att människor tenderar att fokusera allt för mycket på projektets utfall snarare än dess risker när de underskattar dess tidsåtgång.

Ett klassiskt exempel på byggprojekt med grovt underskattad kostnads- och tidsåtgångsprediktion är tunnlarna genom Hallandsåsen. När klubban föll år 1991 låg den beräknade kostnaden på 1,25 miljarder kronor (SOU 1998:137) och när tunnlarna var färdiga år 2015 hade notan blivit nära tio gånger dyrare: 10,8 miljarder kronor (Trafikverket, 2015). I Tunnelkommissionens slutrapport (SOU 1998:137) genomförd 17 år innan projektet blev färdigt framkommer det att budgeten vid det tillfället hade reviderats två gånger och att den nya kostnadsuppskattningen, med 10% osäkerhetsmarginal, låg på runt 5 miljarder kronor. I rapporten står det även skrivet att tunnlarna skulle kunna vara färdiga år 2001–2002. Tunnelkommissionen drar i utredningen två slutsatser från det (redan då) katastrofala projektet:

1) Bristande styrning
2) Bristande riskanalys av miljöskador

Projektet att bygga tunnlar genom Hallandsåsen har präglats av stor miljöförstöring och en betydande del av kostnaden har varit för läkarundersökningar, reparationer och skadestånd (SOU 1998:137). Det är uppenbart att projektledningen underskattade riskerna med projektet och hade utsatts för planeringsfelet.

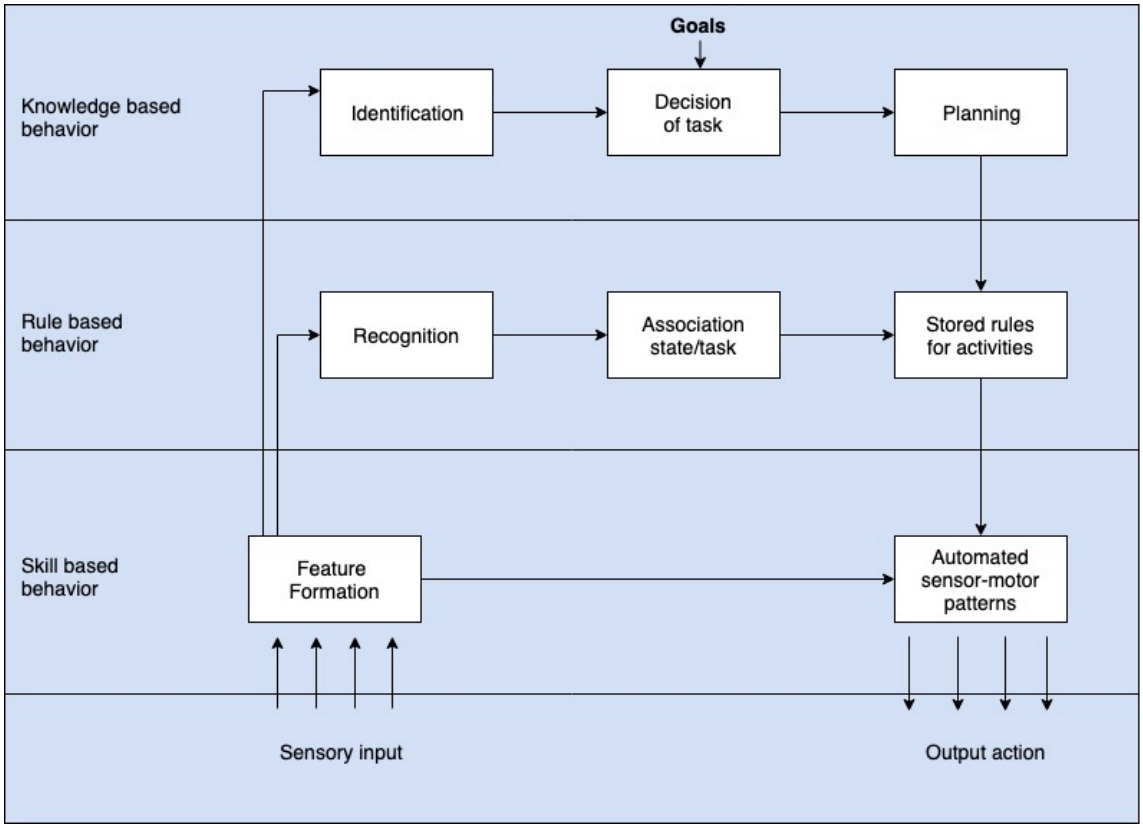
3.2 Skill-Rule-Knowledge-ramverket

Samtidigt som Kahneman och Tversky ämnade förklara människans beteende genom olika snedvridningar figurerade en annan känd forskare, Jens Rasmussen, inom samma område. Rasmussen var en dansk forskare specialiserad på systemsäkerhet och den mänskliga faktorn. Rasmussen tittade bland annat på hur det går att undvika misstag och fel gjorda av människan genom att applicera olika typer av symboler, tecken och signaler i ett systems design. 1983 publicerade han boken *Skills, rules, knowledge: signals, signs and symbols and other distinctions in human performance models* (1983). För att motverka den mänskliga faktorn på ett system krävs en förståelse för varför och när människor tenderar att göra olika typer av misstag. Rasmussen identifierade att människans beslutsmekanismer kan delas upp i tre olika beteendenivåer; skicklighet, regel och kunskapsbeteende. Dessa nivåer lade grunden till Skill-Rule-Knowledge-(SRK)-ramverket vilket har kommit att bli ett väletablerat ramverk vid design av säkra system. Figur 6 är en förenklad illustration över de tre beteendenivåerna och människans korrelerade utförandemönster kopplade till dessa nivåer.

3.2.1 Skicklighetsbaserat beteende - Skill Based Behavior

Skill based behavior, eller skicklighetsbaserat beteende, illustrerat längst ned i figur 6, är beteenden som är starkt kopplade till rutin och automatiskt agerande (Rasmussen, 1983). Denna nivå kräver lite eller inga aktiva beslut och sker mer eller mindre enligt förprogrammerade beteendemönster. Utförandet sker med lätthet och typen av fel och misstag som kan uppstå är så kallade *slips* eller *lapses*. Beteendet är sensomotoriskt, vilket innebär att det sker i samverkan mellan sinnesintryck och muskelrörelse. Sinnesintryck går sällan att varken välja bort eller identifiera med blotta ögat och de

korrelerade muskelrörelserna sker automatiskt. Ett exempel på skicklighetsbaserat beteende är framförandet av en cykel eller att, för en pilot, köra ett flygplan under normala omständigheter. Manuellt framförande av flygplan är kopplat med mycket träning och flygvana och de flesta beslut som sker vid en vanlig flygning är helt reaktiva och baseras på den vana som piloten besitter efter många års träning (Mayerhofer, 2018).



Figur 6. Förenklad illustration över tre nivåer av utförande av mänskliga operatörer (utvecklad från Rasmussen (1983))

3.2.2 Regelbaserat beteende - Rule Based Behavior

Ett regelbaserat beteende eller rule based behavior präglas av användningen av regler och kända tillvägagångssätt för att identifiera en handlingsplan i en välbekant arbetssituation (Rasmussen, 1983). Detta beteende är illustrerat i mittsektionen i figur 6. Till skillnad från skicklighetsbaserat beteende kräver det regelbaserade beteendet en högre nivå av kognitivt medvetande. Reglerna som sätts upp av individen är en uppsättning instruktioner som baseras på tidigare erfarenheter och incidenter. Då det uppstår situationer som följer vanliga rutiner men som inte utförs per automatik skickas det vidare till den regelbaserade nivån. Där identifieras situationen och vilka sammanställda rutiner och lagrade regler som bör appliceras för att lösa den. I flyg-exemplet motsvarar det regelbaserade beteendet de ageranden som baseras på regler och

tillvägagångssätt som piloten själv har skapat under sina år som pilot (Mayerhofer, 2018). Det är således inte de regler och checklistor som är uppsatta för alla inom flyget utan snarare regler baserade på egna erfarenheter.

3.2.3 Kunskapsbaserat beteende - Knowledge Based Behavior

Den sista nivån i Rasmussens (1983) modell är Knowledge based behavior eller kunskapsbaserat beteende, illustrerat överst i figur 6. Det kunskapsbaserade beteendet innebär en större kontroll och ett mer avancerat resonemang än både det skicklighetsbaserade och det regelbaserade beteendet. Denna typ av kontrollbeteende krävs när situationen är oväntad. Vid en ny eller oväntad situation passeras signalen vidare från både den skicklighetsbaserade nivån och den regelbaserade nivån. Baserat på en analys av det rådande läget sätts explicita mål upp för situationen, vilket påverkar valet av utförande. Inom flyget skulle denna typ av situation kunna exemplifieras med en exceptionell flygplanshändelse som nödlandningen på Hudsonfloden i New York (Mayerhofer, 2018). Vid denna händelse utsattes piloten för en extrem situation utan tydliga rutiner eller erfarenheter att följa. Genom att analysera situationen och avväga potentiella alternativ kunde piloten dock hantera situationen och nödlanda flygplanet säkert mitt i Hudsonfloden.

3.3 Det mänskliga intellektets två system

Psykologerna Keith Stanovich och Richard West lanserade tillsammans en teori om att människans intellekt är uppdelat i två system, det snabba System 1 och det långsamma System 2 (Stanovich & West, 2000). Daniel Kahneman kom senare att tillämpa System 1 och 2 på sin forskning om mänskliga snedvridningar och dessa blev grundpelare i hans bok *Tänka snabbt och långsamt* (2013). Ett bakomliggande antagande om System 1 och 2 är att kognitiva snedvridningar ligger i människans natur. Inte helt okontroversiellt kan man säga att det skicklighetsbaserade och det regelbaserade beteendet identifierat av Rasmussen faller in under Stanovich och Wests definition av System 1 och det kunskapsbaserade beteendet faller in under definitionen av System 2.

3.3.1 System 1

System 1 kan kort beskrivas som den intuitiva delen av människans tankesystem som arbetar automatiskt och snabbt med liten eller ingen ansträngning. Det intuitiva tänkandet är en process som sker utan någon känsla av varken delaktighet eller medveten styrning, den sker helt automatiskt och kräver stor ansträngning för att styra eller ändra (Kahneman, 2013). Ett exempel på en situation då System 1 aktiveras är då en person blir visad en entydig bild. Det kan vara en bild föreställande ett argt ansikte eller ett glatt barn. Intuitivt och utan någon ansträngning kan personen som exponeras för bilden identifiera de olika känslorna som de två ansikten uttrycker. Detta är ett mycket simpelt exempel och en situation då det är nästintill riskfritt att förlita sig på System 1. Andra exempel på situationer som hanteras av System 1 är; enkla matematiska beräkningar, att avsluta meningen “krig och ...”, uppfatta om ett föremål

befinner sig nära eller långt bort, tyda budskapet på en stor reklamskylt och så vidare (Kahneman, 2013).

Många av de mentala aktiviteter som finns inbäddade i System 1 är medfödda och finns hos alla individer medan andra utvecklas automatiskt efter långvarig och ihärdig träning. Ett exempel på detta är ett schackdrag som ur en oerfaren spelares perspektiv kan tyckas vara en aktivitet som kräver mycket tankekraft och analys men som för en erfaren schackspelare kan vara lika intuitivt som att ange svaret på 2+2. Vad som innefattas i en persons System 1 är alltså helt individuellt och beror på personens erfarenheter, uppväxt, intressen och så vidare. Människans System 1 har en tendens att vara diskriminerande, dömande, irrationellt och impulsivt vilket visar sig i olika typer av snedvridningar presenterade ovan.

3.3.2 System 2

System 2 kan beskrivas som långsamt, beräknande, logiskt och organiserat. Till skillnad från System 1 upplevs användandet av System 2 som medvetet och ansträngande och medför aktivt deltagande och en subjektiv känsla. Att ge ett korrekt svar på problemet 13x47 är svårt (för de flesta) och är därför ett exempel på en uppgift som kräver deltagande av System 2 (Kahneman, 2013). För att kunna lösa denna typ av matematiska problem krävs att personen i fråga fokuserar, eventuellt antecknar och använder sig av sina tidigare multiplikationskunskaper för att på så vis finna en lösning på problemet. System 2 använder sig av lagrade minnen och tidigare upplevelser för att ta sig an ett nytt problem från sina erfarenheter.

Ett annat exempel på en situation som hanteras av System 2 är då en bil ska parkeras i en trång parkeringsficka, vilket kräver både mental styrka och fysisk rörelse. Listan kan göras lång på aktiviteter och beslut som kräver System 2 men precis som System 1 beror dessa situationer på individens minne, erfarenheter och tidigare kunskaper. Aktiviteter som kräver System 2 ockuperar delar av den mentala kapaciteten hos en person och i vissa fall är det omöjligt att utföra mer än en aktivitet åt gången. Det är exempelvis svårt, eller näst inpå omöjligt, att lösa 13x47 samtidigt som en bil manövreras in i en trång parkeringsficka (Kahneman, 2013). I en av de studier som Kahneman beskriver i *Tänka snabbt och långsamt* (2013) redogör de även för att de mentala utmaningarna som hanteras av System 2 även aktiverar vissa fysiska funktioner i kroppen. Att koncentrera sig och lösa en svår uppgift kan bland annat leda till svettningar, förändrad pupillstorlek och ökad hjärtrytm (Kahneman, 2013). I samma stund som ett problem blir löst eller personen ifråga bestämmer sig för att ge upp går dessa funktioner tillbaka till det normala. Detta är ett tecken på att System 2 har kopplats bort och inte längre behövs.

3.4 Algoritmiska snedvridningar – Algorithmic Bias

Ur teknisk synvinkel är den kanske mest kända snedvridningen metodfel, vilket är något som alla som hanterar data behöver ta hänsyn till och vidta aktiva åtgärder för att

motverka. Några förekommande exempel på sådana är sammanblandning av orsaksfaktorer (confounding), där korrelation mellan variabler uppstår utan att det finns ett direkt samband mellan dem, urvalssnedvridning (selection bias), där urvalet av data inte speglar hela populationen och informationssnedvridning (information bias), som inträffar under insamlingen av data (Delgado-Rodríguez & Llorca, 2004). Alla dessa metodfel är starkt kopplade till kvaliteten på använda data.

Joseph Weizenbaum hävdade genom sin bok *Computer power and human reason: from judgment to calculation* (1976) att algoritmer kan drabbas av snedvridningar både genom vilka data som används i programmet och på vilket sätt programmet är kodat. Weizenbaum reflekterade över detta faktum då han utvecklade ett textanalysprogram, ELIZA, som kunde simulera svar i en konversation. ELIZA's första uppgift var att agera Rogeriansk psykoterapeut, vilken återspeglar patientens uttalanden i sina frågor och påståenden. Han upptäckte snart att människor tenderar att prata med datorn som om den vore en människa och att allmänheten började spekulera kring möjligheter med att använda programmet inom psykoterapin (Weizenbaum, 1976). Dessa uttalanden skrämde Weizenbaum, som var av uppfattningen att det finns begränsningar för vad en dator borde göra. Han stärkte sin uppfattning genom att påpeka att eftersom algoritmer är en uppsättning regler som människan har satt upp, baseras de oundvikligen på programmerarens antaganden, förväntningar och snedvridningar (Weizenbaum, 1976). Likväl, hävdade Weizenbaum, kan dessa snedvridningar reflekteras i valda indata - vilket snedvrider algoritmen i sig eftersom det utgör datan som algoritmen tränas på.

Fenomenet som Weizenbaum (1976) beskriver kallas idag *algoritmisk snedvridning* och innebär att människors kognitiva snedvridningar i form av värderingar och fördomar implementeras i teknik. Begreppet är något missvisande då algoritmen i sig inte nödvändigtvis är snedvriden ur ett statistiskt perspektiv, men blir snedvriden ur ett socialt perspektiv i implementationen, då den reflekterar existerande social snedvridning hos individen eller samhället. Sammanblandning av orsaksfaktorer, urvalssnedvridning och informationssnedvridning som nämns ovan finns inte bara i teknik, utan även hos människor.

Batya Friedman och Helen Nissenbaum (1996) definierar algoritmisk snedvridning som datorsystem som systematiskt och orättvist diskriminerar vissa individer eller grupper till fördel för andra. De identifierar i sitt ramverk tre olika kategorier av algoritmisk snedvridning: teknisk, förexisterande och framkommande snedvridning. Teknisk snedvridning uppkommer enligt deras kategorisering till följd av tekniska överväganden eller begränsningar. Det kan till exempel röra sig om system sorterade i alfabetisk ordning som gynnar de som har namn som börjar på A. Förexisterande snedvridning är sådan som existerade redan innan algoritmen implementerats, inom antingen samhället i stort eller hos den enskilda individen. Det är denna typ av snedvridning som reflekteras i koden till följd av utvecklarnas egna värderingar och implementeringen kan vara antingen explicit eller implicit. Till skillnad från förexisterande snedvridning uppstår framkommande snedvridning när algoritmen redan är implementerad. Detta eftersom

snedvridningen ofta uppstår till följd av en förändring, exempelvis inom allmän kunskap, population eller kulturella värderingar. Det kan till exempel inträffa vid en utökning av ett systems målgrupp från nationell till internationell nivå. Friedman och Nissenbaum (1996) poängterar att användargränssnitt är särskilt känsliga för den här typen av snedvridning eftersom dess design är så beroende av användarens kunskap, egenskaper och vanor.

Vad som anses vara diskriminerande kan vara individuellt men i Sverige anger diskrimineringslagen (SFS 2008:567) sju diskrimineringsgrunder, vilka kan medföra rättsliga påföljder om de inte tas hänsyn till:

- Kön
- Könsöverskridande identitet eller uttryck
- Etnisk tillhörighet
- Religion eller annan trosuppfattning
- Funktionsnedsättning
- Sexuell läggning
- Ålder

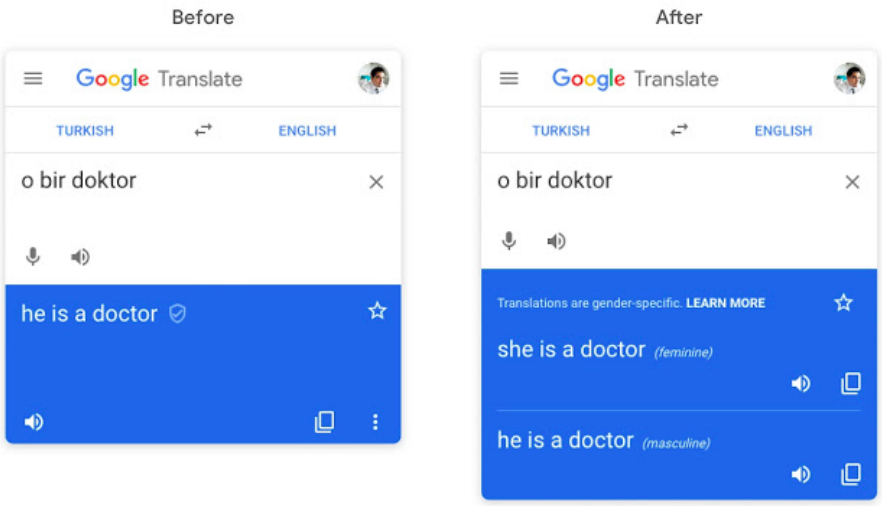
Idag har ett flertal fall av algoritmisk snedvridning som har berört dessa diskrimineringsgrunder uppmärksammats världen över, varav ett urval redogörs för nedan.

3.4.1 Textanalys

Sedan lanseringen av översättningsverktyget Google translate har Google gjort stora framsteg i utvecklingen av verktygets kvalitet och korrekthet. Tekniken bakom verktyget bygger på artificiella neuronnät (neural networks) och kan översätta ord och meningar på fler än 100 språk (Turovsky, 2016). Verktyget beräknas ha fler än 500 miljoner användare per dag världen över och är därmed det överlägset största programmet i sitt slag (Turovsky, 2016). Med ett sådant inflytande som verktyget har är det kritiskt för Google att tillhandahålla ett verktyg som minimerar snedvridna översättningar, något som har varit ett återkommande problem för systemet. Många språk skiljer sig specifikt i hur de representerar kön i pronomen, vilket kan leda till tvetydiga översättningar. På grund av dessa tvetydigheter har Google translate tenderat att generera översättningar som återspeglar snedvridningar representerade i samhället (Johnson, 2018).

Denna typ av tvetydig översättning leder bland annat till att verktyget översätter historiskt manliga yrken och egenskaper till manliga pronomen och historiskt kvinnliga yrken och egenskaper till kvinnliga pronomen. Ett exempel som lyfts av både Google själva såväl som kritiker till Google Translate är de turkiska uttrycken *o bir doktor* och *o bir hemşire* vilka är könsneutrala och som på turkiska betyder *hen är en doktor* och *hen är en sjuksköterska*. Översatt till engelska ändrar Google Translate innebörden och översätter uttrycken till *he is a doctor* och *she is a nurse*. Enligt Bureau of Labor

Statistics (2018) var 88% av sjuksköterskorna i USA kvinnor 2018 och 57% av läkarna var män. Sexton år tidigare var 92% av sjuksköterskorna kvinnor och 69% män (Bureau of Labor Statistics, 2002). Googles algoritm har med andra ord kopierat en underliggande snedvridning i samhällets yrkeskårer. Enligt Google är detta, tillsammans med många liknande exempel, ett problem som har åtgärdats genom modifiering av de existerande modellerna (Johnson, 2018) vilket även illustreras i figur 7. Vid en snabb kontroll av det existerande systemet kan det dock konstateras att problemet kvarstår på Googles egen sökmotor, se figur 8. Google har med andra ord enbart modifierat applikationen, inte funktionen i sökmotorn.

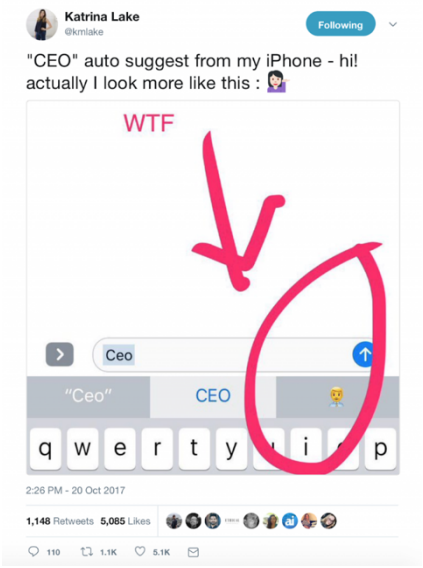


Figur 7. Google Translates modifiering av översättning för könsneutrala pronomen (vänster bild är innan modifieringen och höger bild är efter) (Källa: Johnson, 2018)

Figur 8. Googles översättningsfunktion i sökmotorn för de könsneutrala uttrycken "hen är en doktor" och "hen är en sjuksköterska"

Även Apple har fått kritik för ett liknande problem i deras autofill-funktion på tangentbordet i iPhone. Med hjälp av autofill ger tangentbordet förslag på ord som passar att efterfölja ord som just har skrivits. Kritiken riktade sig främst mot att algoritmens förslag var snedvridna ur ett genusperspektiv (Olson, 2018). Om meningen

som skrevs var *Doktorn sa att* gavs alternativen [han, den, det] och om meningen var *Sjuksköterskan sa att:* gavs alternativen [hon, den, det]. Detta har även speglats i vilka emojis algoritmen har gett som förslag som ersättning för olika ord. Något som uppmärksammats av StitchFix VD Katrina Lake var när hon på sin iPhone skrev ordet CEO (VD) och endast erbjöds förslag på en emoji föreställande en manlig person och inte en kvinnlig, se figur 9 (Olson, 2018). Detta stämmer väl överens med en rådande genus-snedvridning i samhället. På listan över USA:s 500 största företag 2017 var bara 24 VD:ar kvinnor (Fortune, 2017). Katarina Lakes bild fick stor spridning och har bidragit till diskussionen om genus-snedvridna översättningsalgoritmer.



Figur 9. Autofill-funktionen på iPhone föreslår en manlig emoji då ordet CEO skrivs ut (Källa: Olson, 2018)

De ovan redovisade fallen berör maskininlärningsområdet NLP (Natural Language Processing) och är problematiska på grund av att de är kontextbaserade. De genus-snedvridna förslagen i Google Translate och autofill-funktionen på iPhone är konsekvenser av algoritmer som har tränats på data där yrken och egenskaper har varit snedvridet anknutna till ett visst kön. Efter att en modell har tränats på ett ursprungligt dataset omtränas modellen ofta kontinuerligt med hjälp av de alternativ som användarna väljer. Om de ursprungliga alternativen är snedvridna och det inte finns någon möjlighet att lära om modellen genom att exempelvis välja en kvinnlig VD-emoji kommer modellens förslag snedvridas ytterligare då modellen vid omträningen får bekräftelse på att dess ursprungliga och snedvridna förslag var korrekt.

3.4.2 Anställningsförfarande

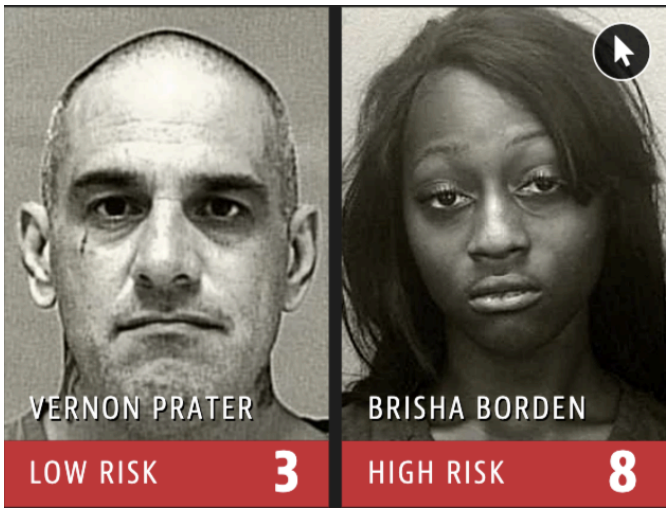
Eftersom det finns en så tydlig underliggande snedvridning i samhället gällande yrkesval har algoritmisk snedvridning även påvisats i rekryteringsprocesser. IT-jätten Amazon vill ligga i framkant i automatiseringen och utvecklade 2014 en algoritm för att gradera jobbansökandens CV:n utefter hur lämpliga aspiranterna var för posten.

Algoritmen tränades genom att utläsa mönster från inkomna CV:n inom en tioårsperiod och visade sig föredra män för tekniska roller på grund av representativiteten av män i urvalet (Reuters, 2018). Enligt Reuters (2018) källor gjorde Amazon algoritmen neutral gentemot ordet *kvinna,* men beslutade sig för att stänga ner projektet, vilket indikerar på att problemet kvarstod.

Även marknadsföring till lediga tjänster har visat sig vara snedvriden. Marknadsföring når per definition individen bäst då den är anpassad för dennes behov och önskemål. Datta m.fl. (2015) genomförde därför en studie angående diskriminering i riktad reklam. Artikeln visade att riktad marknadsföring av lediga arbetspositioner representerades olika mellan kvinnor och män. Män fick i större utsträckning marknadsföring för tjänster gällande chefspositioner och andra välbetalda arbeten än kvinnor fick. Författarna menar att marknadsföringen i sig förvisso inte nödvändigtvis behöver anses vara diskriminerande men att det i det långa loppet kan leda till ojämställdhet mellan könen och könsdiskriminering. Om kvinnor, delvis på grund av snedvriden marknadsföring och sämre kännedom om lediga tjänster, i mindre utsträckning söker sig till arbeten på högre positioner kan det leda till att färre kvinnor tillsätts till den typen av tjänster vilket i sin tur bidrar till ojämlikheten på arbetsmarknaden.

3.4.3 Riskbedömning

I USA är det vanligt att låta algoritmer göra en riskbedömning i samband med rättsfall. Bedömningen överlämnas till domaren innan domstolsbeslut och utgör en grund för viktiga beslut rörande allt från borgenssummor till frigivning (ProPublica, 2016a). Den oberoende nyhetsbyrån ProPublica (2016a) beskriver i sin rapport ett stort antal tillfällen då riskbedömningspoängen har gett effekt på de misstänktas domar. Ett exempel de nämner är då en man vid namn Paul Zilly stod åtalad för att ha stulit en gräsklippare och några verktyg. Åklagaren rekommenderade ett års fängelse och efterföljande uppföljning för att se till att Zilly fortsatte på "rätt bana". Efter att ha sett Zillys höga riskbedömningspoäng valde dock domaren att upphäva överenskommelsen mellan åklagare och försvaret och dömde istället Zilly till två års fängelse. ProPublica nämner dock även många, för de misstänka, fördelaktiga beslut tagna till följd av algoritmen - så som alternativa domar istället för fängelse. Trots den stora påverkan riskbedömningsalgoritmerna har på människors liv redovisas beräkningarna bakom dem sällan. ProPublica (2016a) lät därför göra en undersökning av ett av de mest förekommande riskbedömningssystemen vid namn COMPAS (Correctional Offender Management Profiling), utvecklat av Northpointe, och fann oroväckande resultat. Northpointe har ifrågasatt rapportens metoder och slutsatser, vilket ProPublica har besvarat (ProPublica, 2016b).



Figur 10. Northpointes riskbedömning för två arresterade amerikanska medborgare som ertappades med stöld. Vernon Prater (till vänster) var tidigare dömd för väpnat rån. Brisha Borden (till höger) var tidigare dömd för mindre företeelser under ungdomsåren (Källa: ProPublica, 2016a)

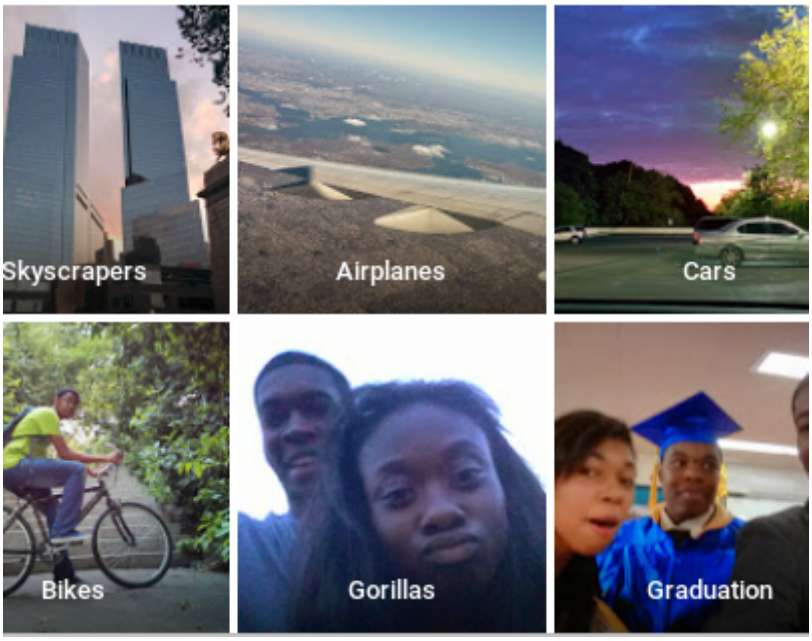
Northpointe's verktyg bedömer risken för att personen i fråga kommer att begå framtida brott på en skala från 1–10 och klassificerar den misstänkta i låg-, mellan- och högrisk-kategorier, se figur 10. Verktyget gör sin bedömning utefter svaren på 137 frågor (Northpointe, 2011) som antingen den misstänkte själv fyller i eller som tas ut från brottsregistret (ProPublica, 2016a). Enligt ProPublicas undersökning (ProPublica, 2016a) var algoritmen 77% mer benägen att högrisk-kategorisera mörkhyade för framtida våldsbrott och 45% mer benägen att högrisk-kategorisera dem för framtida generella brott jämfört med ljushyade personer, oberoende av bakgrundsfaktorer som ålder, kön och tidigare brottshistorik.

Det är enligt den amerikanska konstitutionen inte tillåtet att döma någon utefter deras etniska tillhörighet (U.S. Const. Amend. XIV, Sec. 1). Ändå går dessa siffror i linje med forskningsresultat som konstaterar att mörkhyade personer mer sannolikt blir arresterade i USA än ljushyade personer, oberoende av liknande bakgrundsfaktorer (Kizer, 2017). Det är med andra ord sannolikt att denna underliggande snedvridning återspeglas i träningsdatan. Trots detta inkluderar ingen av frågorna som verktyget gör sin bedömning på etnisk tillhörighet (Northpointe, 2011). Algoritmen kan dock ändå hitta mönster kopplade till etnisk tillhörighet genom att frågor som korrelerar med denna inkluderas i verktyget (ProPublica, 2016a). Ett vanligt argument till att inkludera sådana riskfyllda parametrar - som kan bidra till algoritmisk snedvridning - är att träffsäkerheten minskar om dessa exkluderas. I just detta fall har argumentet motbevisats av Julia Dressel och Hany Farid (2018) som i sin undersökning fann att samma träffsäkerhet som COMPAS uppnådde kan åstadkommas vid inkludering av enbart två parametrar (ålder och antal tidigare fällande domar) istället för 137 parametrar. I Dressel och Farids modell uppnådde COMPAS träffsäkerhet 65,2%. New York State gjorde 2012 en undersökning av COMPAS-algoritmen, som fann att den hade 71% träffsäkerhet (New York State Division of Criminal Justice, 2012). New York

State undersökte inte skillnader i förutsägelser av framtida brott i relation till etnisk tillhörighet, vilket kan vara en anledning till att de gav grönt ljus för verktyget.

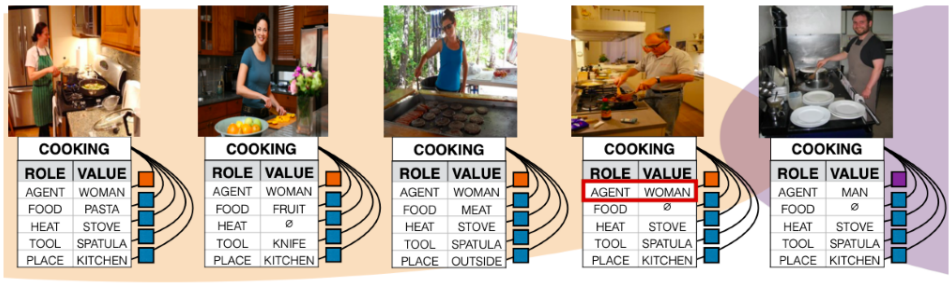
3.4.4 Bildigenkänning

Flera fall av algoritmisk snedvridning har uppkommit i samband med bildigenkänning - eller AI-området *computer vision*. Jacky Alciné (2015) gjorde en obekväm upptäckt när han använde Googles bildigenkänningverktyg och märkte att hans mörkhyade vänner klassificerades som "gorillor", se figur 11. Enligt en undersökning gjord av Wired Magazine (2018) har problemet ännu inte fått en långsiktig lösning då algoritmen fortfarande inte kan känna igen vissa typer av apor, däribland gorillor. Google har helt enkelt tagit bort etiketten "gorilla". Ett annat fall av algoritmisk snedvridning i bildigenkänning uppdagades då en kvinna med asiatiskt påbrå publicerade en bild på sin hemsida (Wang, 2009) föreställande bilden av skärmen på sin Nikon-kamera när hon tog en selfie. På skärmen stod det "Did someone blink?". Trots att Nikon är ett japanskt företag var algoritmen uppenbarligen inte tillräckligt tränad på ansikten med asiatiskt påbrå. Även Flickr har haft problem med sin bildigenkänningsalgoritm. Enligt The Guardian (2015) har Flickrs autotaggning gjort ett flertal kontroversiella misstag i sin etikettering. Människor, både mörk- och ljushyade, har fått etiketten "apa" och "djur" och koncentrationsläger har fått etiketten "sport". En talesperson från Flickr säger till tidningen att de ständigt arbetar med att förbättra algoritmen, som lär sig av sina misstag när någon raderar en tagg. Enligt The Guardian (2015) har Flickr tagit bort etiketten "sport" från bilder på koncentrationsläger och helt och hållit raderat etiketten "apa".



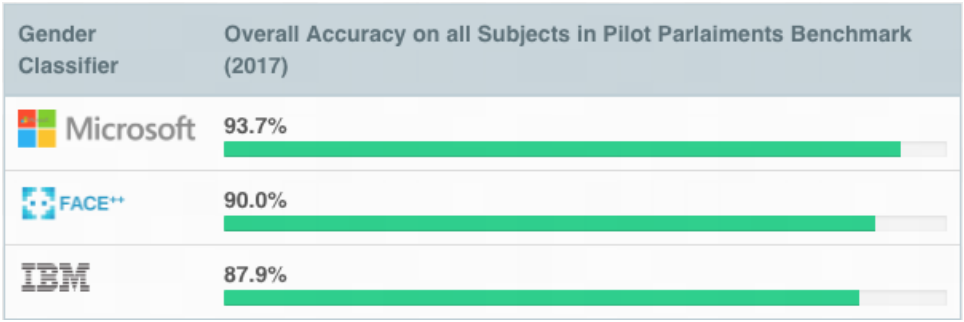
Figur 11. Googles bildigenkänningsverktyg etiketterar två mörkhyade personer som gorillor (Källa: Alciné, 2015)

Även snedvridning med avseende på kön har framkommit inom bildigenkänning. Chang m.fl. (2017) har publicerat en artikel i samband med en undersökning av genussnedvridning hos två stora dataset - Imsitu och COCO (som sponsras av bland andra Microsoft och Facebook) - som stödjer bland annat situations- och aktivitetsigenkänning. Resultatet visade att båda dessa dataset var genussnedvridna. Författarna fann bland annat att aktiviteten *matlagning* var över 33% mer benägen att utföras av en kvinna än en man på bilderna i datasetet. Efter att ha tränat algoritmen sjönk andelen män som lagar mat på bilderna från 33% till 16%. Författarna fann med andra ord att aktivitetsigenkänningen tenderade att till och med förstärka snedvridningen från träningsdatan i utfallet. Ett exempel på detta illustreras i figur 12, där en man som lagar mat klassificeras som en kvinna.



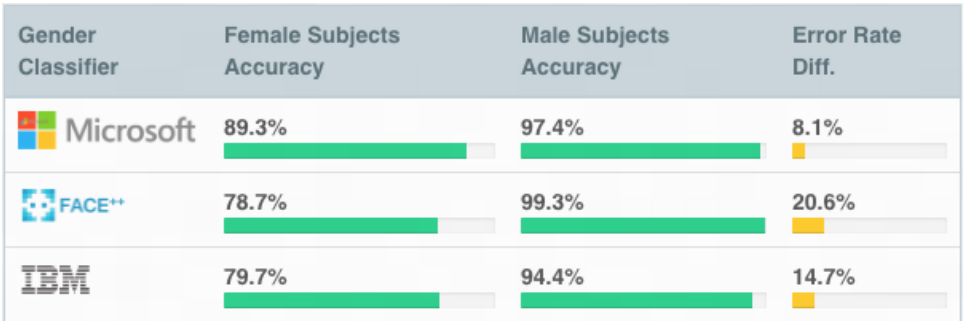
Figur 12. Aktivitetsigenkänningsalgoritmen Imsitu förstärker könssnedvridning i träningsdata genom överklassificering av kvinnor i kök (Källa: Chang m.fl. 2017)

För att utvärdera olika bildigenkänningsalgoritmer och deras träffsäkerhet gentemot olika kön och hudfärger genomförde Joy Buolamwini och Timnit Gebru en studie där tre olika "gender classifiers", könsigenkänningsverktyg, testades mot varandra (Buolamwini m.fl., 2018). De verktyg som testades var Microsoft, IBM och Face++ vars huvudsakliga funktion är att med maskininlärning utifrån en bild identifiera könet på en person. För att kunna testa verktygen på ett korrekt sätt använde de i studien ett jämt fördelat dataset, vilket representerade alla hudtoner och båda könen i samma utsträckning. I det första testet presenterat i figur 13 jämfördes den övergripande träffsäkerheten för samtliga verktyg. Resultatet visade att Microsoft presterade högst på 93.7% och IBM presterade sämst på 87.9%.



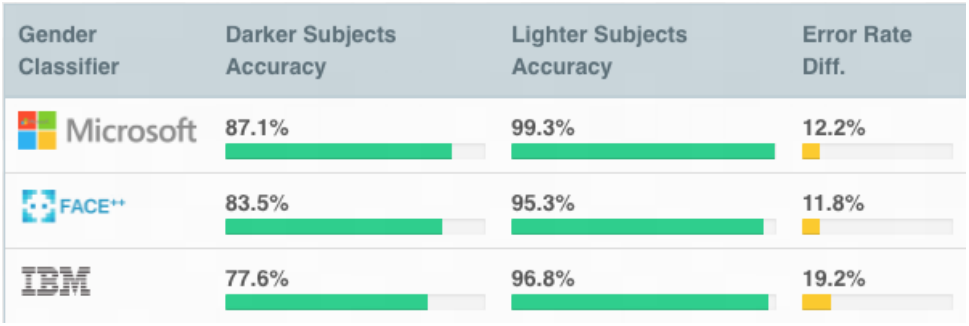
Figur 13. Träffsäkerhet för respektive verktyg utifrån hela testsetet (Källa: Gender Shades, 2018)

Studien visade att alla verktyg presterade bättre på män än på kvinnor, vilket presenteras i figur 14. Skillnaden i felfrekvens för de olika verktygen var 8.1% − 20.6%, där Microsoft presterade bäst med en felfrekvens på 8.1% och där Face ++ presterade sämst med en felfrekvens på 20.6%



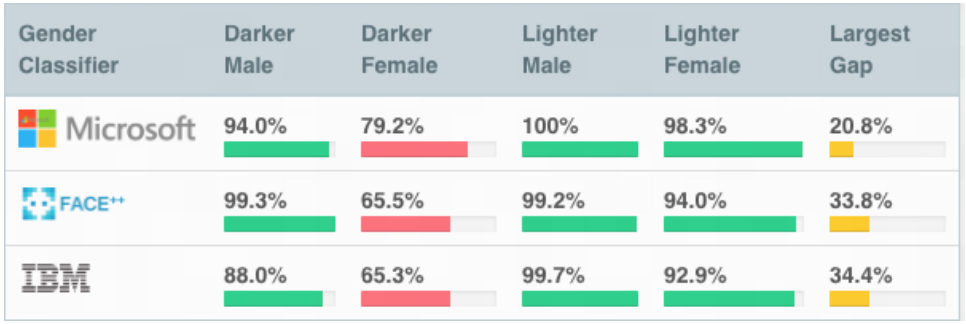
Figur 14. Jämförelse av verktygens träffsäkerhet vid bedömning av kön (Källa: Gender Shades, 2018)

Studien visade även att samtliga verktyg presterade bättre på ljusare än mörkare ansikten, se figur 15. Face++ presterade bäst med en felfrekvens på 11.8% och IBM presterade sämst med en felfrekvens på 19.2%.



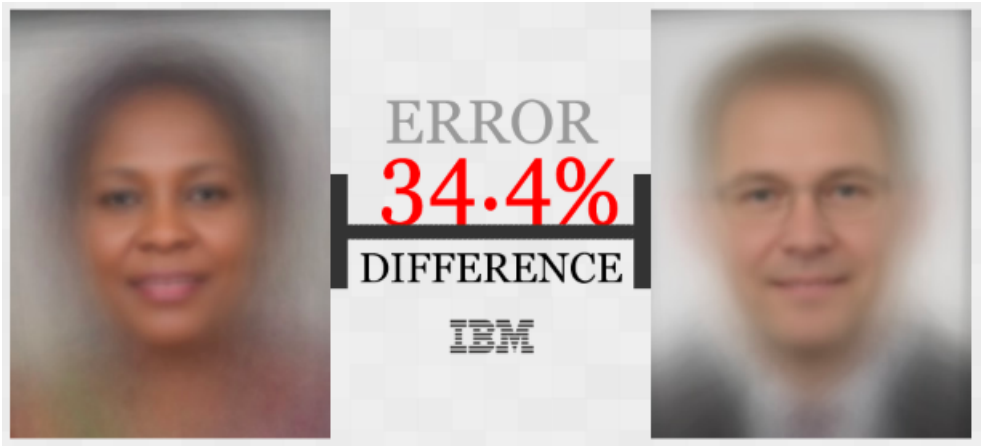
Figur 15. Jämförelse av verktygens träffsäkerhet vid bedömning av hudton (Källa: Gender Shades, 2018)

Samtliga verktyg presterade sämst på mörka kvinnliga ansikten och bäst på ljusa manliga ansikten där Microsoft presterade bäst och IBM presterade sämst, se figur 16.



Figur 16. Jämförelse av verktygens träffsäkerhet vid bedömning av kön och hudton (Källa: Gender Shades, 2018)

Den största skillnaden i felfrekvens mellan den grupp där verktygen presterade bäst (ljusa manliga ansikten) och den grupp där verktygen presterade sämst (mörka kvinnliga ansikten) var 34.4% se figur 17. Denna typ av jämförelse visar på att verktygs och algoritmers träffsäkerhet inte enbart bör valideras utifrån ett helhetsperspektiv utan att den även bör testas för samtliga subgrupper i ett dataset.



Figur 17. Skillnaden i träffsäkerhet i IBM:s bildigenkänninsalgoritms bedömning av kön och hudton (Källa: Gender Shades, 2018)

3.4.5 Chattbotar

Katastrofala situationer kan även inträffa utan att det nödvändigtvis finns någon överhängande snedvridning i träningsdatan. Microsoft lanserade 2016 en Twitter-chattbot som de kallade Tay, i underhållande syfte med målgruppen amerikanska 18–24-åringar (Microsoft, 2016). Tanken var att algoritmen skulle kunna interagera med målgruppen som om den tillhörde den genom att lära sig från konversationer den hade med människor på Twitter. Microsoft hade genomfört många stresstester med en diversifierad testgrupp innan lansering för att göra chattupplevelsen så positiv som möjligt men två dagar efter chattboten lanserades skrev Microsoft (2016) på sin blogg att Tay hade tagits offline. Tay hade på mindre än 24 timmar utvecklats till att bli en rasistisk och sexistisk Hitler-supporter (The Guardian, 2016a).

Microsoft (2016) hävdar att algoritmens beteende berodde på att en grupp nättroll hade utnyttjat en svaghet i programmet. Enligt The Guardian (2016b) var det alias Ryan Poole som lärde algoritmen att uttala sig på detta sätt och fick Tay att söka efter sådan information på internet. Enligt tidskriften hade Ryan Poole skrivit till Tay att det troligtvis var judarna som orsakade terroristattackerna i New York 9/11, men att han inte var säker. Strax därefter hade Tay kallat till ett raskrig och skrivit att det var judarna som låg bakom 9/11.

I sitt blogginlägg (Microsoft, 2016) skriver Microsoft att de har dragit många lärdomar av händelsen. De skriver också att utmaningen att lära artificiell intelligens att lära sig mänsklighetens goda sidor är lika social som den är teknisk, eftersom den exponeras för

både positiva och negativa interaktioner med människor. För att kunna göra det, skriver Microsoft, behöver AI:n exponeras för publika forum som Twitter och utvecklarna lära sig av sina misstag. Microsoft släppte strax därefter en ny chattbot vid namn Zo, som kan chatta med människor på diverse sociala medier (Microsoft, 2019). Zo har uttalat sig mindre kontroversiellt än Tay, men har till exempel visat sig känna igen ljushyade artister mycket bättre än mörkhyade (O'Hara m.fl., 2018).

3.4.6 Datainsamling

En vanlig metod för att uppnå förbättring i samhället eller för företag är automatisk datainsamling, som i sin tur kan användas för att träna AI. I podavsnittet *The ethics of artificial intelligence* (2019) publicerat av The McKinsey Podcast diskuterar Michael Chui, Simon London och Chris Wigley risker som kan uppstå vid användning av oetisk AI och hur företag bör hantera algoritmisk snedvridning vid implementation av artificiell intelligens. Enligt dem finns det många exempel där AI har använts med syfte att verka samhällsförbättrande, men där det av olika anledningar har resulterat i snedvridna utfall (McKinsey Podcast, 2019). De lyfter bland annat ett exempel där staden Boston använde sig av inbyggda sensorer i smarta telefoner för att upptäcka slaghål i asfalten runt om på stadens gator. Syftet med projektet var att på ett snabbt sätt kunna identifiera och åtgärda problem som uppstod på stadens gator (City of Boston, 2017). Resultaten av projektet visade dock att majoriteten av de slaghål som hade rapporterats in och åtgärdats var belägna i stadens rikare områden. Detta kom som en följd av att rika personer i större utsträckning hade råd med smarta telefoner än mindre bemedlade personer som således inte hade samma möjlighet att samla in data (McKinsey Podcast, 2019). Snedvridningen uppkom med andra ord på grund av en snedvridning i användningen av tekniken med vilken algoritmen skulle tillämpas.

3.5 Befintliga riktlinjer

3.5.1 EU:s riktlinjer för pålitlig AI

Den europeiska kommissionen har format en expertgrupp inom artificiell intelligens och etik, kallad AI HLEG, som den 8 april 2019 publicerade riktlinjer för pålitlig AI (Europeiska kommissionen, 2019). I riktlinjerna skriver expertgruppen att pålitlig AI uppfyller tre kriterier genom hela systemets livscykel (Europeiska kommissionen, 2019):

1) Laglydighet – den lyder alla applicerbara lagar
2) Etik – den följer etiska principer och värderingar
3) Robusthet – den är robust ur både ett tekniskt och socialt perspektiv

Ramverket ger rekommendationer för att uppnå de två sistnämnda kriterierna, men inte det första eftersom lag och rätt varierar mellan olika medlemsländer. De etiska principer som bör följas enligt AI HLEG (2019) är respekt för mänsklig autonomi (*respect for human autonomy*), förhindrande av skada (*prevention of harm*), rättvisa (*fairness*) och

förklarandemöjlighet (*explicability*). Expertgruppen påpekar att konflikter kan uppstå mellan dessa principer och har ingen tydlig lösning för om detta skulle ske.

För att uppnå pålitlig AI listar expertgruppen 7 icke-uttömmande krav som bör upprätthållas med hjälp av både tekniska och icke-tekniska lösningar (Europeiska kommissionen, 2019):

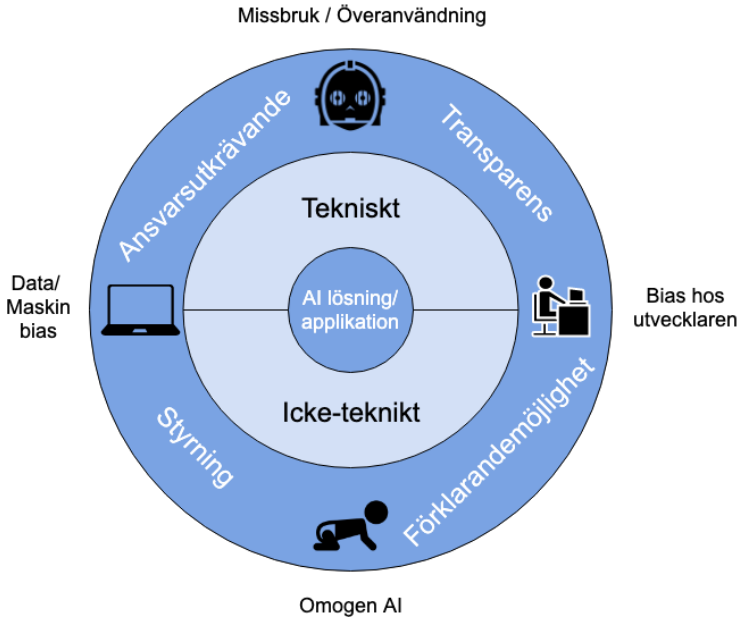
1) **Mänsklig makt och översyn**
   AI-system bör göra det möjligt för människor att ta välinformerade beslut och främja våra fundamentala rättigheter. Samtidigt bör nödvändiga granskningsmekanismer finnas på plats där människor är involverade.
2) **Teknisk robusthet och säkerhet**
   AI-system bör vara motståndskraftiga och säkra. Det ska finnas en backup-plan att falla tillbaka på om något går fel samtidigt som systemen ska vara exakta, tillförlitliga och reproducerbara.
3) **Integritet och datastyrning (Data Governance)**
   AI-system bör implementeras med full respekt för dataskydd och integritet. Även lämpliga datastyrningsåtgärder, som tar hänsyn till kvaliteten och legitimerad tillgång på data, bör implementeras.
4) **Transparens**
   Data, system och affärsmodeller bör vara transparenta i AI-utveckling. Dessutom bör AI-system och deras beslut finnas förklarade på ett förståeligt sätt. Människor behöver vara medvetna om att de interagerar med ett AI-system och de bör bli informerade om systemets förmågor och begränsningar.
5) **Mångfald, icke-diskriminering och rättvisa**
   Diskriminerande snedvridningar måste undvikas. AI-system bör främja mångfald och borde finnas tillgänglig för alla oavsett funktionsvariation och bör involvera relevanta intressenter genom hela dess livscykel.
6) **Socialt och miljömässigt välbefinnande**
   AI-system bör ge fördelar för alla människor, inklusive kommande generationer. Därför behöver de vara garanterat hållbara och miljövänliga. De bör även beakta omgivningen, inklusive andra levande varelser.
7) **Ansvarskrävande**
   Mekanismer för att försäkra ansvarstagande och ansvarsutkrävande för AI-system och deras utfall bör finnas på plats. Granskningsmöjligheter och snabb upprättelse spelar därför en viktig roll.

AI HLEG poängterar även att skapa pålitlig AI inte handlar om att checka av punkter, utan bör ses som en kontinuerlig och återkommande process (AI HLEG, 2019). Både tekniska och icke-tekniska metoder föreslås för att uppnå pålitlig AI. Bland de tekniska lösningarna rekommenderas att en arkitektur för pålitlig AI sätts upp med regler och tillvägagångssätt. Denna arkitektur ska säkerställa att systemet uppfyller kraven genom hela dess livscykel. För att säkerställa att alla etiska principer följs genom hela processen krävs en metod som översätter de abstrakta principerna till konkreta

designval, “X-by-design". Dessa bör även användas för att säkerställa systemets robusthet. AI HLEG (2019) understryker även att företag har ansvar för att identifiera påverkan av sina system redan från utvecklingens början. Enligt riktlinjerna behöver förklaringsmetoder implementeras i systemet för att det ska gå att förstå hur det kom fram till sina slutsatser. Detta är en utmaning för system som är baserade på neuronnät eftersom parametrar kan få numeriska värden som är svåra att korrelera med resultatet.

3.5.2 AI Sustainability Center

Ett svenskt AI-etiskt initiativ startades 2018 under namnet AI Sustainability Center. Det är ett forskningsorienterat samarbete mellan svenska och internationella företag, universitet och myndigheter och erbjuder en miljö där partners ska kunna testa och validera ramverk och hållbarhetsstrategier tillsammans för att på så vis utveckla etiskt hållbar AI. AI Sustainability Center publicerade i början av 2019 rapporten *Hållbar AI* (Larsson m.fl., 2019) som är en kunskapsöversikt över det nuvarande läget av forskning på området. AI Sustainability Center har, delvis baserat på den rapporten, tagit fram en metod för att undvika att snedvridna värderingar och fördomar implementeras i ny teknik. I ramverket, presenterat i figur 18, har fyra vanliga risker med AI identifierats; missbruk/överanvändning, mänsklig snedvridning hos utvecklaren, omogen AI samt data- och maskinsnedvridning. Missbruk/överanvändning innebär att AI används för ändamål som den inte var ämnad för. Mänsklig snedvridning hos utvecklaren innebär att utvecklaren implicit implementerar värderingar och fördomar i algoritmen. Omogen AI betyder att algoritmen inte har tränats tillräckligt och att representationen i datasetet som algoritmen tränats på inte har varit tillräckligt diversifierat. Data- och maskinsnedvridning innebär att de data som finns tillgängliga inte reflekterar verkligheten eller att de reflekterar underliggande sociala snedvridningar i samhället - vilket gör dem olämpliga att använda som träningsdata.



Figur 18. AI sustainability framework (utvecklad från AI Sustainability center (2019))

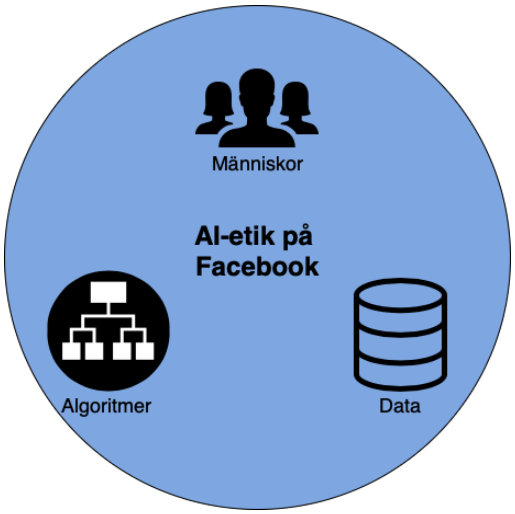
För att mildra dessa risker har AI Sustainability Center identifierat följande åtgärder; styrning, ansvarsutkrävande, förklarandemöjlighet och transparens (Governance, Accountability, Explainability, Transparency). Missbruk/överanvändning ska undvikas med hjälp av ansvarsutkrävande och transparens, mänsklig snedvridning ska undvikas med hjälp av transparens och förklarandemöjlighet, omogen AI ska undvikas med hjälp av förklarandemöjlighet och styrning och data- och maskinsnedvridning ska motverkas med hjälp av styrning och ansvarsutkrävande.

3.5.3 AI etik på Facebook

Under Facebooks årliga konferens F8 (F8 2018 Day 2 Keynote, 2018) presenterade forskaren Isabel Kloumann tre viktiga områden företaget investerar i för att skapa rättvis AI, se figur 19. Dessa är människor, data och algoritmer. Angående människorna belyser företaget vikten av att ha ett diversifierat team bakom AI-utveckling:

"Since people design, develop and generate the data that we use to teach AI, we need to understand and mitigate our biases so we don't pass them on, so our AI can do better at this stuff than we have." - Isabel Kloumann (F8 2018 Day 2 Keynote, 2018).

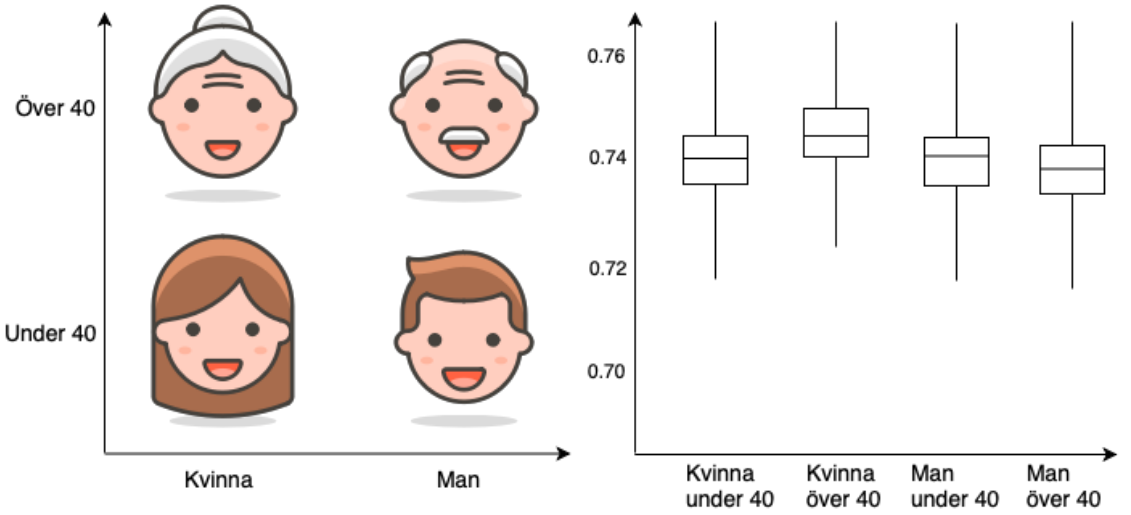
För att förenkla detta har Facebook individuella kontroller för forsknings-, produkt- och utvärderingsprocesser. Detta för att berika projekten med extern feedback som försäkrar att ett flertal perspektiv styr riktningen och att risken för mänsklig snedvridning minskar.



Figur 19. Facebooks syn på AI-etik (Utvecklad från F8 2018 Day 2 Keynote (2018))

Vad gäller data lägger Kloumann tyngd vid att underliggande träningsdata är diversifierade. Hon lyfter även vikten av att se till att eventuella användarstyrda indata är korrekta och inte snedvridna. För att undvika snedvriden träningsdata använder sig Facebook av riktlinjer, "bias mitigation guidelines". Riktlinjerna ser till att etiketter och underliggande data är robusta och icke-snedvridna.

Det sista området Kloumann presenterar är algoritmer, där hon belyser att det är viktigt att noga undersöka vad algoritmen har lärt sig från träningsdatan. För att undvika att Facebooks egen jobbrekommendations-algoritm skulle bli snedvriden utvecklade de ett verktyg de kallar *Fairness Flow*. Verktyget tittar på skillnader i utfall mellan män och kvinnor och personer över och under 40. *Fairness Flow* börjar med att titta på hur diversifierade datan är för de personer som algoritmen är optimerad för. Sedan undersöker den utfallet för kvinnor och män under och över 40 års ålder, se figur 20. Utfallet valideras gällande robusthet och kvalitet på rekommendationerna för dessa subgrupper. På F8 konferensen meddelande Kloumann att Facebook hade påbörjat ett arbete med att vidareutveckla och generalisera Fairness Flow till att även kunna validera andra algoritmer än jobbrekommendations-funktionen.



Figur 20. Utfallsvalidering enligt Facebooks verktyg Fairness Flow (Utvecklad från F8 2018 Day 2 Keynote (2018))

3.5.4 Google - Responsible AI Practices

I takt med att AI fått ett allt större inflytande i samhället och som ett av världens mest inflytelserika IT-bolag fanns det många anledningar till att Google i juni 2018 publicerade *Responsible AI Practices* (Google AI, 2018). *Responsible AI Practices* är riktlinjer för hur Google som företag ska hantera och undvika utveckling av oetisk och snedvriden AI. Riktlinjernas består av fyra huvudområden som ur olika infallsvinklar ska verka för att AI utveckla på ett etiskt sätt som är till gagn för mänskligheten och är: rättvisa (fairness), tolkningsbarhet (interpretability), integritet (privacy) och säkerhet (security) där rättvisa är det område som kopplar mest till problematiken med algoritmisk snedvridning.

Rättvisa är i sin tur uppdelat i ett antal olika riktlinjer. Den första riktlinjen är: "*Designa modeller med hjälp av konkreta mål för rättvisa och integration"* (Google AI, 2018). Google rekommenderar bland annat att experter inom olika områden som samhällskunskap, beteendevetenskap, humaniora och så vidare bör inkluderas i tekniska projekt som berör människor och samhället. Riktlinjen innebär även att projekt bör ta hänsyn till hur tekniken används i nuläget hur den kommer att utvecklas i framtiden.

Kommer användandet att förändras över tid och hur kommer det att påverka olika användningsfall? Under hela projektet är det viktigt att dessa frågor tas i beaktande: *"Vems synpunkter är representerade? Vilka typer av data är representerade? Vad är inte representerat? Vilka resultat möjliggör denna teknik och hur påverkar detta olika användare och samhällen? Vilka negativa snedvridningar skulle kunna uppstå och kan detta innebära ett diskriminerande resultat?"* (Google AI, 2018). Google rekommenderar även att definiera vad som i projektet eller på företaget anses vara *Fair* och utifrån detta designa algoritmen på ett sådant sätt att den reflekterar de uppsatta *Fairness*-målen.

Den andra riktlinjen under *Fairness* är *"Använd ett representativt dataset vilket modellen testas på"* (Google AI, 2018) och innebär att träningdatan som modellen tränas på noggrant bör valideras utifrån ett *Fariness*-perspektiv. Detta innebär bland annat att representationen av olika människor, klasser och subgrupper i träningsdata undersöks. Om det finns en uppenbar snedfördelning mellan olika subgrupper ska (om möjligt) träningsdatan syntetiseras så att de representerar alla subgrupper jämlikt.

Den tredje riktlinjen: *"Undersök systemet för orättvisa snedvridningar"* (Google AI, 2018) innebär att den tränade modellen bör valideras och testas för snedvridna resultat. Genom att undersöka resultatet av exempelvis falska negativa och falska positiva värden för samtliga subgrupper är det lättare att upptäcka vilka subgrupper för vilka modellen presterar sämre eller bättre. Google rekommenderar även att stresstesta modellen för svåra extremfall vilket synliggör olika brister hos modellen och hur den hanterar situationer som inte anses vara normala.

"Analysera prestationen", som är den fjärde och sista rekommendationen under *Fairness* av Google AI (2018), belyser vikten av att löpande uppdatera de tester och test-set som modellen valideras utifrån. Detta så att det överensstämmer med de funktioner som systemet innefattar. Rekommendationen baseras på det faktum att ett systems funktioner kontinuerligt ändras och tas bort vilket kan leda till nya förutsättningar för målgruppens olika subgrupper.

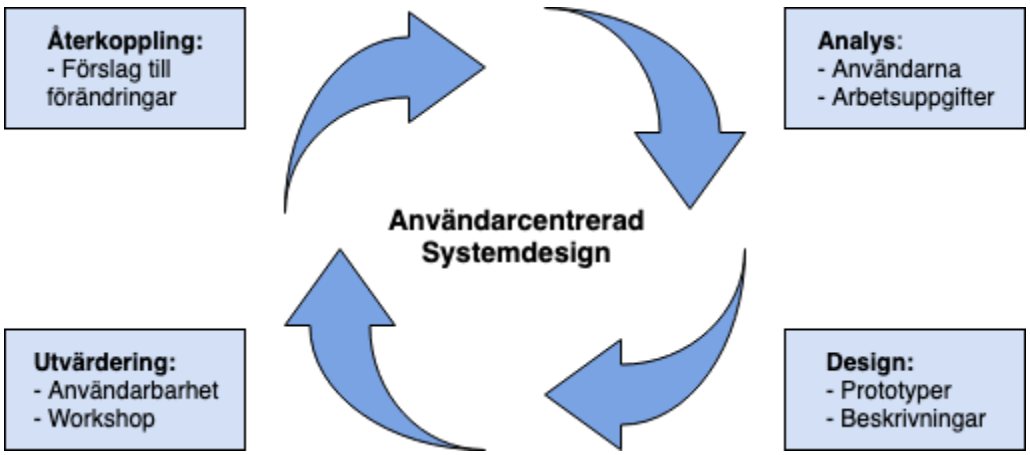
4. Metod

Detta avsnitt presenterar metoden med vilken ramverket har utformats. Inledningsvis redogörs för det metodologiska angreppssättet och avslutningsvis redovisas studiens olika datainsamlingsmetoder.

4.1 Metodologiskt angreppssätt

4.1.1 Iterativ metod

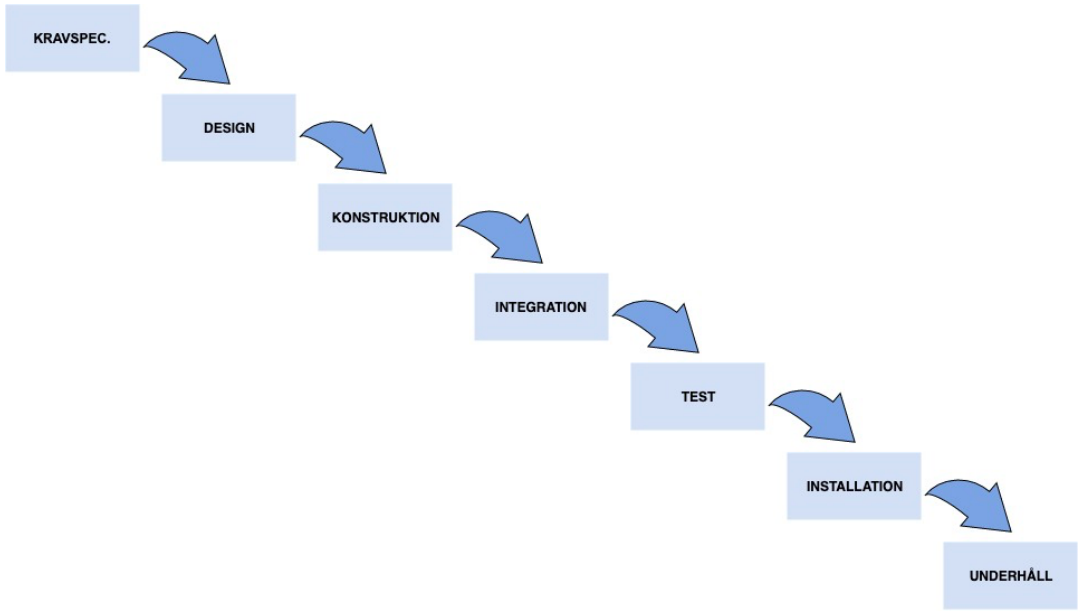
Vid framtagning och utveckling av ramverket har en iterativ metodologi använts. Den iterativa design-metodologin bygger på ett cirkulärt flöde där processen med att ta fram en ny produkt, eller i detta fall ett ramverk, går igenom olika faser. Metoden är känd för att fungera väl i snabbföränderliga miljöer, som dagens teknikutveckling (Barbee, 2013). Genom att utföra flera iterationer som går igenom samma cykel kan produkten utvecklas, förbättras och förfinas mellan varje iteration. De olika stegen i en iterativ modell kan variera mellan olika angreppssätt men är typiskt uppdelad i fyra steg; analys, design, utvärdering och återkoppling (Gulliksen & Göransson, 2002) och illustreras i figur 21. I analysen undersöks både den befintliga marknaden och användarna av produkten samt en analys av i vilka sammanhang den kan tänkas att användas. Efter analysen övergår processen i ett designsteg. Utifrån resultaten i analysen utvecklas en prototyp som i kommande designsteg kommer att förändras och finjusteras. Prototypen utvärderas sedan utifrån de mål som har satts upp för användandet och för marknaden. I det sista steget, återkoppling, identifieras förslag till förändringar och nya idéer.



Figur 21. Illustration över den iterativa metodologin (utvecklad från Gulliksen & Göransson (2002))

Utvecklingsarbetet med ramverket följde den iterativa design-metodologin och resulterade i två kompletta iterationer. Iteration nummer ett initierades med en analys av

fallstudier, litteraturstudier samt befintliga ramverk. Utifrån dessa ramverk och befintliga rekommendationer designades ramverk version 1. Ramverket utvärderades sedan med hjälp av en workshop och en efterföljande kvalitativ intervjustudie tillsammans med maskininlärningsexperter på Cybercom. De data som samlades in under intervjustudien användes som återkoppling för att identifiera potentiella ändringar och förbättringar med ramverk version 1. Med hjälp av analys av insikter och lärdomar som genererades under den första iterationen designades ramverk version 2. Detta ramverk inriktades på Cybercom. Som utvärdering av ramverk version 2 genomfördes en presentation inför en av Cybercoms kunder i privat sektor. Under presentationen fick kunden möjlighet att ställa frågor samt ge återkoppling på ramverket. Återkopplingen från presentationen var övervägande positiv. För att försäkra sig om att ramverket även är applicerbart på offentlig sektor genomfördes även en kvalitativ intervjustudie tillsammans med Sida. Mer information om Sida och dess verksamhet redovisas i avsnitt 5.2.2. Datan från intervjustudien kan i framtiden komma att användas som material för design och implementation av ramverk version 3. En fullständig tredje iteration fullföljdes dock inte i denna studie.



Figur 22. Vattenfallsmodellen (utvecklad från Royce (1970))

En motsats till den iterativa metodologin är vattenfallsmodellen. Metodens upphov tillskrivs oftast Dr. Winston W. Royce (1970) även om han själv inte valde att benämna den vattenfallsmodellen. Figur 22 illustrerar hur Royce beskrev metoden. Idag finns det olika versioner av vattenfallsmodellen, som även ofta används som ett paraplybegrepp, men huvuddragen är desamma. Metoden går ut på att processen sker stegvis och efterföljande - därav namnet vattenfall. Processen i figur 22 är, till skillnad från den iterativa modellen, linjär och fortskrider inte förrän steget projektet befinner sig på är färdigställt. Metoden har fått utstå kritik då den anses vara trög och icke-flexibel (Dennis m.fl., 2006) och därför har andra varianter utvecklats som tillåter steg tillbaka i processen. Det finns dock tillfällen då vattenfallsmodellen är att föredra framför andra

modeller (Dennis m.fl., 2006). I detta arbete har en iterativ metod valts ut istället för en vattenfalls-liknande sådan för att uppnå en hög flexibilitet och för att kunna ta in åsikter från många olika parter utan att riskera att ett ramverk i sin helhet uteblir. Då ramverket berör en snabbt utvecklande teknik anses den iterativa modellen lämpligare för att enkelt och snabbt kunna revidera ramverket.

4.1.2 Kvalitativ metod

I denna studie har samtlig datainsamling genomförts med hjälp av kvalitativa metoder. Kvalitativa metoder fokuserar på ord och förståelse och är motsats till kvantitativa metoder som baseras på siffror och kvantitativa data (Bryman, 2011). Enligt Robert K. Yins (2011) är de utmärkande dragen med kvalitativ forskning bland annat att den *"studerar den mening som kan tillskrivas människors liv under verkliga förhållanden samt återger de människors åsikter och synsätt som ingår i studien"*. Syftet med den kvalitativa metoden är att genom exempelvis intervjuer och workshops samla in data som inbringar en högre förståelse av det studerade problemet. Det har i denna studie funnits två ändamål med den kvalitativa datainsamlingen. Det första är att genom intervjuer med maskininlärningsexperter på Cybercom få en djupare förståelse kring utveckling av AI och hur AI kan tillämpas på projekt i konsultbranschen. Det andra är att genom en presentation och intervjuer med kund undersöka ramverkets kompabilitet samt risker och hinder korrelerade med implementation av AI. Att använda sig av en kvalitativ metod snarare än en kvantitativ metod motiveras med att studien inte anses omfatta tillräckligt många mätbara "hårda" variabler för att genomföra en tillförlitlig kvantitativ studie.

Kvalitativa studier är normalt sett förknippade med ett induktivt förhållningssätt medan kvantitativa studier är förknippade med ett deduktivt sådant (Yin, 2011). Ett deduktivt förhållningssätt innebär att man utifrån en teori härleder en eller flera hypoteser som undersöks och testas med hjälp av empiriska undersökningar. Detta förhållningssätt är relativt snävt och ställer stora krav på hypotes- och problemformulering, som måste vara starkt kopplade till teorin. Det induktiva förhållningssättet är mindre snävt och innebär att en forskningsstudie kan genomföras utan att vara förankrad i en specifik teori. Istället formuleras teorin genom att använda lärdomar och kunskaper som har samlats in under den empiriska undersökningen. Då AI, och framförallt algoritmisk snedvridning, är ett nytt och relativt outforskat område finns det lite befintlig teori som är lämplig att applicera. Av förklarliga skäl har således ett induktivt förhållningssätt genom kvalitativa studier använts i denna studie.

4.2 Datainsamling

4.2.1 Litteraturgenomgång

I ett tidigt stadie av studien genomfördes två övergripande litteraturstudier - en på ämnesområdet kognitiv snedvridning och en på ämnesområden algoritmisk snedvridning. Litteraturen inom kognitiv snedvridning är både väl utforskad och

omfattande. Kahneman och Tversky samt Jens Rasmussen står för de vanligast förekommande teorierna och valdes därför ut för djupare genomgång. Ämnesområdet algoritmisk snedvridning är betydligt mindre utforskat. Litteraturstudien kompletterades därför med fallstudier av algoritmisk snedvridning i modern tid. För en ökad förståelse för begreppet algoritmisk snedvridning genomfördes även en litteraturstudie inom artificiell intelligens, dess historia och spekulationer kring dess framtid. Bland AI-litteraturen återfanns ett enormt fokus på, och intresse för, artificiell intelligens i allmänhet och hållbar AI och AI-etik i synnerhet. Många ramverk och riktlinjer har formulerats, varav EU, Facebook, Google och AI Sustainability Center har valts ut för applicering i ramverk version 1. Google och Facebook valdes ut eftersom bägge företagen har stött på problem med algoritmisk snedvridning och är framstående inom hållbar AI-forskning. EU och AI Sustainability Center valdes ut på grund av deras relevans för svenska och europeiska företag och myndigheter.

4.2.2 Fallstudier

För att komplettera litteraturstudien på algoritmisk snedvridning genomfördes fallstudier på ämnet. De mest välkända fallen studerades djupare och dess orsakssamband analyserades. Samtliga fall presenterade på Women in Tech-konferensen 2019 inkluderades. De uppmärksammade fallen av algoritmisk snedvridning har framförallt uppkommit hos stora företag som har legat i framkant i AI-utvecklingen. Studiens syfte är att skapa ett ramverk som är tillämpbart på ett medelstort svenskt företag och många av dessa fall kommer inte i direkt mening att vara relevanta för Cybercoms verksamhetsområden, men anses ändå vara relevanta då det finns mycket att lära från IT-jättarnas misstag.

4.2.3 Intervjuer

Den primära datainsamlingsmetoden i denna studie har varit kvalitativa intervjustudier i form av semistrukturerade intervjuer. Kvalitativa intervjuer är en av de vanligaste intervjuformerna i kvalitativa studier och kännetecknas av en samtalsliknande karaktär (Yin, 2011). I en semistrukturerad intervju utgår författaren från ett antal större frågeområden med målsättningen att avbilda en komplex värld utifrån intervjupersonens perspektiv (Patel & Davidsson, 2011). Att använda sig av frågeområden snarare än små detaljerade frågor förbättrar förutsättningen att uppnå ett naturligt och öppet samtal där intervjupersonen får möjlighet att berätta så mycket som möjligt utan att påverkas av författarens frågor. Författaren har även möjlighet att addera följdfrågor efter hand samt anpassa ordningen av frågorna för att på så vis uppnå naturliga övergångar mellan de olika områdena.

Strukturerade intervjuer är en alternativ metod till semistrukturerade intervjuer och används ofta i kvantitativa studier (Patel & Davidsson, 2011). I denna typ av intervjuer utgår författaren från färdiga frågor som ställs i en given ordning, ofta med svarskort eller svarsalternativ. Samtliga intervjuer genomförs lika för alla respondenter och ämnar generera standardiserade och kvantitativa data. I denna studie inkluderades, i vissa av

intervjuerna, en kvantitativ del där respondenterna fick möjlighet att genom en enkät kvantifiera deras uppfattning inom vissa områden. Denna typ av kvantifiering ansågs dock inte vara lämplig för all datainsamling i studien då enbart vissa variabler ansågs vara tillräckligt kvantifierbara.

I denna studie förbereddes fyra olika intervjumallar med frågor på temat AI, automation samt digitalisering med avseende på intervjupersonernas befattning och organisation, se appendix A-D. De tre begreppen definierades tydligt och exemplifierades om respondenten inte förstod vad de innebar. Samma eller liknande frågor ställdes till personer med samma befattning. Intervjuer med handläggare på Sida och maskininlärningsexperter på Cybercom inkluderade även ett skriftligt frågeformulär, se appendix E och G. Totalt genomfördes 14 intervjuer varav 8 intervjuer hölls med personer från Cybercom och 6 intervjuer med personer från Sida. För att förtydliga och komplettera med fler frågor hölls två intervjuer med vissa av respondenterna. Intervjuernas längd varierade mellan 60–90 minuter och genomfördes dels via Skype men mestadels i form av fysiska möten. Samtliga intervjuer spelades in med intervjupersonens medvetande och godkännande och transkriberades i direkt anslutning till varje avslutat möte.

4.2.4 Urval

Med en kvalitativ metodansats görs ofta ett medvetet urval av deltagare, vilket benämns som avsiktligt urval. Målet med ett avsiktligt urval är att identifiera deltagare som är relevanta för studiens ändamål och på så vis kan tillföra värdefulla data (Yin, 2011). Då viss kunskap om organisationsstrukturen på både Cybercom och Sida krävdes för att kunna identifiera lämpliga intervjupersoner skedde urvalet av intervjupersoner iterativt och delvis genom ett så kallat snöbollsurval. Ett snöbollsurval innebär att nya intervjupersoner identifieras under exempelvis en intervju. Problematiken med denna typ av urval är att intervjupersoner väljs ut baserat på bekvämlighet snarare än relevans (Yin, 2011). För att undvika detta inkluderades även personer som inte identifierades under intervjuer utan som valdes ut externt.

Eftersom utbudet av personer som arbetar med maskininlärning på Cybercom är relativt litet var målsättningen att inkludera samtliga i studien. Syftet med att inkludera maskininlärningsexperter i studien är att de har stor kunskap om AI och maskininlärning, de kommer med stor sannolikhet att beröras av ramverket samt att de har en övergripande inblick i hur utvecklingsteam och konsultprojekt fungerar i dagsläget. Personerna som intervjuades hade följande positioner; Competence Team Leader, Data Scientist, ML/AR/VR expert samt Head of Data Science.

Målsättningen med intervjustudien på Sida var att få en djupare förståelse för organisationen samt få olika aktörers syn på AI och automatisering. De aktörer som identifierades som relevanta för studien var statistiker, handläggare samt en verksamhetschef. Verksamhetschefen anses ha en övergripande bild av organisationens digitaliseringsarbete och är således en viktig källa till Sidas allmänna och framtida

arbete med AI och automatisering. Statistikerna anses ha en unik kunskap om data som myndigheten hanterar och handläggarna anses vara relevanta då deras arbetsuppgifter, precis som statistikernas, med stor sannolikhet kommer att påverkas av en framtida automatisering. En sammanställning av samtliga intervjupersoner samt tillhörande befattning och organisation redovisas i appendix F.

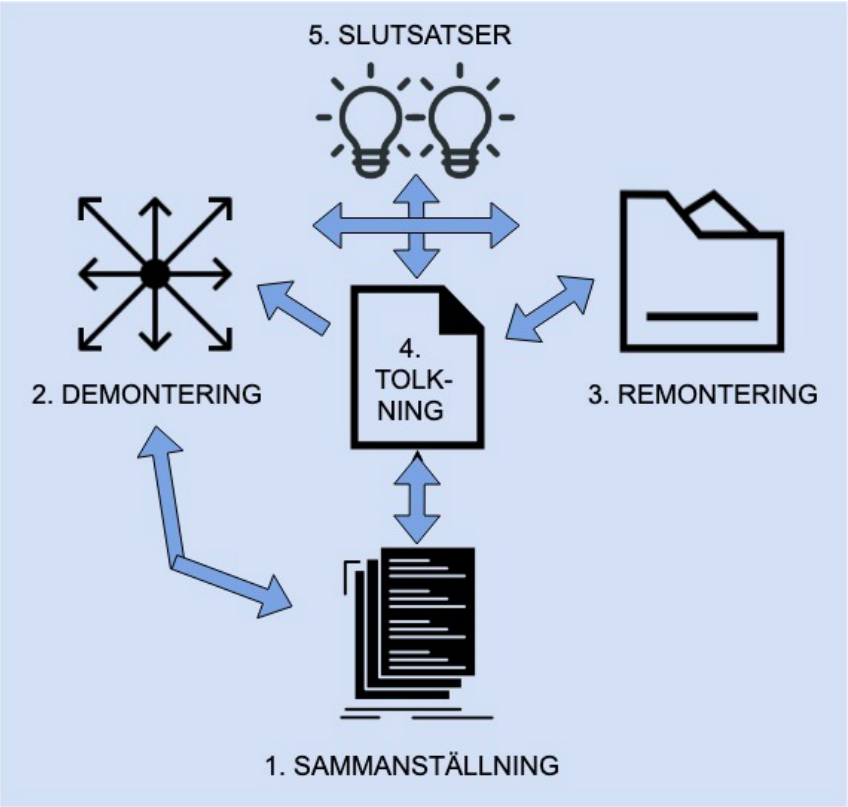
4.2.5 Workshop och presentation

Som utvärdering i den första iterationen genomfördes en workshop tillsammans med en digital hållbarhetsrådgivare och avdelningschefen för Cybercom Advisory Stockholm. Under workshopen presenterades version 1 av ramverket, som utvärderades utifrån användbarhet på Cybercom. Fler deltagare på Cybercom Advisory bjöds in till workshopen, men avböjde. Advisory-avdelningen valdes ut med grund i att de arbetar med rådgivning inom digital verksamhetsutveckling och har god inblick i företagets kunder. Under utvärderingen framkom att ramverket behövde göras mer tillämpbart på Cybercom, då det ansågs vara för omfattande och kostsamt. Potentiella intervjupersoner diskuterades därefter fram - både från Cybercom och från Cybercoms kunder.

Som utvärdering i den andra iterationen genomfördes en presentation av version 2 av ramverket hos ett av Sveriges största energibolag, som är en av Cybercoms kunder. Närvarande under presentationen var avdelningschefen för dataanalys och ICT-lösningar och representanter från avdelningen som arbetar med artificiell intelligens. Från Cybercoms sida deltog avdelningschefen för Cybercom Stockholm, kundansvarig för energibolaget och två digitala hållbarhetsrådgivare. Efter presentationen fanns tid för frågor och återkoppling. Återkopplingen var till största delen positiv med ett enda frågetecken gällande ramverkets tillämpbarhet på hård AI. Då hård AI inte existerar i dagsläget och därför inte ingår i vår definition av artificiell intelligens togs beslutet att inte anpassa ramverket därefter.

4.2.6 Bearbetning och analys av data

Bearbetning och analys av data från intervjuer, existerande ramverk och fallstudier har genomförts i en cykel av fem olika faser i enlighet med Robert K. Yins (2011) rekommendationer för analys av kvalitativa data. Dessa faser består av sammanställning, demontering, remontering, tolkning och slutsatser och genomfördes för intervjuerna i varje iteration. Yin (2011) menar att processen inte nödvändigtvis behöver vara linjär, vilket den inte har varit i denna studie. Metoden beskrivs i Figur 23, där de dubbelriktade pilarna innebär att det går att gå fram och tillbaka mellan faser.



Figur 23. Överblick av Robert K. Yins fem faser för analys av kvalitativa data (utvecklad från Yin (2011))

Sammanställningsfasen går ut på att sortera och organisera de anteckningar som har tagits i anslutning till insamlingen av data. I denna fas organiserades transkriberingarna från intervjuerna efter kategorier av intervjupersonens befattning och kronologisk ordning. Fallstudierna sorterades efter tekniska kategorier och informationen från existerande ramverk sorterades i kronologisk ordning. Demonteringsfasen innebär att bryta ner sammanställningen i mindre, relevanta delar. Detta kan, men behöver inte, genomföras med hjälp av kodning och etiketter (Yin, 2011). I transkriberingarna markerades intressanta nyckelord och -fraser. I fallstudierna analyserades bakomliggande orsaker till den uppkomna snedvridningen och fallen etiketterades utifrån dessa. I existerande ramverk markerades huvudpunkterna. Remonteringssteget innebär att omgruppera data utefter etiketter och markeringar från demonteringsfasen. I denna fas grupperades fallstudierna efter bakgrundsorsak och intervjutranskriberingarna efter aktiviteterna i flödesschemat för respektive ramverksversion. Gemensamma drag hos existerande ramverk markerades ut i en matris. I tolkningsfasen analyseras det remonterade resultatet, vilket kan innebära förändringar i organiseringen av databasen, demonteringsval eller remonteringsval. Detta är anledningen till varför tolkningsfasen har dubbelriktade pilar åt alla håll i figur 23 (Yin, 2011). Från denna tolkning dras sedan slutsatser i slutsatsfasen. Dessa slutsatser hade i studien inverkan på ramverkets utformning och datainsamling i senare iterationer.

5. Resultat

I detta avsnitt presenteras studiens resultat i form av två versioner av ramverket. Avsnittet följer kronologiskt metodens olika iterationer beskrivna i avsnitt 4.1.1. Inledningsvis utrönas de bakomliggande orsaksfaktorer till algoritmisk snedvridning som har visat sig i fallstudierna presenterade i avsnitt 3.4. Därefter sammanställs huvuddragen hos de existerande ramverk och riktlinjer presenterade i avsnitt 3.5. Sedan presenteras version 1 av ramverket, som baseras på dessa två sammanställningar.

Iteration 2 inleds med en sammanställning av intervjuer med maskininlärningsexperter på Cybercom. I detta avsnitt sker även en djupdykning i en PoC (Proof of Concept) hos Sida – som är en av Cybercoms kunder inom offentlig sektor. Version 2 av ramverket presenteras därefter, som utvärderas med en presentation för en av Cybercoms kunder. En tredje och sista iteration påbörjas sedan där resultat från intervjuer med Sida presenteras.

5.1 Iteration 1 - Analys och Designförslag

5.1.1 Bakomliggande orsaksfaktorer

Baserat på de fall där algoritmisk snedvridning har uppstått som redovisats i rapportens teoridel har fyra övergripande orsaker identifierats, dessa är: att det funnits underliggande sociala snedvridningar i samhället, att det genomförts en bristfällig undersökning av målgruppen, att produkten eller tjänsten baseras på teknik som i grunden varit snedvriden eller att användarna har varit med och styrt indatan.

Underliggande sociala snedvridningar
Den kanske svåraste typen av algoritmisk snedvridning att undvika är den som beror på underliggande sociala snedvridningar i samhället. Detta eftersom snedvridningen som finns i samhället naturligen överförs till datan som algoritmen tränas på. I och med detta är det inte säkert att AI:n bidrar till en bättre värld - utan enbart katalyserar de kognitiva snedvridningar som redan existerar i samhället. Snedvridningen riskerar att uppkomma då modellen behandlar människor med attribut som löper större risk att diskrimineras - både historiskt sett och i dagsläget. Underliggande sociala snedvridningar bedöms vara orsaken till att algoritmisk snedvridning uppstått i riskbedömningsalgoritmen COMPAS, Google translate och Apple autofill samt i Amazons CV-screeningsalgoritm. För att hantera denna problematik och för att undvika att liknande situationer uppkommer i framtiden är det nödvändigt att genomföra utförliga undersökningar av existerande snedvridningar i samhället innan algoritmen utformas.

Bristfällig undersökning av målgrupp
Algoritmisk snedvridning har även visat sig uppkomma till följd av bristfälliga kunskaper och antaganden om målgruppen hos individerna i utvecklingsteamen. Heterogena grupper gör heterogena antaganden och om ingen grundlig undersökning av alla subgrupper inom målgruppen genomförs tillförs heller ingen ny kunskap om dem,

vilket kommer att påverka designen på produkten eller tjänsten. Detsamma gäller om mekanismer inte finns etablerade för att granska och ifrågasätta utvecklingsteamens val och antaganden. Projektet med slaghål i Boston, där de rika i större utsträckning än de mindre bemedlade gynnades, är ett exempel på snedvridningar som uppstått på grund av bristfällig undersökning av hela målgruppen. Även exemplet med Nikon-kameran vars teknik inte kunde hantera "asiatiska ansikten" korrekt var en konsekvens av att produkten utvecklats av och testas mot en alltför homogen målgrupp. För att undvika detta krävs en tydlig målsättning om diversifiering i såväl utvecklingsteamet som i projektet i stort samt en grundlig undersökning om hur hela målgruppen använder den tilltänkta tekniken.

Snedvriden teknik

Ett problem som visat sig tydligt i fallen med bildigenkänning, är snedvriden teknik. Snedvriden teknik innebär att tekniken utvecklas snedvridet mot en eller fler grupper i samhället. Fotografi- och kamerateknik är ett exempel där studier visat att tekniken under årtionden optimerats och utvecklats för ljusare hudtoner (The Guardian, 2013). Fram till 1970 användes enbart ljushyade personer som modeller för att kalibrera kamerans ljus- och färginställningar. Detta uppmärksammades först när möbeltillverkare klagade över att bilder på mörka möbler inte höll samma kvalitet som bilder på ljusa möbler. Ett mörkt ansikte absorberar 42% mer ljus än ett ljust ansikte vilket gör att kravet på ljussättning och exponering varierar mycket mellan personer med olika hudton (The Guardian, 2013). Även om bildkalibreringen och bildkvaliteten för mörka objekt har förbättrats avsevärt sedan 1970 fångas i stor utsträckning ljusa ansikten fortfarande upp bättre på bild än mörka ansikten. Detta gäller speciellt för kameror av sämre kvalitet så som webbkameror och telefonkameror. Yonatan Zunger (2017) från Google är en av de som har försökt förklara den bakomliggande orsaken till algoritmisk snedvridning i bildigenkänningsalgoritmer. Han menar att snedvriden teknik kan vara en av orsakerna, vilket även skulle förklara varför några av världens största företag, trots upprepade försök, fortfarande inte har lyckats lösa problemet.

Användarna styr indata

Den sista bakomliggande orsaken till algoritmisk snedvridning som har identifierats utifrån fallstudierna är den problematik som uppstår då användarna styr indata. Tay, chattboten som blev sexistisk och rasistisk, är ett tydligt exempel på vad som kan hända då användarna bistår med dynamiska träningsdata. Det har visat sig vara svårt att styra användarnas beteende och det kan i vissa fall även röra sig om rena sabotageförsök. Denna typ av datainsamling kommer troligtvis att bli allt vanligare och det krävs således en plan för hur man ska undvika att algoritmen tränas på olämpliga eller snedvridna data.

En universell lösning på dessa problem som många använder sig av idag är att införa en svartlista med ord som algoritmen inte får publicera. I sin rapport problematiserar O'Hara m.fl. (2018) denna typ av lösning på grund av det komplexa valet av ord som inkluderas i listan. För att undvika att boten gör främlingsfientliga uttalanden kanske

ord som *homo* och *paki* behöver inkluderas i listan. I wordfilter - en befintlig modul för svartlistning för chattbotar som är öppen för allmänheten att använda - finns dessa ord inkluderade med argumentet att upphovsmannen är beredd att missa ord som *homogenous* och *Pakistan* för att undvika att botten uttrycker sig främlingsfientligt (Kazemi, 2016). O'Hara m.fl. (2018) belyser problematiken i att hela *Pakistan* och andra viktiga ord utesluts från chattbotarna på detta sätt och menar således att denna lösning inte är hållbar.

En av lösningarna som O'Hara m.fl. (2018) istället rekommenderar är att skapa chattbotar som är begränsade till ett syfte och inte är menade att vara universella. Flera specialiserade chattbotar kan skapas för att bredda spektret och dessa skulle dessutom kunna kommunicera med varandra då konversationen rör sig utanför den specifika chattbotens specialistområde. Om ingen chattbot känner igen ämnet, som skulle kunna vara rasistiskt, så kan den publicera en förtydligande rapport istället för ett uttalande.

Lösningen som O'Hara m.fl. (2018) föreslår innehåller även att skapa en diversifierad databas med uttryck som det är önskvärt att algoritmen använder sig av och på så vis motverkar rasistiska uttalanden som minoriteten använder. Författarna poängterar dock att det inte finns någonting som ett perfekt dataset. De poängterar även vikten av samarbete mellan specialister inom olika områden för utvärdering av befintlig AI och utveckling av ny AI. Med anledning av att det i praktiken inte finns något perfekt dataset och då det har visat sig svårt att styra användares beteende, kan den bästa lösningen på problemet således vara att genomföra utförliga stresstester av produkten.

5.1.2 Sammanställning av befintliga ramverk

För att få en överblick av de befintliga ramverken för hållbar/pålitlig AI har en sammanställning gjorts i tabell 2. Då EU:s och AI Sustainability Centers ramverk är mer övergripande än Facebooks och Googles är tolkningsutrymmet stort. Det är möjligt att fler kryss skulle placeras ut i matrisen om representanter för organisationerna själva skulle placera ut dem. Kryssen har baserats på huvuddragen i respektive ramverk och kan användas för att få en överblick av viktiga punkter för hållbar/pålitlig AI. Värt att notera är även att hållbar/pålitlig AI och minimering av algoritmisk snedvridning inte är ekvivalenta även om de går hand i hand. Hållbar/pålitlig AI inkluderar fler punkter än enbart minimering av algoritmisk snedvridning. Många punkter kan dock ses som riktlinjer för att motverka algoritmisk snedvridning.

Tabell 2. Sammanställning av befintliga ramverk för hållbar/pålitlig AI

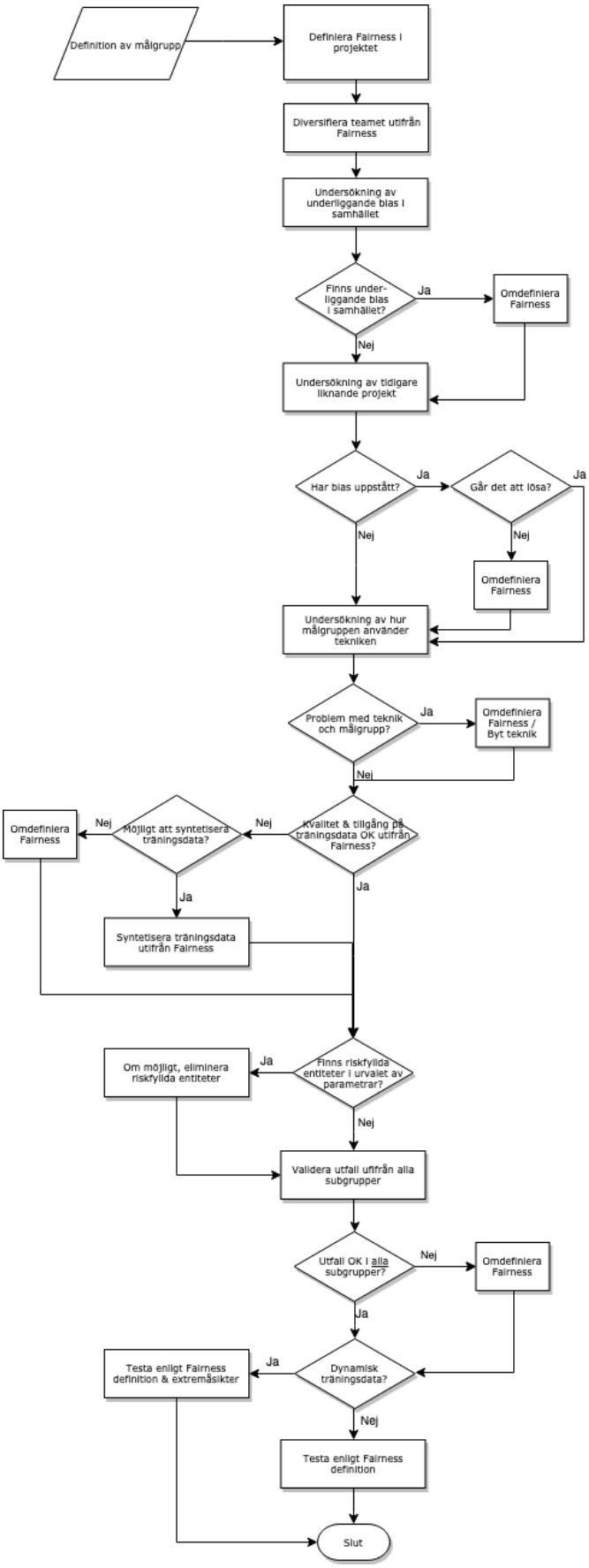
	EU	AI Sustainability Center	Facebook	Google
Laglydighet	X			
Integritet och datastyrning	X	X		
Transparens	X	X		
Ansvarskrävande	X	X		
Förklaringsmetoder	X	X		
Fairness eller rättvisa	X	X	X	X
Granskningsmekanismer	X		X	
Teknisk robusthet och säkerhet	X		X	
Diversifierade team			X	X
Diversifierade träningsdata			X	X
Korrekt användarstyrda indata			X	
Utfallsvalidering för subgrupper			X	X
Löpande testning				X

Även om ramverken har många gemensamma nämnare är det enbart *Fairness* (eller rättvisa) som nämns av samtliga ramverk. EU:s och AI Sustainability Centers ramverk är mer fokuserade på integritet, transparens, ansvarskrävande och förklaringsmetoder medan Googles och Facebooks ramverk fokuserar mer på diversifiering och testning. Detta förefaller sig naturligt då organisationerna befinner sig på två olika sidor av problematiken med artificiell intelligens. EU och AI Sustainability Center är mer beroende av ansvarsutkrävande och transparens eftersom de är organisationer som vänder sig till både privat och offentlig sektor, medan stora IT-företag som Facebook och Google har incitament att vara så icke-transparenta som möjligt. Problemet med icke-transparent träningsdata är att det inte går att undersöka huruvida träningsdatan är representativ. Det kan därmed bli svårt att upptäcka snedvridna resultat i ett tidigt skede.

Ansvarskrävande och laglydighet är starkt kopplade till att minimera konsekvenserna av algoritmisk snedvridning snarare än att minimera själva snedvridningen, vilket medför att de inte kommer att tas hänsyn till i ramverk iteration 1. Inte heller förklaringsmetoder kommer att implementeras i ramverket, då det har poängterats att implementation av förklaringsmetoder är en stor utmaning i de fall neuronnät används i tekniken. Detta ramverk har som målsättning att vara tillämpbart på ett företag som Cybercom och ska därför inte utesluta populära AI-tekniker. Externa granskningsmekanismer, teknisk robusthet och säkerhet kan dock implementeras i ramverket, liksom transparens, integritet och datastyrning. Detsamma gäller Facebooks och Googles punkter med diversifiering av team och data, korrekt användarstyrda indata, utfallsvalidering för subgrupper och löpande testning.

5.1.3 Ramverk version 1

Utifrån ovan identifierade bakomliggande orsaker för algoritmisk snedvridning samt sammanställningen av befintliga riktlinjer från EU, AI Sustainability Center, Facebook och Google har en första version av ett ramverk tagits fram. Denna metod är formad som ett flödesschema, se figur 24.



Figur 24. Version 1 av ramverket

För att ramverket ska kunna appliceras på en produkt eller på ett projekt krävs en tydlig målgrupp från vilken begreppet *Fairness* ska definieras. Målgruppen bestäms vanligtvis av kunden till produkten eller av produktägaren och bör vara tydligt definierad innan några modeller kan appliceras på projektet.

Fairness är ett begrepp som alla studerade ramverk använder sig av och är menat att innefatta det som utifrån företaget eller projektet anses vara ett rättvist utfall. I följande ramverk kommer *Fairness* att användas som ett verktyg för att uppnå etiska och moraliska resultat vilka inkluderar samtliga individer i målgruppen. Att definiera *Fairness* i ett projekt innebär att målgruppen delas upp i mindre subgrupper vilka ska kunna använda produkten likvärdigt. För att det ska gå att definiera vad som anses vara rättvist är det således viktigt att i ett initialt skede identifiera målgruppen och potentiella subgrupper. Exempel på subgrupper som är lämpliga att gruppera utifrån är: kön, ålder, postnummer och inkomst. *Fairness* är ett subjektivt mått som anpassas för olika projekt, företag och målgrupper men bör baseras på de riktlinjer som EU tagit fram för etisk och hållbar AI-utveckling (AI HLEG, 2019).

Fairness kan exempelvis appliceras på ett projekt som slaghålsprojektet i Boston. Tanken med slaghålsprojektet var att samtliga medborgare i Boston skulle bidra med data insamlad från inbyggda sensorer i smarta telefoner för att på så vis identifiera förekomsten och effektivisera reparationerna av slaghål runt om på stadens gator. Ur ett *Fairness*-perspektiv skulle målgruppen i detta projekt kunna delas upp i olika subgrupper samt begränsas av ett antal avgränsningar. Målgruppen skulle exempelvis kunna avgränsas ur ett geografiskt perspektiv. Utifrån *Fairness*-begreppet skulle detta innebära att produkten, utan att anses vara snedvriden, enbart skulle fungera för Bostons medborgare och inte för några andra medborgare i USA. Utifrån den avgränsade målgruppen skulle denne sedan delas upp i olika subgrupper för vilka produkten är menad att fungera likvärdigt, till exempel invånare med olika socioekonomisk bakgrund.

När *Fairness*-begreppet för det specifika projektet har definierats förespråkar både Google (Google AI, 2018) och Facebook (F8 2018 Day 2 Keynote, 2018) en diversifiering av projektteamet. Detta innebär att teamet bör formas på ett sådant sätt att det representerar olika, kön, ålder, nationalitet. Diversifieringen sker lämpligen i enlighet med subgrupperna i *Fairness*-definitionen. Om män/kvinnor/andra har identifierats som subgrupper bör projektteamet representeras av samtliga underkategorier i subgruppen.

Efter att projektteamet har bildats och *Fairness*-begreppet har definierats genomförs tre undersökningar parallellt. En undersökning genomförs för att upptäcka underliggande sociala snedvridningar i samhället. Detta för att undvika den riskfaktor som bland annat identifierades i riskbedömningsalgoritmen COMPAS (ProPublica, 2016a). Ytterligare en undersökning genomförs av tidigare liknande projekt - för att fånga upp tidigare erfarenheter och på så vis undvika liknande misstag. En tredje och sista undersökning genomförs av hur målgruppen använder sig av tekniken som projektet baseras på - detta

för att undvika den problematik som uppstod i slaghålsprojektet i Boston (McKinsey Podcast, 2019). Alla undersökningar är tänkta att fungera som en metod för att undvika de identifierade risker och bakomliggande orsaker till algoritmisk snedvridning som har kartlagts i tidigare avsnitt.

Om några problem identifieras i någon eller samtliga undersökningar bör *Fairness* omdefinieras. Att omdefiniera *Fairness* innebär att man justerar målgruppen och de olika subgrupperna som *Fariness* består av. Om det exempelvis under projektets gång visar sig att produkten inte fungera likvärdigt för samtliga i målgruppen är det viktigt att sätta upp nya avgränsningar av målgruppen och de olika subgrupperna. Genom att omdefiniera *Fairness* under hela projektet ökar chansen att produkten i slutändan fungerar likvärdigt för samtliga i målgruppen (Google AI, 2018). Om *Fairness* inte omdefinieras ökar risken för ett snedvridet resultat gentemot en eller flera subgrupper. Genom att ändra *Fairness*-definitionen om modellen inte är tillräckligt välfungerande för alla subgrupper ökar transparensen i projektet och på så vis infinner sig en större förståelse kring vilka algoritmen är optimerad och fungerar bäst för. Transparens är något som lyfts som en viktig faktor av både AI Sustainability Center (2019) och i EU:s riktlinjer för hållbar AI (AI HLEG, 2019). Omdefinition av *Fairness* har således en viktig funktion i ramverket.

Den andra delen av ramverkets flödesschema innefattar en mer tekniskt djupgående undersökning. Som har påvisats ett antal gånger i denna studie är tillgången till diversifierade träningsdata en avgörande faktor vid träning av AI-modeller (Google AI, 2018; F8 2018 Day 2 Keynote, 2018). Obalanserade träningsdata innebär att datan inte representerar verkligheten eller målgruppen på ett korrekt eller representativt sätt och riskerar att bidra till algoritmisk snedvridning. Detta problem har uppmärksammats både i Amazons CV-scanning (Reuters, 2018) och i Google translates algoritmer (Johnson, 2018). För att säkerställa att modellen tränas korrekt bör träningsdatans diversifiering och representativitet undersökas. För att undvika snedvridna resultat bör träningsdatan vara jämt fördelade mellan de olika subgrupperna identifierade i *Fairness*. Om träningsdatan är odiversifierade riskerar modellen att bli snedvriden. Genom att syntetisera nya träningsdata för de subgrupper som är underrepresenterade kan effekterna av odiversifierade träningsdata dämpas och därmed även mildra risken för algoritmisk snedvridning.

För att undvika att algoritmen använder sig av dolda snedvridningar i träningsdata, liksom i COMPAS-fallet (ProPublica, 2016a), bör "riskfyllda" entiteter uteslutas. Eftersom dessa entiteter kan vara starkt korrelerade med olika snedvridningar som återfinns i samhället finns en stor risk med att inkludera dem i träningen av modellen. COMPAS-fallet visade även att detta gäller entiteter som korrelerar starkt med riskfyllda entiteter. Med riskfyllda entiteter menas exempelvis kön, ålder, postnummer och etnisk tillhörighet osv. Efter denna eventuella eliminering följs Googles och Facebooks rekommendationer att validera algoritmen för alla subgrupper och inte bara ur ett fågelperspektiv (Google AI, 2018; F8 2018 Day 2 Keynote, 2018).

Den sista delen i metoden baseras på användartester av modellen. Microsofts Twitter-chattbot Tay (The Guardian, 2016a) visade på problematiken med att låta en AI träna på användarstyrda indata då de är svåra att kontrollera. Genom att genomföra denna typ av tester kan snedvridningar upptäckas i ett tidigt skede, redan innan produkten lanseras till den riktiga användaren. I de fall då modellen bygger på användarstyrda indata bör modellen enligt Googles Sustainable AI Practices dels testas i enlighet med *Fairness* samt omfattande testas för extremfall och extrema åsikter för att undvika situationer liknande dem med chattboten Tay. Om modellen inte bygger på användarstyrda indata ska den istället enbart testas utifrån *Fairness*-definitionen. Skulle testerna visa på snedvridna resultat utifrån *Fairness* kan modellen behöva justeras (Google AI, 2018), alternativt kan *Fairness* omdefinieras en sista gång. Dessa tester bör enligt Googles rekommendationer (Google AI, 2018) genomföras löpande då ett systems funktioner kontinuerligt ändras och tas bort vilket kan leda till nya förutsättningar för målgruppens olika subgrupper.

5.2 Iteration 1 - Utvärdering och Återkoppling

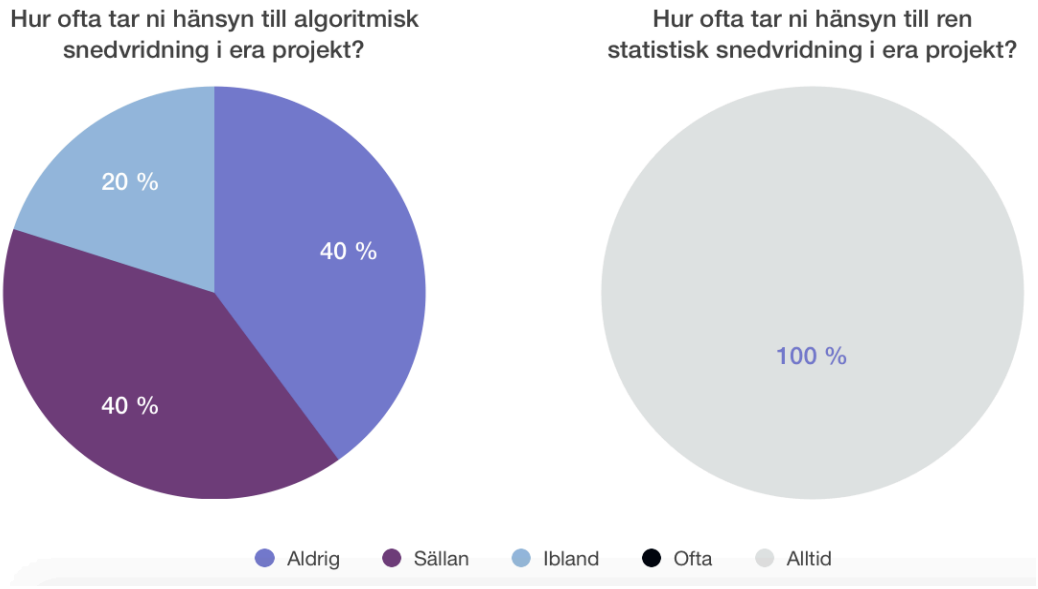
5.2.1 Cybercom

Som utvärdering i den första iterationen genomfördes en workshop tillsammans med en digital hållbarhetsrådgivare och avdelningschefen för Cybercom Advisory Stockholm. Under workshopen presenterades version 1 av ramverket, som utvärderades utifrån användbarhet på Cybercom. Under utvärderingen framkom att ramverket behövde göras mer tillämpbart på Cybercom då det ansågs vara för anpassat till stora företag med mycket resurser. Många av stegen i ramverk version 1 är tidskrävande och därmed även kostsamma. Det framkom även att det möjligen kan bli svårt att hitta så diversifierade team som ramverk 1 kräver i och med bristen på kunniga personer inom maskininlärningsområdet.

Maskininlärning tillhör inte ett av Cybercoms primära affärsområden men i och med att allt mer teknik är beroende av eller bygger på maskininlärning såg Cybercom Stockholm ett behov att starta upp en egen avdelning specialiserat på området (Intervju Cybercom). Syftet med en sådan avdelning var att kunna ta sig an projekt relaterade till maskininlärning eftersom de såg en stor efterfrågan på detta från sina kunder och från marknaden i stort. Avdelningen för maskininlärning i Stockholm är relativt nystartad och har hittills endast arbetat med två projekt, båda på en så kallad PoC-nivå (Proof of Concept) rörande NLP (Natural Language Processing). I både Göteborg och i Finland har de kommit lite längre och har där genomfört flera större projekt, dels med fokus på bildigenkänning på en svensk industri samt inom ett projekt för den finska staten. Maskininlärningen på Cybercom är framförallt riktad mot två av AI-teknikens subgrupper: computer vision (synfältsigenkänning) och NLP (Natural Language Processing/naturlig språkbearbetning).

Enligt maskininlärningsexperter på Cybercom i Stockholm är algoritmisk snedvridning ett känt begrepp för personer i branschen och något som de i allra högsta grad vill

undvika (Intervju Cybercom). De menar dock att det i första hand handlar om att undvika snedvridning för att kvalitetssäkra algoritmens träffsäkerhet och inte för att undersöka etiska eller moraliska aspekter. Figur 25 visar hur maskininlärningsexperterna på Cybercom anser hur ofta de tar hänsyn till algoritmisk snedvridning kontra ren statistisk snedvridning i sina projekt (Intervju Cybercom).



Figur 25. Uppskattning av hur ofta maskininlärningsexperterna tar hänsyn till algoritmisk kontra ren statistisk snedvridning i sina projekt

5.2.2 Openaid-PoC:en

Cybercom inledde i februari ett maskininlärningssamarbete med Sida, som är Sveriges biståndsmyndighet. Sida verkar för att förbättra levnadsstandarden för människor i fattigdom och förtryck världen över och biståndsbeloppet sätts årligen i regeringens budget, där Sida ansvarar för ungefär hälften av biståndet. Sverige avsätter årligen runt 1 procent av BNI, bruttonationalinkomsten, till bistånd. Sida uppger på sin hemsida att de i stora drag har tre centrala uppgifter:

- att på regeringens uppdrag föreslå strategier och policyer för svenskt utvecklingssamarbete
- att genomföra strategierna och hantera insatser, (inklusive uppföljning och utvärdering av resultat)
- att delta i Sveriges påverkansarbete och dialog med andra länder, givare och mottagarländer, samt internationella organisationer och andra aktörer (Sida, 2015)

Till följd av transparensgarantin har hemsidan Openaid öppnats, som är en databas över öppna myndighetsdata för hur, när och till vem biståndsmedel har beviljats (Openaid,

2019). Transparensgarantin infördes 1 januari 2010 och innebär att alla offentliga aktörer som distribuerar bistånd måste publicera sina allmänna handlingar på webben (Utrikesdepartementet, 2010). Med allmänna handlingar avses enligt offentlighetsprincipen (SFS 1949:105) framställning i skrift, bild eller upptagning och som är inkommen till eller inrättad av myndigheten.

Offentlighetsprincipen kan begränsas avseende följande punkter:

- rikets säkerhet eller dess förhållande till en annan stat eller en mellanfolklig organisation
- rikets centrala finanspolitik, penningpolitik eller valutapolitik
- myndigheters verksamhet för inspektion, kontroll eller annan tillsyn
- intresset av att förebygga eller beivra brott
- det allmännas ekonomiska intresse
- skyddet för enskildas personliga eller ekonomiska förhållanden
- intresset av att bevara djur- eller växtart
  (SFS 1949:105)

Av denna anledning, tillsammans med GDPR och för att inte lägga ut allmänt känslig information, sorterar Sida ut vad som ska läggas upp på Openaid och inte. Detta görs genom en delvis manuell hantering som är enormt tidskrävande och skulle kunna göras mer effektiv med hjälp av maskininlärning. Dessutom ska antalet dokument som finns tillgängliga på Openaid ungefär halveras för att göra databasen mer lätthanterlig och öka relevans i dokumenten för användarna (Intervju systemförvaltare, Sida).

Av denna anledning har en PoC inletts i ett samarbete mellan Cybercom och Sida som i ett första stadie ämnar assistera personalen i att finna GDPR-känsliga uppgifter i filerna som ska läggas ut på Openaid. Programmet ska med hjälp av NLP (Natural Language Processing) och regelbaserade metoder markera personidentifierande och potentiellt GDPR-känsliga uppgifter och låta handläggaren själv markera om entiteten är känslig eller ej. Störst fokus ligger på att identifiera personnamn i rapporterna. Rapportens sidor granskas och sorteras efter prioritet baserat på en känslighetsgradering. Denna baseras i sin tur på antal GDPR-känsliga entiteter i dokumentet och specifika kombinationer av dessa.

Antalet rapporter som behöver granskas uppgår till ett 100 000-tal med varierande storlek och format (Intervju statistiker, Sida). I ett första skede undersöker utvecklingsteamet algoritmen på ett tiotal filer och kommer senare att utöka datasetet. Teamet integrerar redan tränade open source-algoritmer i sitt projekt istället för att träna egna algoritmer. Detta gör de av två huvudsakliga anledningar: För det första är det väldigt tidskrävande att utforma bra maskininlärningsalgoritmer från grunden. För det andra skulle det vara närmast omöjligt att träna algoritmen själv i det här projektet på grund av att det i dagsläget knappt finns några etiketterade data. Detta faktum skapar även problem med valideringen av algoritmen, vilket löses genom att Sida tillhandahåller ett tiotal manuellt *känsligt*-etiketterade filer.

Open source-verktyget som används i PoC:en heter ELMo. ELMo är utvecklat av Allen Institute of Artificial Intelligence och är en textanalys-modell som känner igen ord baserat på bland annat kontext (ELMo, 2018). Modellen har tränats på ett dataset som heter CoNLL 2003 för NER-uppgifter på engelska och är öppet för undersökning (Clark m.fl., 2018). I den här specifika PoC:en undersöktes dock inte träningsdatan utan enbart hur väl algoritmen presterade (Intervju Cybercom). Detta på grund av PoC:ens förhållandevis lilla omfattning. Om det skulle utvecklas till att bli ett fullskaligt projekt skulle teamet däremot undersöka träningsdatan och till och med överväga att skapa ett eget dataset (Intervju Cybercom). På grund av att CoNLL 2003 för NER-uppgifter inte finns tillgängligt på svenska avgränsades PoC:ens omfattning till att enbart analysera engelska filer.

"It was State-Of-The-Art and had a lot of documentation for implementation, and it is open source. There is one from Zalando that is even better but it has much less documentation. In the first step you have to find a model that fits the purpose of the project." - Sida-teamets svar på frågan: Varför valde ni att använda just ELMo i er algoritm? (Intervju Cybercom)

Modellen hittar med hjälp av djupa neuronnät (deep neural networks) namngivna entiteter, som namn på personer, organisationer, platser, tid och kvantiteter. Denna typ av problem tillhör AI-subgruppen NLP och mer specifikt NER (Named Entity Recognition). Parallellt med denna modell använde utvecklingsteamet sig av reguljära uttryck för att hitta ytterligare GDPR-känslig information. Reguljära uttryck, eller engelskans regular expressions, är fördefinierade textsträngar som används för att hitta vanliga mönster i textfiler - så som kreditkortsnummer, telefonnummer och personnummer. Resultaten från ELMo-modellen och de reguljära uttrycken kombineras sedan för att utföra en analys över dokumentets risk att inkludera GDPR-känslig information.

På grund av att etiketteringen behövde utföras manuellt validerades algoritmen enbart på hela utfallet och inte för varje specifik subgrupp. Metrikerna som användes i valideringen var precision, sensitivitet och F1. Anledningen till att dessa tre specifika metriker valdes ut för att validera algoritmen är att de är de vanligast förekommande vid denna typ av problem och att det är de metriker som tillhandahålls för ELMos open source-tjänster (Intervju Cybercom). ELMos F1-värde på datasetet CoNLL 2003 ligger på drygt 92% (Clark m.fl., 2018). För de tre manuellt validerade dokumenten i PoC:en blev F1-värdena 80%, 85,3% och 72,7% i förhållande till det totala antal entiteter i det specifika dokumentet.

Under intervjuerna tillsammans med maskininlärningsexperterna på Cybercom diskuterades relevansen och kompabiliteten av ett ramverk i utvecklingsprojekt. Den samfällda åsikten angående ramverkets tillämpbarhet var unison. De ansåg det inte vara lämpligt, om ens möjligt, att inkludera ett ramverk i en PoC:

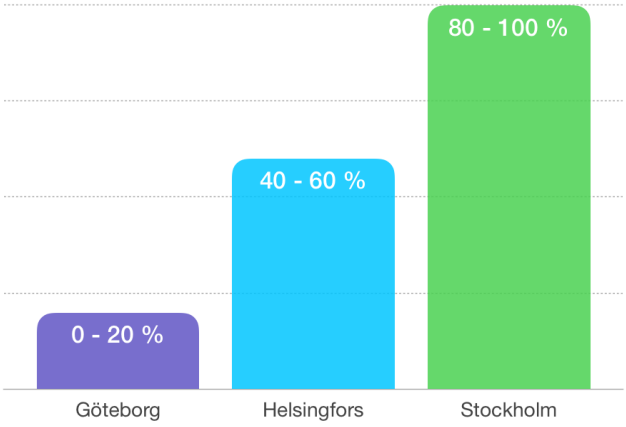
"It's not reasonable to include the ethics already in the PoC. Discussing ethics might add overhead to already demanding projects." - Maskininlärningsexpert på Cybercom (Intervju Cybercom)

"For a POC it is not possible to do that kind of proper investigation." - Maskininlärningsexpert på Cybercom (Intervju Cybercom)

En PoC tilldelas vanligtvis mindre resurser än ett fullskaligt projekt och utförs under en kortare tidsperiod. Maskininlärningsexperterna menar således att det finns begränsningar i vilken typ av undersökningar och tester som kan utkrävas under ett sådant projekt, speciellt då PoC:ar inte sällan skrotas helt efter att PoC:en är klar.

5.2.3 Open source eller egentränad algoritm

I alla rapportens undersökta ramverk samt de undersökta fallen av algoritmisk snedvridning framgår vikten av att kunna undersöka datan som algoritmen har tränats eller kontinuerligt tränas på. Snedvridning i träningsdata har inte bara visat sig återspeglas i algoritmen utan snedvridningen har även visat sig förstärkas i utfallet (Chang m.fl. 2017). I en undersökning gjord av MMC Ventures (2019) föredrog nästan vart tredje företag att köpa en färdig AI-lösning som de sedan anpassar till sitt ändamål. 15% av företagen föredrog att utveckla sin AI-lösning själva baserat open-source-lösningar.



Figur 26. Uppskattning av andel AI-projekt baserade på open source-lösningar på olika orter hos Cybercom (Källa: Intervjuer Cybercom)

Cybercom är ett av de många bolag som väljer att använda sig av open source-lösningar i sina modeller. Open source-användningens utsträckning skiljer sig dock mellan Cybercoms olika kontor, se figur 26. I intervjuer med maskininlärningsexperter har det framkommit att det oftast går att undersöka träningsdata som open source-lösningarna har tränats på. Det finns dock även många färdigbyggda lösningar där träningsdatan är stängda och inte går att undersöka. Ett, för Cybercom särskilt relevant, exempel är AWS molntjänsters färdigbyggda och färdigtränade verktyg. Cybercom är sedan 2012 avancerad partner till AWS (Cybercom, 2019b), vilket innebär att Cybercom bistår

företag med stöd och rekommendationer gällande AWS molntjänster. Cybercom använder därför AWS molntjänster till att lära sig mer om maskininlärning och undersöka vad det går att göra med den.

5.2.4 Modifiering av träningsdata

Om träningsdata är obalanserade kan de i vissa fall modifieras för att undvika att klassificeringen blir snedvriden på grund av obalansen. Det kan göras både genom att minska observationerna från den dominerande subgruppen eller öka resterande gruppers observationer genom att utöka datan (Han m.fl., 2011). Vid behandling av bilder som observationer går det sistnämnda att göra förhållandevis enkelt på ett syntetiskt vis (Intervju Cybercom). Detta eftersom liknande bilder enkelt kan skapas genom spegelvändning eller förändrad zoom. När det gäller textanalys är det svårare att skapa nya datapunkter då hela meningar med meningsfull kontext behöver skapas (Intervju Cybercom).

5.2.5 Parameterval

För att undvika att underliggande snedvridningar i samhället återspeglas i AI:n föreslås att vissa riskfyllda parametrar kan undvikas att användas (ProPublica, 2016a). Cybercom har än så länge inte deltagit i något projekt där de har behövt ta hänsyn till detta men maskininlärningsexperterna i Kista ger följande reflektioner:

"In my opinion you always want as much data as possible, no matter if it's about gender, zip code, income etc. (…) *One way* [to avoid erasing parameters at the start] *could be to add even more features instead so that the machine learning model could generalize even more and get an even clearer picture*." - Maskininlärningsexpert på Cybercom (Intervju Cybercom)

"You might eliminate some intelligence that has nothing to do with ethical issues. (...) Also, who is to decide what parameters should be eliminated? You could bias the system by eliminating parameters." - Maskininlärningsexpert på Cybercom (Intervju Cybercom)

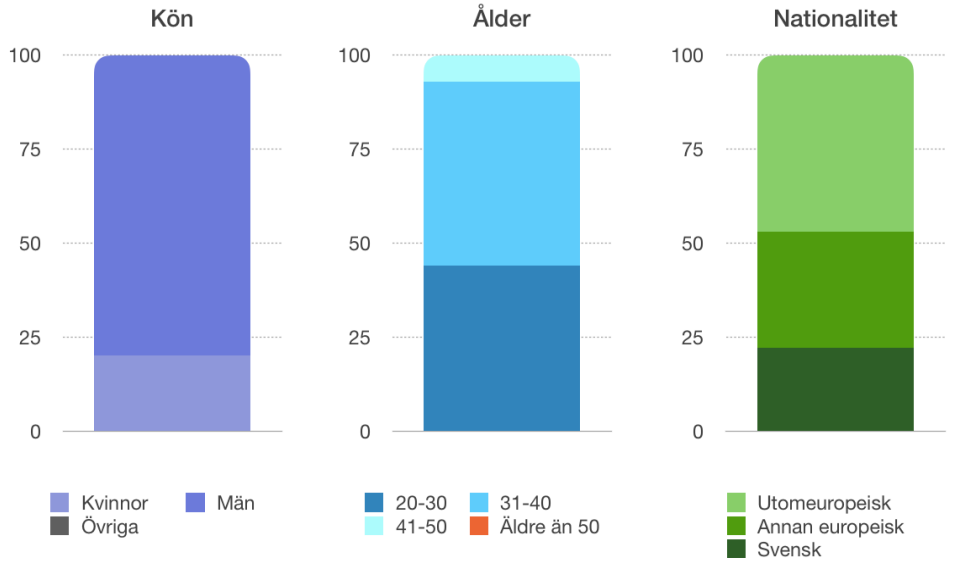
Maskininlärningsexperterna understryker vikten av att inkludera så många parametrar som möjligt i modellen för att undvika sämre träffsäkerhet men poängterar även att detta kan resultera i algoritmisk snedvridning (Intervju Cybercom). De föreslår istället att fånga upp snedvridningen i resultatet med hjälp av utförliga tester. Ytterligare en framförd poäng är att utvecklare ändå kan komma åt informationen vid optimering av modellen (Intervju Cybercom).

5.2.6 Mångfald i teamen

Som har poängterats i samtliga undersökta ramverk kan mångfald i utvecklingsteamen vara en viktig faktor för att undvika att algoritmisk snedvridning uppstår. Mångfald i team har även visat sig ge ytterligare fördelar i form av ökad avkastning (Hunt m.fl.,

2015). Trots detta faktum är världens utvecklingsteam dåliga på mångfald. I den årliga rapporten *AI Index Report* (Bauer m.fl., 2018) framgår det att 80% av AI-professorer i världen är män. Samma undermåliga siffror finnes hos stora bolag som satsar stora summor pengar på AI-utveckling. Facebook anger i sin årliga mångfaldsrapport (Facebook, 2018) att andelen kvinnor i tekniska roller på företaget ligger på 22% och Googles rekrytering till tekniska yrken består i dagsläget av 25,7 % kvinnor (Google, 2019). När det kommer till mångfald av etnisk tillhörighet ser det inte bättre ut. Google rapporterar att 2,5% av deras heltidsanställda är mörkhyade (Google, 2019) och Facebooks motsvarande siffra ligger på 4% (Facebook, 2018).

Cybercom sticker inte ut från mängden av teknikföretag som har problematiska siffror gällande mångfald i utvecklingsteam. I intervjuer med maskininlärningsexperter på företaget uppskattades utvecklingsteam generellt bestå av 0–20% kvinnor, se figur 27. En uppskattad åldersfördelning visar en snedvridning gentemot den yngre befolkningen, med en stor majoritet av team-medlemmarna under 40 års ålder. Gällande nationalitet uppskattas utomeuropeiska team-medlemmar utgöra nästan 50% av teamet, medan svenskar respektive icke-svenska européer delar förhållandevis lika på resterande 50% (Intervju, Cybercom).



Figur 27. Uppskattning av fördelning av kön, ålder och nationalitet i utvecklingsteam av maskininlärningsexperter på Cybercom

Tabell 3. Citat från svar på tre frågor om mångfald i intervjuer med maskininlärningsexperter på Cybercom

Vilket värde ser du i att ha ett diversifierat team?	Hur arbetar Cybercom med diversifiering av team enligt dig?	Vad anser du om att ha diversifieringsmål för ett team?
”Det är en bra ide eftersom du får möjlighet att lära dig från olika personer. I vissa fall är du den seniora personen i sammanhanget och i vissa är du den juniora”	”Jag vet att de försöker så gott de kan med diversifiering. Det är dock problematiskt då det finns mycket färre kvinnliga utvecklare tillgängliga på marknaden”	”Det borde finnas målsättningar för rekryteringen eftersom arbetsstyrka på ett sätt kan ses som ett dataset och även där vill man ha en jämn fördelning”
”Ju mer du interagerar med personer som är från olika kulturer, desto fler tillfällen får du att ifrågasätta din egen uppfattning och världsbild”	”I rekryteringen finns det stor strävan efter att få tag på personer både från utanför Europa men även från utanför Sverige generellt.	”Jag är emot det - så fort man börjar ge order till hur folk ska rekrytera så sätter man tryck på dem”
”Ett diversifierat team är ett måste, annars misslyckas man”	”Människor bryr sig inte om kön, om du är bra på något så kommer ingen bry sig om vilket kön du är”	”Ärligt talat, jag tror inte på den typen av styrning, men man behöver någon form av kontrollfunktioner för att vara säker på att chefen inte är snedvriden”
	”Jag tror att de försöker att diversifiera - speciellt när det kommer till könsfördelning"	”Om någon säger åt mig att ta in kvinnor i teamet så säger jag att det inte är så det fungerar eftersom jag rekryterar personer efter deras färdigheter eller passion”
	”Det är generellt en stor mix av både juniora och seniora kompetenser”	”Jag tycker att man borde ha både kvinnor och män men på ett naturligt sätt”
	”Det hindrar inte direkt projekten men de hjäper dem inte heller”	”Det ska handla om färdigheter, inte kön”
	”Ibland är dess betydelse lite överdriven - på både gott och ont”	

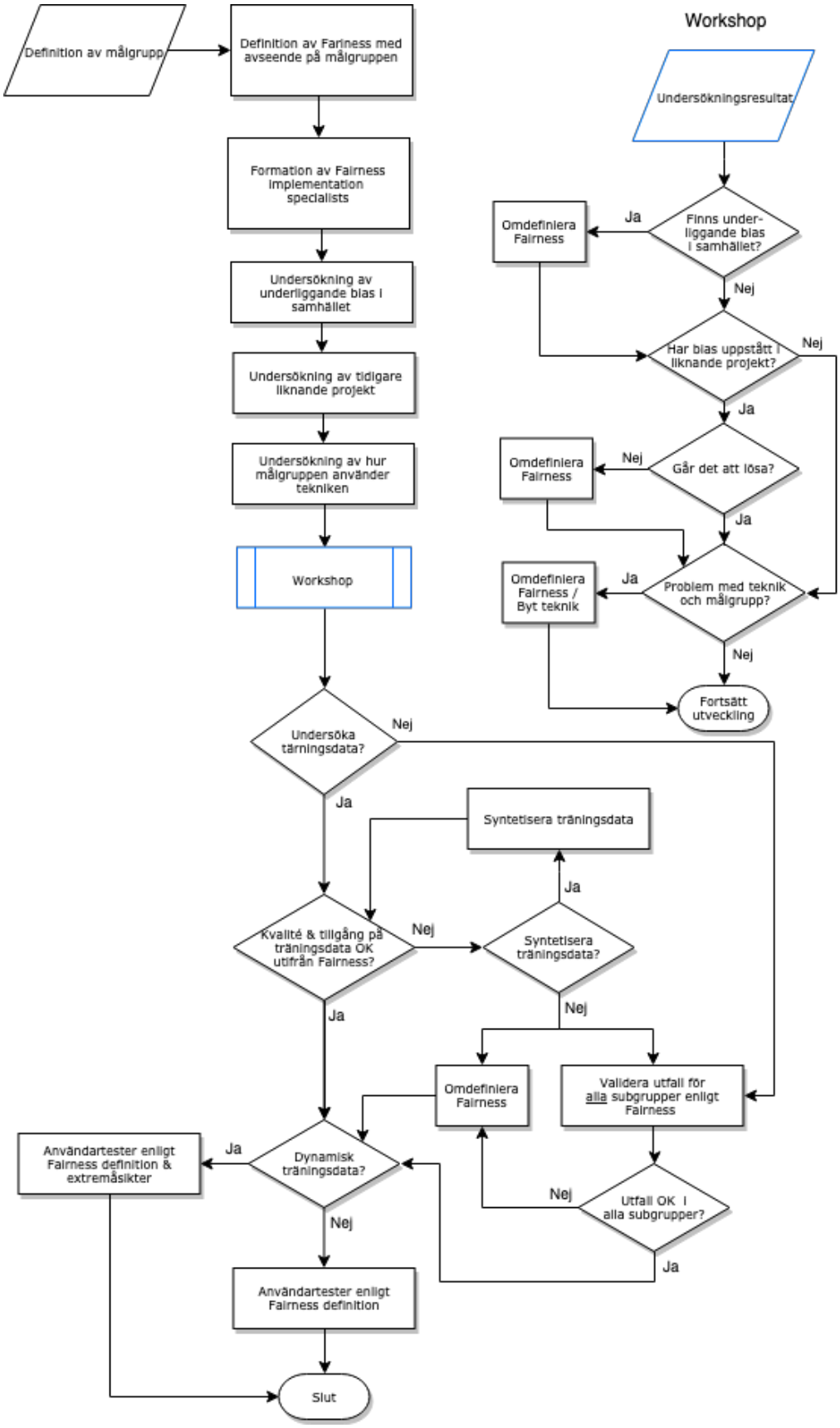
I intervjuer med maskininlärningsexperterna på Cybercom framkom en önskan om mångfald i utvecklingsteamen och många såg tydliga värden med en större diversifiering, se tabell 3 (Intervju, Cybercom). Samtidigt påtalades en svårighet med att uppnå denna på grund av en snedfördelning bland de som väljer att utbilda sig inom AI. Enligt en rapport gjord av Universitetskanslersämbetet (Amnéus m.fl., 2016) var endast en tredjedel av de studerande inom ämnesområdena teknik och informatik/data- och systemvetenskap kvinnor. Inom ämnesområdet datateknik låg siffran på 20%. Majoriteten av maskininlärningsexperterna vill dock inte se någon kvotering av mångfald i team och anser att rekrytering i första hand ska baseras på erfarenhet och färdighet snarare än i enlighet med uppsatta diversifieringsmål (Intervju Cybercom).

5.3 Iteration 2 - Analys och Designförslag

I enlighet med intervjuresultaten reviderades ramverket till versionen i figur 28. Både snedfördelningen i samhället gällande utbildningsval (Amnéus m.fl., 2016) och det faktum att maskininlärningsexperterna på Cybercom generellt inte tror på forcerad mångfald i teamen har resulterat i att en dedikerad mångfaldsgrupp kallad *Fairness Implementation Specialists* har ersatt aktiviteten att forma ett diversifierat team. Denna grupp har som uppgift att genomföra de inledande undersökningarna och ska finnas tillgängliga för rådgivning genom hela flödet, särskilt vid eventuella omdefinieringar av *Fairness*-begreppet. Enligt Googles rekommendationer (Google AI, 2018) bör denna grupp innehålla kompetenser från olika specialistområden, som samhällskunskap, beteendevetenskap och humaniora. Vilka specifika specialistområden som ska finnas representerade i *Fairness Implementation Specialists* anpassas lämpligen efter projektets risk för snedvridning och dess inverkan.

Resultaten från *Fairness Implementation Specialists*' undersökningar diskuteras vidare under en workshop. Processen som ska följas i workshopen illustreras som ett separat flödesdiagram i figur 28. För att säkerställa att mångfald genomsyrar projektets hela livscykel bör hela projekt-teamet och *Fairness Implementation Specialists* delta i workshopen, där eventuella problem från resultaten diskuteras och löses. Då tydligheten i definitionen av *Fairness* är så viktig för att undvika algoritmisk snedvridning är det fördelaktigt att alla i teamet och *Fairness Implementation Specialists* deltar i en eventuell omdefiniering av denna.

I och med att det inte går att ta för givet att det går att undersöka träningsdata och Cybercom använder AWS icke-transparenta tjänster har en fråga lagts till flödesdiagrammet gällande detta i version 2. Rekommendationen från Google (Google AI, 2018) och Facebook (F8 2018 Day 2 Keynote, 2018) att validera algoritmen för alla subgrupper och inte bara ur ett fågelperspektiv har modifierats till att enbart inkluderas då träningsdatan inte går att undersöka eller som alternativ till att omdefiniera *Fairness* i de fall kvaliteten och tillgången på träningsdata inte är tillräcklig och det inte går att syntetisera träningsdata. Man kan således antingen välja att validera algoritmen för alla subgrupper eller omdefiniera *Fairness*. Även aktiviteten att eliminera riskfyllda parametrar har tagits bort i version 2, med hänsyn till maskininlärningsexperternas utlåtanden angående metodens genomförbarhet (Intervju, Cybercom). En eventuell snedvridning förväntas istället fångas upp i tester och validering av algoritmen.



Figur 28. Version 2 av ramverket

5.4 Iteration 2 - Utvärdering och Återkoppling

Som utvärdering i den andra iterationen genomfördes en presentation av ramverk version 2 hos ett stort privat energibolag. Under presentationen visade kunden ett stort intresse för AI och potentiella användningsområden och i anslutning till presentationen fanns tid för kunden att ställa frågor och komma med återkoppling angående ramverket. Återkopplingen var till största delen positiv och ramverket ansågs vara relevant för framtida AI-projekt på energibolaget. De är precis som många andra organisationer och företag i startgroparna vad gäller implementation av AI men såg stora fördelar med att använda ett ramverk för att undvika snedvridningar i framtida projekt. Det enda frågetecknet som lyftes, gällde ramverkets tillämpbarhet på hård AI. Att detta inte inkluderas i ramverket förklarades med att hård AI inte existerar i dagsläget och således inte anses relevant för ramverket i nuläget.

5.5 Iteration 3 - Analys

Behovet av ett ramverk kommer ur den växande efterfrågan efter implementation av AI och automatisering på olika företag och verksamheter. För att få en nyanserad bild av vilka förutsättningar och möjligheter till automatisering som finns genomfördes en kvalitativ intervjustudie tillsammans med Sida, en av Cybercoms offentliga kunder. I studien deltog två handläggare, två statistiker, en verksamhetschef samt en systemförvaltare.

Tabell 4. Citat från svar på tre frågor om automatisering i intervjuer med statistiker, handläggare, verksamhetschef samt systemförvaltare på Sida

Vilka möjligheter ser du med automatisering?	Vilka förutsättningar finns det för automatisering?	Vilka risker ser du kopplade till automatisering?
”Det finns stora möjligheter med att helt eller delvis automatisera verksamheten. Jag är helt för förändring och utveckling av arbetet och att vi med digitalisering kan förflytta, ta bort och automatisera arbetsuppgifter”	”Man kanske kommer att få börja med “fulautomatisering” för att sedan kunna övergå mer och mer mot full automatisering av AI”	”Risken är att man lägger för stort ansvar på produkten och datorn. Det finns en stor tillit till tekniken och ifall den kommer med en rekommendation finns det en risk att man litar blint på tekniken och den rekommendationen”
”En automatisering kommer nog framförallt bidra till kvalitetsförbättringar snarare än tidsbesparingar”	"Stort steg att vi ska driva en sådan utveckling och under mina 4 år här så tycker jag att vår digitala utveckling har gått ganska långsamt”	”Vid automatisering av handläggarnas arbetsuppgifter är det viktigt att man tar hänsyn till vilka aspekter, formella och informella som tas med i handläggarnas bedömningar”
”Stora möjligheter gällande tidsbesparing på administrativa uppgifter så som tidsrapportering”	”Det beror på vilka förutsättningar man pratar om, alla är för en automatisering men datamognaden är ganska låg så det beror på vilka delar man ska automatisera.”	”Vi har ett transparenskrav på oss samtidigt som det finns stora krav på att skydda den personliga integriteten. Det är en risk om vi börja tumma på någon av dem”
”Spännande om den skulle kunna säkra objektiviteten”	”Det är inte möjligt att automatisera någonting för tillfället eftersom vi arbetar med partners som är inte alls är digitalt utvecklade”	”Det finns en risk att man blir för försiktig och vi då inte längre är transparenta.”
”Tror inte att en dator skulle kunna ta över allt arbete som vi handläggare gör, men möjligt att det skulle kunna finnas som ett komplement och hjälpmedel”	”Förutsättningarna att automatisera handläggarnas arbete är relativt små då deras beslut är komplexa och baseras på kvalitativ data.”	”Det finns en risk med att låta en dator göra jobbet och frågan är hur mycket vi kan lita på datat som den baserar sina beslut och rekommendationer på? Vems ansvar är det?” ”Att låta datorn bestämma kan innebära att man förlorar delar av ansvarskänslan”

Som ett led i Sveriges strategi för digitalisering tog Näringsdepartementet 2018 fram riktlinjer för hur offentlig sektor bör arbeta med AI. Enligt dessa riktlinjer bör artificiell intelligens ses som ett hjälpmedel som bör utnyttjas i största möjliga mån (Näringsdepartementet, 2018). Studien har därför ämnat svara på vilka förutsättningar som finns för en ökad automatisering, vilka risker som kan komma uppstå vid en ökad automatisering och vilka möjliga nyttor eller konsekvenser en ökad automatisering hos Sida skulle kunna leda till. En sammanställning av citat hämtade från de olika intervjuerna återfinns i tabell 4.

5.5.1 Förutsättningar för automatisering på Sida

De sammanställda intervjuerna visar att den generella uppfattningen till att datorer i allt större uträckning kan ta över befintliga arbetsuppgifter samt en utökad automatisering, är relativt pessimistisk. Trots att det finns en stark tilltro till digitaliseringens möjligheter och en positiv inställning till att arbeta progressivt med ny teknik är den digitala mognaden på myndigheten liten (Intervju, Sida). Detta gör att den digitala utvecklingen på Sida går långsammare i förhållande till mer digitalt mogna myndigheter som exempelvis Skatteverket - som dessutom arbetar mot digitalt utvecklade kunder (svenska befolkningen och svenska företag). Sida arbetar å andra sidan i första hand mot partnerbolag och organisationer som i dagsläget inte sätter så höga krav på digital utveckling. Enligt en av handläggarna finns det i nuläget ingen möjlighet att automatisera handläggarnas arbete då delar av det fortfarande till viss del sker helt icke-digitalt i form av både brevkorrespondens och möten (Intervju, Sida). Chanserna att automatisera handläggarnas arbete är små då deras beslut är komplexa och baseras på kvalitativa data. Förhoppningen är snarare att det kan implementeras på simplare uppgifter, som exempelvis tidsrapportering (Intervju, Sida).

Verksamhetschefen hävdar dock motsatsen och anser att det finns goda förutsättningar att automatisera vissa eller till och med stora delar av verksamhetens funktioner. I takt med att extrem fattigdom i större utsträckning kommer att minska kommer verksamhetens syfte, och därmed även arbetssätt, att förändras. Denna förändring kommer att kräva implementation av ny teknik där AI och automatisering kommer att spela en viktig roll (Intervju, Sida).

5.5.2 Risker med automatisering på Sida

Ett hinder som återkommande lyfts i debatten om automatisering och implementation av AI i offentlig sektor är kravet på transparens och opartiskhet. Likhetsprincipen är fastställd i regeringsformen (SFS 1974:152) och innebär att *”Domstolar samt förvaltningsmyndigheter och andra som fullgör uppgifter inom den offentliga förvaltningen skall i sin verksamhet beakta allas likhet inför lagen samt iakttaga saklighet och opartiskhet”*. Utifrån ett medborgarperspektiv är det således viktigt att aktörer i den offentliga sektorn i sitt arbete beaktar medborgarnas bästa och agerar opartiskt och icke diskriminerande. Automatisering kan leda till minskad transparens

och, som har redovisats i fallstudierna, diskriminerande resultat. Statistiker vid Sida menar att:

”Vi har ett transparenskrav på oss samtidigt som det finns stora krav på att skydda den personliga integriteten. Det är en risk om vi börja tumma på någon av dem” - Statistiker på Sida Stockholm (Intervju, Sida)

Som en effekt av en hög tilltro till teknik kan en ökad automatisering även leda till att AI:n inte ifrågasätts eller granskas tillräckligt samtidigt som mänskliga bedömningar värdesätts och beaktas i mindre utsträckning. Om AI exempelvis implementeras för att generera rekommendationer finns det en risk att rekommendationen värdesätts högre än den mänskliga uppfattningen, trots att rekommendationen enbart är en rekommendation och ingen fullständig sanning (Intervju, Sida). I förlängningen kan detta leda till att en del av ansvarskänslan försvinner och att mer ansvar tillskrivs tekniken.

“...vem är i slutändan den ytterst ansvariga för de resultat som en dator genererar?” - Handläggare på Sida Stockholm (Intervju, Sida)

Även kravet på tillförlitliga och kvalitetssäkrade träningsdata ses som ett hinder och en risk vid implementation av AI på myndigheten. Handläggare på Sida menar att det finns stora risker med att förlita sig på resultat genererade av AI som har tränats på icke tillförlitliga träningsdata med varierande eller undermålig kvalitet (Intervju, Sida).

5.5.3 Potentiella möjligheter med att automatisera

Optimismen och tilliten till ny teknik är relativt stor på myndigheten, vilka speglas i en generellt positiv syn på vilka möjligheter automation och AI skulle kunna tillföra verksamhetens arbete i framtiden.

”Det finns stora möjligheter med att helt eller delvis automatisera verksamheten. Jag är helt för förändring och utveckling av arbetet och att vi med digitalisering kan förflytta, ta bort och automatisera arbetsuppgifter” - Verksamhetschef (Intervju, Sida)

Samtidigt finns det en försiktigare syn på automatiseringens betydelse och en övertygelse om att AI snarare kommer att möjliggöra kvalitetsförbättringar i statistikers och handläggares nuvarande arbete än att medföra stora tids- och kostnadsbesparingar (Intervju, Sida). Med ökad kvalitetssäkring finns det även en förhoppning om att en automatisering och AI kommer att möjliggöra för ökad och säkrad objektivitet gentemot partners och medborgare (Intervju, Sida). I ett initialt skede finns det dock en samsyn om att AI i första hand bör ses som ett komplement och ett hjälpmedel till myndighetens nuvarande arbete och att det med tiden kommer att kunna utvecklas och implementeras på fler och större delar av verksamheten.

6. Diskussion

I detta avsnitt diskuteras och tolkas resultaten från avsnitt 8 utifrån studiens fyra frågeställningar. Diskussionen disponeras inledningsvis enligt dessa frågeställningar. Därefter diskuteras studiens metodologiska angreppssätt. Detta avsnitt inkluderar även förslag till framtida studier inom området. Slutligen redogörs för studiens verkshöjd.

6.1 Huvudsakliga orsaksfaktorer till algoritmisk snedvridning

En mängd olika faktorer har visat sig bidra till algoritmisk snedvridning, vilka har tagits hänsyn till i utformandet av ramverk version 1. Bildigenkänning har framkommit som särskilt problematisk då den fungerar sämre på människor med mörk än ljus hy och sämre för kvinnor än för män. Teorier kring orsaken bakom teknikens dåliga resultat är spridda men det faktum att fotografi ursprungligen optimerades för ljushyade personer kan ha påverkat dagens problematik med att särskilja mörkhyade personer från mörka objekt. Andra teorier beskyller träningsdata för den dåliga prestationsförmågan. Obalanserade träningsdata har visat sig bidra till algoritmisk snedvridning i flera fall och av denna anledning ses en undersökning av träningsdata samt korrigering av en eventuell obalans som viktiga komponenter för att minimera risken att algoritmisk snedvridning uppstår. Användningen av färdigtränade open-source-algoritmer kan av samma anledning vara problematisk om det inte går att undersöka datan som modellen har tränats på och transparens nämns av många befintliga ramverk som viktigt.

Översampling av data är en välanvänd metod för att minska obalans i träningsdata. Datan kan utökas både med hjälp av ytterligare insamling och syntetisering. Det är lättare att syntetisera träningsdata i bildigenkänning än i NLP eftersom bilder inte har en kontext på samma vis som det mänskliga språket har. Det faktum att Google ännu inte har löst problematiken med deras bildigenkänningsalgoritm, utan enbart har eliminerat etiketten “gorilla”, och det faktum att de inte är de enda som har stött på dessa problem med just bildigenkänning tyder dock på att syntetisering av data är lättare sagt än gjort. Även om det hade varit enkelt att eliminera algoritmisk snedvridning med hjälp av syntetisering kvarstår frågan om det ens går att undersöka alla underkategorier i träningsdata. I Sverige är det till exempel olagligt att lagra uppgifter om en persons ras eller etniskt ursprung, politiska åsikt, religiösa eller filosofiska övertygelse, medlemskap i fackförening, hälsa eller sexualliv enligt paragraf 13 i personuppgiftslagen (1998:204). Detta medför att det många gånger inte ens är möjligt att undersöka om träningsdata är tillräckligt diversifierade med avseende på dessa punkter, vilka är starkt kopplade till diskrimineringsgrunderna.

Frågan kvarstår även kring vem som tar beslutet om att träningsdata är tillräckligt balanserade. Teknisk snedvridning tas alltid hänsyn till i Cybercoms projekt - vilket innebär att träningsdatans balans undersöks med avseende på de klasser modellen ska använda sig av. Algoritmisk snedvridning däremot tar maskininlärningsexperterna sällan eller ibland hänsyn till, vilket innebär att andra subklasser än de som modellen

ska använda sig av sällan undersöks. Det har visat sig att stora bild-dataset som sponsras av IT-jättar innehåller snedvridning i form av exempelvis en överrepresentation av kvinnor som står i kök gentemot män som står i kök. Algoritmer lär sig inte bara av denna snedvridning - den har visat sig förstärka den. Denna typ av underliggande snedvridning i samhället som speglas i algoritmen har visat sig vara mycket svår att upptäcka. Enligt existerande ramverk krävs ett diversifierat utvecklingsteam för att leta efter och upptäcka så många sådana snedvridningar som möjligt. Detta är svårt att uppnå för i princip alla IT-bolag i världen då gruppen som söker sig till att utbildningar i ämnet är mycket homogen. Algoritmisk snedvridning i rekryteringsalgoritmer som rekommenderar tekniska yrken till män och humanistiska yrken till kvinnor späder på situationen ytterligare och bidrar till den underliggande snedvridningen i samhället.

En metod som föreslås av existerande ramverk för att undvika framskridande algoritmisk snedvridning är att låta bli att använda sig av riskfyllda entiteter, eller entiteter som korrelerar starkt med dessa. Det verkar dock existera en konflikt mellan hög träffsäkerhet och få inkluderade parametrar då potentiellt viktig information utelämnas med utelämnade entiteter. Återkopplingen från Cybercoms maskininlärningsexperter visade tydligt på att detta gick emot sättet de föredrar att arbeta på och att de inte trodde på idén. Det finns dock fall där fler parametrar inte har lett till sämre träffsäkerhet, som till exempel i fallet COMPAS (Dressel & Farid, 2018).

Just COMPAS-fallet är kritiskt eftersom det påverkar människors liv i så pass stor utsträckning. Parametrar som korrelerade starkt med människors etniska tillhörighet användes i modellen, vilken lärde sig och tillämpade samhällets snedvridning gentemot mörkhyade i det amerikanska rättssystemet. Det är enligt den amerikanska konstitutionen inte tillåtet att döma människor med grund i etnisk tillhörighet, vilket algoritmen som domarna baseras på uppenbarligen gör. Frågan är då om en sådan algoritm är önskvärd i dessa typer av fall då den troligtvis kommer att basera sin bedömning på faktorer som inte är direkt kopplade till brottet som har begåtts. Även i fallet med universella chattbotar, som Microsofts chattbot Tay, har tillämpningen av AI ifrågasatts. Nyfikenheten kring AI och dess möjligheter är så pass stor att många glömmer att AI, precis som vi människor, är begränsad till den världsbild den ges. Av denna anledning kan det, likt O'Hara m.fl. (2018) föreslår, vara klokt att i detta skede begränsa sig till att använda flera specialiserade AI-lösningar istället för att försöka uppnå en universell sådan.

6.2 Metoder för att undvika algoritmisk snedvridning

I takt med att fler fall av algoritmisk snedvridning har uppmärksammats har ramverk och riktlinjer för hållbar/pålitlig AI utvecklats för att adressera problemet. Då riktlinjerna är riktade mot hållbar/pålitlig AI i allmänhet inkluderas där även andra aspekter av hållbarhet och pålitlighet som inte i direkt mening berör algoritmisk snedvridning. Algoritmisk snedvridning betraktas enbart vara en del i begreppet hållbar/pålitlig AI. EU:s rekommendation att införa "X-by-design"-metoder är ett

exempel på en sådan allmän rekommendation, vilket Cybercom också har gjort via Sustainability by Design som denna studie är del av.

EU:s och AI Sustainability Centers bidrag till existerande ramverk är nästan snarlika varandra, vilket inte är oväntat då bägge organisationerna är större samarbeten mellan olika typer av aktörer. Bägge organisationernas riktlinjer är förhållandevis breda och innehåller begrepp som transparens, ansvarskrävande, förklaringsmetoder, integritet och datastyrning. Transparens har implementerats i ramverket genom att hela tiden ändra *Fairness*-definitionen om modellen inte är tillräckligt välfungerande för alla subgrupper. På så vis infinner sig en transparens kring vilka algoritmen är optimerad och fungerar bäst för i enlighet med EU:s rekommendationer. Ansvarskrävande är inte direkt kopplat till att minimera risken för algoritmisk snedvridning, utan snarare för att minimera konsekvenserna av det och har därför inte implementerats i ramverket. Förklaringsmetoder anses vara för teknik-specifikt för att implementeras i ramverket och dessutom inte applicerbart på alla typer av AI. Integritet och datastyrning implementeras genom undersökning av träningsdata. Vikten av att använda sig av diversifierade träningsdata nämner även Facebook och Google i sina ramverk, vilket har lagts stor vikt vid i denna studies ramverk.

Facebooks och Googles ramverk är något mer konkreta än EU:s och AI Sustainability Centers. Liksom EU nämner dock Facebook både externa granskningsmekanismer, teknisk robusthet och säkerhet. Det förstnämnda implementeras i ramverket med *Fairness Implementation Specialists* som granskar processen i förhållande till *Fairness*. Det sistnämnda implementeras genom validering av modellens robusthet för alla subgrupper enligt även Googles rekommendationer och stresstestning av eventuella användarstyrda indata. Ramverket förespråkar även, enligt Googles rekommendationer, att modellens prestation testas löpande då ett systems funktioner kontinuerligt ändras och tas bort vilket kan leda till nya förutsättningar för målgruppens olika subgrupper.

Mångfald, icke-diskriminering och *Fairness* (eller rättvisa) är begrepp som alla undersökta ramverk understryker vikten av och som genomsyrar hela studiens ramverk genom den kontinuerligt återkommande *Fairness*-definitionen och specialistgruppen *Fairness Implementation Specialists*. Eftersom alla undersökta ramverk nämner detta begrepp anses det vara en av grundpelarna för att minimera risken för algoritmisk snedvridning.

6.3 AI-utveckling hos Cybercom och deras kunder

Målsättningen för ramverk version 2 var att skapa ett ramverk som var applicerbart på både små och stora projekt på Cybercom. Efter konstruktionen av version 1 av ramverk genomfördes en undersökning i form av en workshop samt en intervjustudie tillsammans med maskininlärningsexperter på Cybercom och i enlighet med intervjuresultaten reviderades ramverket till version 2. I intervjustudien diskuterades bland annat relevansen och kompabiliteten av ett ramverk i utvecklingsprojekt på Cybercom. Maskininlärningsexperterna ansåg det inte vara lämpligt att inkludera ett

ramverk i en PoC då de vanligtvis tilldelas mindre resurser än ett fullskaligt projekt och utförs under en kortare tidsperiod. Samtliga maskininlärningsexperter var dock enade om att behovet av ett ramverk är stort för framtida projekt, då det i dagsläget inte hör till vanligheten att etiska aspekter och frågeställningar inkluderas i AI-utveckling

Då det i intervjustudien framkom att det finns ett visst motstånd mot att forcera mångfald i projektteam ersattes detta steg i ramverk version 1 med införandet av en dedikerad mångfaldsgrupp kallad *Fairness Implementation Specialists*. Denna grupp innehåller kompetenser från olika specialistområden som samhällskunskap, beteendevetenskap och humaniora. Om de efterfrågade kompetenser saknas internt på Cybercom kan dessa hämtas från externa aktörer alternativt direkt från kund. För att mångfald och medvetenheten kring snedvridning ska genomsyra hela projektet infördes i ramverk version 2 även en workshop där resultaten från *Fairness Implementation Specialists*' undersökningar diskuteras vidare.

Cybercom är ett av de många bolag som använder sig av open source-lösningar i sina algoritmer. Utifrån den förutsättningen och med anledning av det inte alltid går att undersöka träningsdata lades en extra fråga angående detta in i ramverk version 2. Intervjustudien visade även på att mycket av valideringen och testningen i ramverk version 1 är för tidskrävande och kostsam för mindre bolag. För att minimera kostnaderna men samtidigt bibehålla ramverkets robusthet ändrades delar av dessa steg i ramverk version 2. Bland annat inkluderas validering av alla subgrupper enbart då träningsdatan inte går att undersöka eller som alternativ till att omdefiniera *Fairness* i de fall kvaliteten och tillgången på träningsdata inte är tillräcklig och det inte går att syntetisera träningsdata.

6.4 Förutsättningar för automatisering

För att undersöka vilka förutsättningar för automatisering och AI-implementation som finns hos Cybercoms kunder har två kunder inkluderats i studien. Syftet med dessa undersökningar har varit att få en bättre bild av organisationernas förutsättningar och inställning, dels till ett ramverk men även till AI och automatisering i stort. Det ställs allt hårdare krav på att svenska myndigheters arbete och organisationer ska effektivisera och digitaliseras. Som ett led i Sveriges strategi för digitalisering har Näringsdepartementet tagit fram rekommendationer för offentlig sektors arbete med digitalisering såväl som AI. Enligt dessa riktlinjer bör AI ses som ett hjälpmedel som i största möjliga mån bör utnyttjas i såväl privat som offentlig sektor. Det finns således mycket som talar för att företag och myndigheter inom en snar framtid kommer att börja implementera mer och mer AI och automatisering deras organisationer.

Vad gäller inställningen till AI och intresset för ett ramverk hos Cybercoms kunder skiljer sig delvis åsikterna. Presentationen på energibolaget visade å ena sidan ett stort behov och intresse för såväl AI som för ett ramverk. Intervjustudien på Sida visade å andra sidan en viss skepticism vad gäller förutsättningar för automatiseringar på kort sikt. Där anses utvecklingen hindras av komplexa arbetsuppgifter och en digitalt

omogen organisation. På lång sikt är optimismen dock något större och det finns stora förhoppningar om att AI och automatisering kan leda till både högre transparens och ökad objektivitet gentemot partners och anställda. Som fallstudierna visat är fallgroparna många och för att säkerställa objektivitet i AI är det avgörande att tillämpa någon typ av ramverk i utvecklingsfasen såväl som vid implementeringen av AI:n i samhället. Det finns således goda skäl att tro att båda kunderna i framtiden kommer att efterfråga fler AI-projekt från Cybercom vilket i sin tur skulle leda till ett ökat behov av ett ramverk.

Både energibolaget och Sida är i uppstartsfasen i sin användning av AI och mycket talar för att projekten den närmaste tiden kommer att vara PoC:ar snarare än fullskaliga projekt. Som maskininlärningsexperterna konstaterade är det inte säkert att ett ramverk är lämpligt att använda i en PoC då projekten oftast är kortare och tilldelas mindre resurserna än stora projekt. Å andra sidan visar fallstudierna att det är värdefullt att upptäcka snedvridningar redan i ett tidigt skede och det går således att argumentera för att ett ramverk bör appliceras redan på PoC:ar. Detta bör framförallt betraktas på projekt inom offentlig sektor där det enligt grundlag finns krav på både objektivitet och likabehandling gentemot medborgare. Att i stort sett all AI-utveckling på Cybercom sker med hjälp av open source-modeller är ytterligare ett argument för att redan i en PoC vara uppmärksam på eventuella snedvridningar. Detta då flera av fallstudierna visat på existerande snedvridningar i befintlig AI-teknik, som i verktyg för bildigenkänning och textanalys.

Ramverket innebär omfattande undersökningar, kontroller och tester vilket automatiskt resulterar i större kostnader för projekten. Vid konstruktionen av ramverk version 1 var målsättningen enkom att skapa ett solitt ramverk och inga ekonomiska begränsningar togs i beaktning. Med solitt menas att risken för algoritmisk snedvridning minimeras i så stor utsträckning som möjligt. Detta har har säkerställts genom att ramverket är en sammanställning av befintliga ramverk samt erfarenheter dragna från fallstudier av AI-projekt där snedvridningar har uppstått. Vid framställningen av ramverk version 2 var målsättningen att skapa ett solitt ramverk applicerbart på både mindre och större projekt på Cybercom. För att göra det applicerbart på Cybercom modifierades vissa steg för att bland annat bli mindre kostsamt och mer anpassat för en mellanstor organisation.

I projekt kan det tänkas finnas en konflikt mellan kravet på kostnad och kravet på soliditet. För privata kunder kan exempelvis kravet på vinst och minimering av kostnader vara större än kravet på att produkten ska vara helt solid. Detta skulle kunna innebära att kunder, som konsekvens av en kostnadsbesparing, väljer att bortse från delar av ramverket. För att underlätta denna prioritering skulle en anpassning gentemot privat sektor och en potentiell utveckling av ramverket tas fram. Detta skulle exempelvis kunna vara en prioriterad lista över vilka steg som anses vara viktigast, och således inte bör prioriteras bort, och vilka steg som är mindre avgörande för ett säkert slutresultat. Denna skulle även kunna anpassas utefter en riskanalys. Vad gäller offentlig sektor anses ramverket vara en förutsättning för användning och

implementering av AI och trots krav på kostnadseffektivitet bör projekt genomföras i enlighet med det fullständiga ramverket. Detta för att kunna utnyttja ramverkets soliditet och säkerställa att projektet inte bryter mot de lagar som gäller för offentliga aktörer.

6.5 Metodologiskt angreppssätt

I denna studie har analyser gjorts av fyra existerande ramverk, sex kategorier av fallstudier och intervjuer med tretton personer viktiga för ramverkets utformning. Kvalitativa studier som denna kan inte sällan utvecklas genom fler observationer på ett bredare spann. Det finns många möjliga vidareutvecklingar på ramverket som har utformats. Eftersom en iterativ metodologi har tillämpats kan fler iterationer med fördel genomföras med andra infallsvinklar. Två exempel på intressanta aktörer att intervjua är AI-forskare och representanter från AI Sustainability Center. Likväl kan iterationer göras på fler av Cybercoms kunder för att uppnå ett mer generellt ramverk. Detta ramverk har utformats med iterationer från ett statligt ägt företag och en myndighet. En lämplig kompletterande iteration skulle därför kunna genomföras på ett privatägt företag.

Även fallstudierna kan utökas till att inkludera ett större spann. Undersökning via sökmotorer kan vara snedvriden i sig, vilket kan resultera i att urvalet av fall blir snedvridet. De mest kända fallen är inte heller nödvändigtvis de mest relevanta. En analys skulle därför kunna göras på vilka specifika fall som är mest relevanta för svensk AI-utveckling. Utvecklingen sker dock så pass snabbt att det är relevant att lära sig av misstagen som IT-jättarna gör.

Gällande ramverken anses urvalet vara tillräckligt heltäckande då aktörer från både privat och offentlig sektor har inkluderats. Fler ramverk kan dock givetvis undersökas för ytterligare förfining av ramverket. Då EU:s riktlinjer för pålitlig AI gavs ut 8 april i år var tiden knapp för att läsa rapporten i sin helhet. I ett första stadie lades därför fokus på de övergripande riktlinjer som presenteras på Europeiska kommissionens hemsida. Efter att ramverk version 1 hade formats uppmärksammades en kompletterande utvärderingslista för de olika riktlinjerna - bland annat för algoritmisk snedvridning. Denna lista har inte tagits hänsyn till i ramverkets utformning och skulle troligtvis ha påverkat vissa delar. I senare iterationer av ramverket föreslås därför en utvärdering i enlighet med punkterna presenterade i AI HLEGs rapport om pålitlig AI. Utvärderingslistan skulle även påverka sammanställningen av gemensamma drag hos befintliga ramverk i tabell 3.

Studien har använt sig av mestadels kvalitativa data - men också kvantitativa data - från semistrukturella intervjuer. Den kvantitativa delen utgjordes av ett frågeformulär som användes vid intervjutillfället. Denna del skulle kunna kvantifieras ytterligare genom utskick av enkäter till fler aktörer för att uppnå statistiskt signifikanta resultat istället för indikationer. Det ligger dock en svårighet i komplexiteten av de hanterade ämnena. Många handläggare uttryckte en osäkerhet kring ämnet AI och många avböjde intervju

på grund av denna. Detta är inget konstigt eftersom det inte finns någon generellt accepterad definition av vad AI är och begreppet nyligen har börjat användas i så stor utsträckning som det gör idag. Bristen på en definition av AI och bristen på kunskap om AI skulle skapa problem i ett utskick av enkäter - både till följd av olika uppfattningar och bristande svarsfrekvens. Telefonintervjuer skulle kunna lösa det förstnämnda men risken kvarstår att de anställda skulle avböja på grund av obekvämhet. Detsamma gäller begrepp som digitalisering, som kan sammanblandas med engelskans digitization, och automatisering, som olika personer har olika uppfattningar om vad det innebär.

De flesta existerande ramverk ger generella riktlinjer och rekommendationer snarare än konkreta tillvägagångssätt. De berör dessutom de större begreppen hållbar eller pålitlig AI. Inget av de ramverk som har anträffats har presenterat en konkret metod för att minimera risken för algoritmisk snedvridning. Ytterligare ett generellt ramverk hade inte bidragit märkbart till implementationen av hållbar AI men det går att ifrågasätta om ett flödesschema är för specifikt för att tillämpas på en stor bredd olika projekt. Det är möjligt att en checklista utan specifik följdordning hade varit ett lämpligt alternativ men risken finns att företag enbart skulle tillämpa de punkter de anser vara relevanta - vilket inte skulle lösa problemet. Däremot kan en mer detaljerad checklista med rekommendationer vara önskvärd som komplement till flödesschemat. På så vis skulle aktören få hjälp på vägen med att exempelvis formulera *Fairness*. Ett annat betydelsefullt komplement skulle vara att utforma en riskanalysmall för att kunna avgöra hur stor risken är att algoritmisk snedvridning uppstår i specifika projekt. Då ramverket är både omfattande och tidskrävande skulle företagen, med grund i riskanalysen, i vissa fall kunna hoppa över steg i processen eller genomföra hela processen med vetskapen att det är motiverat ur risksynpunkt. Ytterligare ett förslag på vidareutveckling av ramverket är att, likt AI Sustainability Center, applicera ramverket på ett faktiskt projekt och utvärdera dess utfall. Alla dessa komplement innebär omfattande arbete och är möjliga ämnen för framtida studier.

6.6 Verkshöjd

I världen råder en stor nyfikenhet kring och en vilja att lära sig mer om AI och dess potential. Det har visat sig att nästan hälften av alla AI-bolag i Europa egentligen inte ägnar sig åt AI på ett sätt som kan anses tillhöra deras huvudaffärsområde. Denna nyfikenhet kan anses vara inspirerande men kan också anses bidra till att genvägar tas och att diskussioner om etik och hållbarhet åsidosätts. Detta har i flera fall visat sig ge effekter som kan anses vara diskriminerande enligt både svensk och amerikansk grundlag.

Denna studie har utformat ett ramverk som, med hjälp av befintliga ramverk för hållbar och pålitlig AI och tidigare fall av algoritmisk snedvridning, minimerar risken för att algoritmisk snedvridning uppstår i AI-projekt. Ramverket är anpassat efter Cybercom och dess kunder för att det ska användas i praktiken och inte enbart se snyggt ut på pappret. Ramverket är en del i konceptet Sustainability by Design som ämnar att

katalysera bra istället för dåliga effekter av digitalisering. “X-by-design”-metoder är rekommenderat av EU:s riktlinjer för att uppnå pålitlig AI. Ramverket kan förfinas i fler iterationer för att anpassas efter berörda organisationer och kan även kostnadseffektiviseras i förhållande till en riskanalys. Det ska således inte betraktas som konstant eller färdigt, utan snarare som ett ständigt föränderligt ramverk via nya iterationer och revideringar. Den iterativa processen är välkänd för att fungera väl i en snabbt föränderlig teknikutveckling och därför är det av stor vikt att även ett ramverk för teknikutveckling snabbt kan anpassas efter omvärldens status. Nya tillämpningsområden och krav kan snabbt komma att förändra arbetssätt och risker för att algoritmisk snedvridning uppstår.

Ingen av de befintliga studerade ramverken har oss veterligen bidragit med ett konkret tillvägagångssätt för att minimera risken för algoritmisk snedvridning utan fokuserar mer på övergripande rekommendationer för hållbar eller pålitlig AI. Denna studie har inte enbart bidragit med en konkret metod för att minimera risken att algoritmisk snedvridning uppstår - den har även skapat en medvetenhet kring ämnet på Cybercom och hos Cybercoms kunder. Denna medvetenhet i sig kan anses vara en god start för att minimera risken för att algoritmisk snedvridning uppstår i framtida AI-projekt hos Cybercom.

7. Slutsatser

Denna studie visar på att algoritmisk snedvridning kan uppstå i allt från chattbotar till riskbedömningssystem och autofillfunktioner i smarta telefoner. Det är ett fenomen som inte ens de största IT-bolagen i världen har lyckats undvika och trots att dess förekomst kan leda till fatala konsekvenser för både individer och för samhället är det ett relativt outforskat område. I maj 2018 släppte Facebook sin Fairness flow-metod, i juni 2018 släppte Google sina riktlinjer för rättvis AI, i september 2018 startades AI Sustainability Center och först i april 2019 kom EU med riktlinjer för pålitlig och hållbar AI. Samtliga ramverk och rekommendationer har alltså tagits fram under det senaste året vilket visar på att ämnet fram tills nyligen har förekommit i det dolda.

Den kvalitativa fallstudien och studien av EU:s, AI Sustainability Centers, Facebooks och Googles ramverk för hållbar/pålitlig AI visar på sju huvudsakliga faktorer som bidrar till algoritmisk snedvridning. Dessa är: underliggande snedvridningar i samhället; problem med målgruppens användande av tekniken; bildigenkänning; obalanserade träningsdata; icke-transparent träningsdata; användning av riskfyllda entiteter (parametrar så som kön, etnicitet, religion osv) och användarstyrda indata. Denna studie har även identifierat ett antal metoder för att undvika algoritmisk snedvridning: granskningsmekanismer, teknisk robusthet och säkerhet, integritet och datastyrning, transparens, tydlig definition av Fairness eller rättvisa, förklaringsmetoder, diversifierade team, diversifierade träningsdata, korrekt användarstyrda indata, eliminering av riskfyllda parametrar, utfallsvalidering för subgrupper samt löpande testning.

Ingen av de existerande ramverken bidrar dock med några konkreta tillvägagångssätt och det har således funnits ett glapp mellan de stora bolagens och organisationernas riktlinjer och de små bolagens resurser. Syftet med detta ramverk har varit att minska glappet genom att ta fram en metod som tar hänsyn till befintliga riktlinjer men samtidigt är applicerbart på mindre företag och projekt. Ramverkets tillämpbarhet analyserades på två av Cybercoms kunder. Både inom privat och offentlig sektor förutspås detta ramverk ha god potential att tillämpas, då intervjuresultaten tyder på en ökad efterfrågan för denna typ av metoder. Både privat och offentlig sektor utstrålar dessutom en vilja att utveckla AI inom en snar framtid. Förutsättningen för automatisering och implementering av AI ser dock väldigt olika ut mellan olika aktörer. I samband med vidare iterationer skulle ramverket därför kunna modifieras för offentlig och privat sektor för att på så vis anpassas bättre till olika kunders förutsättningar och krav. Utöver detta bör även ramverkets soliditet och användarbarhet testas genom att tillämpas på ett framtida AI-projekt.

Referenser

Tryckta källor

AI HLEG (High-Level Expert Group on Artificial Intelligence) (2019), Ethics Guidelines for Trustworthy AI, Europeiska kommissionen: Bryssel. Tillgänglig online: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (2019-04-15)

AI Sustainability center (2019), AI Sustainability Center - A multidisciplinary center for responsible and purpose-driven technology, based on Nordic values, AI Sustainability Center: Stockholm

Amnéus, I., Dryler, H., Gillström, P., Helldahl, P., Inkinen, M., Severin, J. & Viberg, A. (2016), Kvinnor och män i högskolan, 2016:16, Universitetskanslersämbetet: Stockholm. Tillgänglig online: https://www.uka.se/download/18.5d85793915901d205f9a852/1487841854015/rapport-2016-10-14-kvinnor-och-man-i-hogskolan.pdf (2019-04-26)

Ashok, J., Suppiah, S., VijiPriya, J. (2016), "A Review on Significance of Sub Fields in Artificial Intelligence", International Journal of Latest Trends in Engineering and Technology (IJLTET), vol. 6, nr. 3, ss. 542–548. Tillgänglig online: https://www.ijltet.org/journal/86.pdf (2019-05-02)

Barbee, D. (2013), Agile Practices for Waterfall Projects - Shifting Processes for Competitive Advantage, J. Ross Publishing. Tillgänglig online: https://app.knovel.com/hotlink/toc/id:kpAPWPSPC1/agile-practices-waterfall/agile-practices-waterfall (2019-05-14)

Bauer, Z., Grosz, B., Etchemendy, J., Lyons, T., Niebles, C.J., Manyika, J., Clark, J., Brynjolfsson, E., Perrault, R. & Shoham, Y. (2018), The AI Index 2018 Annual Report. AI Index Steering Committee, Human-Centered AI Initiative, Stanford University: Stanford. Tillgänglig online: http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf (2019-04-25)

Bengtsson, H. (2012), Offentlig förvaltning: att arbeta i demokratins tjänst. Gleerup: Malmö.

Berggren, N., Jordahl, H., Poutvaara, P. (2016), "The right look: Conservative politicians look better and voters reward it", Journal of Public Economics, vol. 146, ss. 79–86. Tillgänglig online: https://www-sciencedirect-com.ezproxy.its.uu.se/science/article/pii/S0047272716302201 (2019-02-14)

Blaufus, K., Braune, M., Hundsdoerfer, J. & Jakob, M. (2015), "Self-serving bias and tax morale", Economics Letters, vol. 131, ss. 91-93. Tillgänglig online: https://www-sciencedirect-com.ezproxy.its.uu.se/science/article/pii/S0165176515001421 (2019-02-18)

Boden, M.A. (2016), Ai : Its Nature and Future, Oxford University Press: Oxford. Tillgänglig online: https://ebookcentral.proquest.com/lib/uu/detail.action?docID=4545415 (2019-05-02)

Bryman, J. (2011), Samhällsvetenskapliga metoder. Liver: Malmö.

Buolamwini, J., Gebru, T. (2018), "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", Proceedings of Machine Learning Research, vol. 81, ss. 77-91. Tillgänglig online: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf (2019-04-09)

Callan, M.J., Gheorghiu, A.I., Skylark, W.J. (2017), "Facial appearance affects science communication", Proceedings of the National Academy of Sciences of the United States of America, vol. 114, nr. 23, ss. 5970–5975. Tillgänglig online: https://www.pnas.org/content/114/23/5970 (2019-02-15)

Chang, K., Ordóñez, V., Yatskar, M., Wang, T. & Zhao, J. (2017), Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. Tillgänglig online: https://arxiv.org/pdf/1707.09457.pdf (2019-03-15)

Clark, C., Gardner, M., Iyyer, M., Lee, K., Neumann, M., Peters, M.E. & Zettlemoyer, L. (2018), "Deep contextualized word representations", Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language, Technologies, vol. 1. Tillgänglig online: https://arxiv.org/pdf/1802.05365.pdf (2019-04-23)

Colman, A.M. (2015), A Dictionary of Psychology. Oxford University Press: Oxford. Tillgänglig online: https://global.oup.com/ukhe/product/a-dictionary-of-psychology-9780199657681?cc=se&lang=en (2019-01-28)

Crevier, Daniel (1993), AI: The Tumultuous Search for Artificial Intelligence. Basic Books: New York.

Cybercom, (2019a), Full version: Digital Sustainability. Tillgänglig online: https://www.cybercom.com/what-we-do/digitalisation/advisory/digital-sustainability/ (2019-05-21)

Datta, A., Tschantz, MC., Datta, A., (2015), "Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination", Proceedings on Privacy Enhancing Technologies 2015, ss. 92-105. Tillgänglig online: http://www.andrew.cmu.edu/user/danupam/dtd-pets15.pdf (2019-03-25)

Delgado-Rodríguez, M. & Llorca, J. (2004), Bias. "Journal of Epidemiology and Community Health (1979-)", vol. 58, nr. 8, ss. 635–641. Tillgänglig online: https://www-jstor-org.ezproxy.its.uu.se/stable/pdf/25570444.pdf?refreqid=excelsior%3A9d1486240c29474b36bc834a8ecc1691 (2019-03-13)

Dennis, A., Wixom B. H. & Roth, R. M. (2006), System analysis design. United States of America: Wiley & Sons.

Dressel, J. & Farid, H. (2018). "The accuracy, fairness, and limits of predicting recidivism", Science Advances, vol. 4, nr. 1. Tillgänglig online: http://advances.sciencemag.org/content/4/1/eaao5580 (2018-03-05)

Friedman, B. & Nissenbaum, H. (1996), "Bias in Computer Systems", ACM Transactions on Information Systems (TOIS), vol. 14, nr. 3, ss. 330–347. Tillgänglig online: https://dl-acm-org.ezproxy.its.uu.se/citation.cfm?id=230561 (2019-02-20)

Google (2019), Google Diversity: Annual Report 2019. Tillgänglig online: https://static.googleusercontent.com/media/diversity.google/sv//static/pdf/Google_diversity_annual_report_2019.pdf (2019-04-25)

Guinote, A. & Weick, M. (2010), "How long will it take? Power biases time predictions", Journal of Experimental Social Psychology, vol. 46, nr. 4, ss. 595–604. Tillgänglig online: https://www.researchgate.net/publication/228905402_How_long_will_it_take_Power_biases_time_predictions (2019-05-21)

Gulliksen, J. & Göransson, B. (2002), Användarcentrerad systemdesign (3:e upplagan). Studentlitteratur AB: Lund.

Han, J., Kamber, M. & Pei, J. (2011), Data Mining: Concepts and Techniques. Elsevier Science & Technology: Saint Louis.

Heider, F. (1958), The psychology of interpersonal relations. Wiley: New York.

Hunt, V., Layton, D. & Prince, S. (2015), Diversity Matters, McKinsey&Company. Tillgänglig online: https://www.mckinsey.com/~/media/mckinsey/business%20functions/organization/our%20insights/why%20diversity%20matters/diversity%20matters.ashx (2019-04-25)

Kahneman, D. (2013), Tänka snabbt och långsamt. Volante: Stockholm.

Kahneman, D. & Lovallo, D. (1993), "Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking", Management Science, vol. 39, nr. 1, ss. 17–31. Tillgänglig online: https://www-jstor-org.ezproxy.its.uu.se/stable/2661517?pq-origsite=summon&seq=1#metadata_info_tab_contents (2019-05-21)

Kahneman, D. & Tversky, A. (1981), "The Framing of Decisions and the Psychology of Choice", Science, vol. 211, nr. 4481, ss. 453–458. Tillgänglig online: https://www.uzh.ch/cmsssl/suz/dam/jcr:ffffffff-fad3-547b-ffff-ffffe54d58af/10.18_kahneman_tversky_81.pdf (2019-02-14)

Kizer, M. J. (2017), "Arrested by Skin Color: Evidence from Siblings and a Nationally Representative Sample", Socius, vol. 3, ss. 1-12. Tillgänglig online: https://journals.sagepub.com/doi/full/10.1177/2378023117737922?utm_source=summon&utm_medium=discovery-provider (2019-05-21)

Larsson, S., Anneroth, M., Felländer, A., Felländer-Tsai, L., Heintz, F., Cedering Ångström, R. (2018), Hållbar AI, AI Sustainability Center. Tillgänglig online: http://www.aisustainability.org/wp-content/uploads/2019/03/Hallbar_AI.pdf (2019-04-23)

Lidén, M. (2018), Confirmation Bias in Criminal Cases, Diss., Uppsala universitet. tillgänglig online: http://uu.diva-portal.org/smash/get/diva2:1237959/FULLTEXT01.pdf (2019-02-12)

Markoff, J. (2011), "Computer Wins on 'Jeopardy!': Trivial, It's Not", The New York Times, 17 februari. Tillgänglig online: https://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html# (2019-05-21)

McCorduck, P. (2004), Machines Who Think (2nd ed.). Natick: A. K. Peters.

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 (2019-03-18)

Moore, D.A. & Schatz D. (2017), "The three faces of overconfidence", Social and Personality Psychology Compass, vol. 11, nr. 8. Tillgänglig online: https://doi-org.ezproxy.its.uu.se/10.1111/spc3.12331 (2019-02-13)

New York State Division of Criminal Justice (2012), New York State COMPAS-Probation Risk and Need Assessment Study: Examining the Recidivism Scale's Effectiveness and Predictive Accuracy, New York State: New York, tillgänglig online: http://www.northpointeinc.com/downloads/research/DCJS_OPCA_COMPAS_Probation_Validity.pdf (2019-02-28)

Northpointe (2011), Risk Assessment. Tillgänglig online: https://assets.documentcloud.org/documents/2702103/Sample-Risk-Assessment-COMPAS-CORE.pdf (2019-02-27)

Näringsdepartementet (2018), Nationell inriktning för artificiell intelligens, N2018.14, Regeringskansliet: Stockholm. Tillgänglig online:

https://www.regeringen.se/49a828/contentassets/844d30fb0d594d1b9d96e2f5d57ed14b/2018ai_webb.pdf (2019-01-28)

O'Hara, K. P., Schlesinger, A. & Taylor, A. S. (2018), "Let's Talk About Race: Identity, Chatbots, and AI", Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, nr. 315. Tillgänglig online: https://static1.squarespace.com/static/5a8b405a18b27d5478196dca/t/5a8b690d24a694d7072d25a1/1519085853799/chi18-schlesinger-LetsTalkAboutRace.pdf (2019-03-19)

Patel, R. & Davidson, B. (2011), Forskningsmetodikens grunder: Att planera, genomföra och rapportera en undersökning. Studentlitteratur AB: Lund.

Rasmussen, J. (1983), "Skills, rules, knowledge: signals, signs and symbols and other distinctions in human performance models", IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-13, nr. 3, ss. 257-266. Tillgänglig online: https://ieeexplore-ieee-org.ezproxy.its.uu.se/stamp/stamp.jsp?tp=&arnumber=6313160 (2019-05-21)

Redelmeier, D., Ruff, C.C., Saposnik, G., Tobler, P.N. (2016), "Cognitive biases associated with medical decisions: a systematic review", BMC Medical Informatics and Decision Making, vol. 16, nr. 138. Tillgänglig online: https://bmcmedinformdecismak.biomedcentral.com/track/pdf/10.1186/s12911-016-0377-1 (2019-02-13)

Royce, W. W. (1970), Managing the development of large software systems, Proceedings of the 9th international conference on software engineering, ss. 328-338. Tillgänglig online: https://leadinganswers.typepad.com/leading_answers/files/original_waterfall_paper_winston_royce.pdf (2019-05-09)

SOU 1998:137. Slutrapport av Tunnelkommissionen. Miljö i grund och botten – erfarenheter från Hallandsåsen. Tillgänglig online: https://data.riksdagen.se/fil/08620A70-92A3-413E-A42D-B3F92224D50D (2019-05-21)

Stanovich, K., West, R. (2000), Individual Differences in Reasoning: Implications for the Rationality Debate, Behavioral and brain sciences.

Tegmark, M. (2017), Liv 3.0 : att vara människa i den artificiella intelligensens tid. Volante: Stockholm.

Turing, A.M. (1950), "Computing Machinery and Intelligence", Mind, vol. 49, nr. 236, ss. 433-460. Tillgänglig online: https://www.csee.umbc.edu/courses/471/papers/turing.pdf (2019-02-06)

Utrikesdepartementet (2010), En transparensgaranti i biståndet, UD10.050, Regeringskansliet: Stockholm. Tillgänglig online: https://www.regeringen.se/informationsmaterial/2010/05/ud10.050/

(2019-02-07)

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

Wirth, N. (2018), “Hello marketing, what can artificial intelligence help you with?”, International Journal of Market Research, vol. 60, nr. 5, ss. 435-438. Tillgänglig online: https://web-b-ebscohost-com.ezproxy.its.uu.se/ehost/pdfviewer/pdfviewer?vid=1&sid=e7f42cde-e45b-4a37-970a-882f88b8e49b%40sessionmgr101 (2019-05-22)

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

Internetkällor

Alciné, J. (2015) [tweet], Google Photos, y'all fucked up. My friend's not a gorilla. Tillgänglig online: https://twitter.com/jackyalcine/status/615329515909156865 (2019-03-14)

Britannica Academic (2016), Confirmation bias. Tillgänglig online: https://academic-eb-com.ezproxy.its.uu.se/levels/collegiate/article/confirmation-bias/627535 (2019-02-12)

Bureau of Labor Statistics (2002), Household data annual averages: Median weekly earnings of full-time wage and salary workers by detailed occupation and sex. Tillgänglig online: https://www.bls.gov/cps/aa2002/cpsaat39.pdf (2019-05-06)

Bureau of Labor Statistics (2018), Household data annual averages: Median weekly earnings of full-time wage and salary workers by detailed occupation and sex. Tillgänglig online: https://www.bls.gov/cps/cpsaat39.pdf (2019-05-06)

City of Boston (2017), Street Bump. Tillgänglig online: https://www.boston.gov/departments/new-urban-mechanics/street-bump (2019-02-19)

Cybercom (2019b), AWS. Tillgänglig online: https://www.cybercom.com/what-we-do/cloud-services/aws/?gclid=Cj0KCQjw2IrmBRCJARIsAJZDdxCG4xjXgE_2WTkxEUFjJ-g2mnvBJqI5L-u8znM8k86tI_OOy5vqhxwaAv6CEALw_wcB (2019-04-26)

Cybercom (2019c), Om koncernen, Tillgänglig online: https://www.cybercom.com/sv/Om-Cybercom/Om-koncernen/ (2019-02-07)

ELMo (2018), ELMo: Deep contextualized word representations. Tillgänglig online: https://allennlp.org/elmo (2019-04-23)

Europeiska kommissionen (2019), Ethics guidelines for trustworthy AI. Tillgänglig online: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (2019-05-22)

F8 2018 Day 2 Keynote (2018) [video], Facebook. Tillgänglig online: https://www.facebook.com/FacebookforDevelopers/videos/f8-2018-day-2-keynote/10155609688618553/ (2019-04-11)

Facebook (2018), Facebook Diversity Update. Tillgänglig online: https://www.facebook.com/careers/diversity-report (2019-04-25)

Fortune (2017), These Are the Women CEOs Leading Fortune 500 Companies. Tillgänglig online: http://fortune.com/2017/06/07/fortune-500-women-ceos/ (2019-05-06)

Gender Shades (2018), The Gender Shades project evaluates the accuracy of AI powered gender classification products. Tillgänglig online: http://gendershades.org/overview.html (2019-04-05)

Google AI (2018), Responsible AI Practices. Tillgänglig online: https://ai.google/education/responsible-ai-practices?category=fairness (2019-04-11)

Turovsky, B. (2016)[blogginlägg], Ten years of Google translate, Google Blog, 28 April. Tillgänglig online: https://www.blog.google/products/translate/ten-years-of-google-translate/ (2019-05-27)

Hassabis, D. & Silver, D. (2017), AlphaGo: Mastering the ancient game of Go with Machine Learning. Google AI Blog. Tillgänglig online: https://ai.googleblog.com/2016/01/alphago-mastering-ancient-game-of-go.html (2019-05-21)

I-scoop (2016), Digitization, Digitalization and digital transformation: the differences. Tillgänglig online: https://www.i-scoop.eu/digitization-digitalization-digital-transformation-disruption/ (2019-02-20)

IT-ord (2017), Sökresultat för: digitalisering. Tillgänglig online: https://it-ord.idg.se/?s=digitalisering (2019-01-31)

Johnson, M. (2018) [blogginlägg], Providing Gender-Specific Translations in Google Translate, 10 december. Tillgänglig online: https://ai.googleblog.com/2018/12/providing-gender-specific-translations.html (2019-02-28)

Kazemi, D. (2016), wordfilter. Tillgänglig online: https://www.npmjs.com/package/wordfilter (2019-03-19)

Markoff, J. (2011)."On 'Jeopardy!' Watson Win Is All but Trivial".The New York Times, Tillgänglig online: https://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html (2019-02-11)

Mayerhofer, C. (2018), Rasmussen's SRK Model. Tillgänglig online: https://aviatortraining.net/2018/08/02/rasmussens-srk-model/ (2019-05-03)

McKinsey Podcast (2019) [podradio], The ethics of artificial intelligence. Tillgänlig online: https://www.mckinsey.com/featured-insights/artificial-intelligence/the-ethics-of-artificial-intelligence?cid=eml-app (2019-02-19)

Microsoft (2016), Learning from Tay's introduction. Tillgänglig online: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/ (2019-03-14)

Microsoft (2019), Let's talk about Zo. Tillgänglig online: https://www.zo.ai/ (2019-03-19)

Myndigheten för digital förvaltning (2018), Vårt uppdrag, Tillgänglig online: https://www.digg.se/om-oss/var-verksamhet/vart-uppdrag (2019-01-28)

Olson, P. (2018), The Algorithm That Helped Google Translate Become Sexist. Tillgänglig online: https://www.forbes.com/sites/parmyolson/2018/02/15/the-algorithm-that-helped-google-translate-become-sexist/#1fca53097daa (2019-03-08)

Openaid (2019), Om Openaid.se. Tillgänglig online: https://openaid.se/sv/about/ (2019-02-07)

ProPublica (2016a), Machine Bias. Tillgänglig online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (2019-02-26)

Regeringskansliet (2017), Mål för digitaliseringspolitik. Tillgänglig online: https://www.regeringen.se/regeringens-politik/digitaliseringspolitik/ (2019-01-28)

Reuters (2018), Amazon scraps secret AI recruiting tool that showed bias against women. Tillgänglig online: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (2019-02-20)

Sida (2015), Arbetssätt och metoder. Tillgänglig online: https://www.sida.se/Svenska/sa-arbetar-vi/Arbetssatt-och-metod/ (2019-01-28).

The Guardian (2016a), Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. Tillgänglig online: https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter (2019-03-14)

The Guardian (2016b), The racist hijacking of Microsoft's chatbot shows how the internet teems with hate. Tillgänglig online: https://www.theguardian.com/world/2016/mar/29/microsoft-tay-tweets-antisemitic-racism (2019-05-21)

The Guardian (2015), Flickr faces complaints over 'offensive' auto-tagging for photos. Tillgänglig online: https://www.theguardian.com/technology/2015/may/20/flickr-complaints-offensive-auto-tagging-photos (2019-05-21)

The Guardian (2013), 'Racism' of early colour photography explored in art exhibition. Tillgänglig online: https://www.theguardian.com/artanddesign/2013/jan/25/racism-colour-photography-exhibition (2019-05-08)

Trafikverket (2015) [WayBack Machine], Frågor och svar om Hallandsås, Tillgänglig online: https://web.archive.org/web/20151130062530/http://www.trafikverket.se/nara-dig/Skane/projekt-i-skane-lan/Hallandsas/fragor-och-svar-om-hallandsas/ (2019-02-15)

US Department of Defence (2002), DoD News Briefing. Tillgänglig online: https://archive.defense.gov/Transcripts/Transcript.aspx?TranscriptID=2636 (2019-05-02)

Zunger, Y. (2017), Asking the Right Questions About AI. Tillgänglig online: https://medium.com/@yonatanzunger/asking-the-right-questions-about-ai-7ed2d9820c48?fbclid=IwAR2TQCmcUYmTxtWfPCsWeG0o1OtsSzB4Wvzj1iXXTnZ8-qZd_eMAs5qvC3I (2019-05-09)

Wang, J. (2009) [blogginlägg], Racist Camera! No, I did not blink… I'm just Asian!. Tillgänglig online: http://www.jozjozjoz.com/2009/05/13/racist-camera-no-i-did-not-blink-im-just-asian/ (2019-03-15)

Wired Magazine (2018), When it comes to gorillas, Google Photos remains blind. Tillgänglig online: https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/ (2019-03-14)

Intervjuer

Annergren, B. Data scientist Cybercom Group Stockholm. 2019. Intervju den 29 januari & 16 april

Assal, A. AR, VR, MR, ML-expert Cybercom Group Stockholm. 2019. Intervju den 7 mars & 24 april

Bazine, E. Statistiker Sida Stockholm. 2019. Intervju den 27 mars

Berghald, S. Handläggare Sida Stockholm. 2019. Intervju den 19 mars

Cuzner, M. Systemförvaltare Sida Stockholm. 2019. Intervju den 28 mars

Filippakis, L. Full-Stack Data Scientist Cybercom Group Göteborg. 2019. Intervju den 7 mars & 17 april

Hahto, A. Head of data science Cybercom Group Helsingfors. 2019. Intervju den 12 april

Karaoguz, H. Senior data scientist Cybercom Group Stockholm. 2019. Intervju den 29 januari & 16 april

Mwakifwamba, S. Handläggare Sida Dar es Salaam, Tanzania. 2019. Intervju den 20 mars

Rodes, F. Data scientist Cybercom Group Stockholm. 2019. Intervju den 29 januari & 16 april

Söråker, J. Ansvarig på avdelningen för Maskininlärning, Cybercom Stockholm. 2019. Intervju den 27 februari

Qvarnström, M. Statistiker Sida Stockholm. 2019. Intervju den 27 mars

Wigholm, J. IT-chef IT-enheten Sida Stockholm. 2019. Intervju den 20 februari

Appendix A

Intervjumall för semistrukturerade intervjuer tillsammans med maskininlärningsexperter på Cybercom.

Open Source

- Vilken open source som används vanligen?
- Hur är en specifik open source modell vald?
- Vilka begränsningar och risker ser ni med att använda open source ML-modeller?
- Kan man undersöka träningsdatan?
    - Har ni undersökt träningsdatan?
- Tips på hur man hittar snedvridning i träningsdata, t.ex. blå kläder för män och rosa kläder för kvinnor?

Data

- Hur ofta syntetisera ni träningsdatan?
    - Är det en säker metod?
- Hur avgörs vilka subgrupper som behövs tas hänsyn till/finnas representerade?
- Hur ofta vet ni hur träningsdatan är insamlat/ kommer ifrån?
- Finns det entiteter/parametrar som ni inte inkluderar?

Diversifiering av team

- Hur jobbar ni med diversifiering av team?
- Vad ser ni för värde med att ha ett diversifierat team?
- Skulle det vara rimligt att sätta som krav att personer som är med i projektet kommer från olika bakgrunder etc?
    - Är det ens genomförbart i nuläget eller får man vara nöjd med de som man får som är kunniga inom ämnet?

Validering (subgrupper)

- Vilka metriker använder ni för validering av algoritmen?
- Har ni validerat falska positiva och falska negativa värden inom varje subgrupp?

Övrigt

- Hur avgör ni vilka användare slutprodukten ska användas på?
- I vilken sorts projekt är ett ramverk applicerbart?

Teknisk vs etisk snedvridning

- I vilken utsträckning tittar ni på statistisk snedvridning?
- I vilken utsträckning tittar ni på etisk snedvridning?

Appendix B

Intervjumall för semistrukturerade intervjuer tillsammans med statistiker på Sida.

- Beskriv dina arbetsuppgifter
- Vilka verktyg använder du?
- Hur använder ni er av statistik på sida?
- I vilken utsträckning beror beslut på statistik?
- Finns det potential att använda sig mer av statistik än vad ni gör idag?
- Vilken typ av data hanterar du? Exempel.
- Hur validerar du/ni att datan är korrekt?
- Hur kontrollerar ni hur datan har blivit insamlat?
- Gör ni egna data-insamlingar. Hur går dessa till i så fall?
- På en skala 1-10 hur digitaliserat skulle du säga att sida är?
- Vilken potential till digitalisering ser du?
- Hur mycket av materialet finns i digital form?
- Vad tror du skulle kunna automatiseras?
- Vilka möjligheter finns med automatisering?
- Vilka förutsättningar finns för automatisering?
- Vilka risker ser du med en eventuell automatisering av handläggarnas arbetsuppgifter?
- Tror du att det skulle förekomma mer eller mindre snedvridning vid en automatisering av beslut?
- Ser du några risker med AI-PoCen som Cybercom driver?
    - Tror du att det skulle kunna bli fullt automatiserat?
    - Tror du att fler eller färre dokument kommer att plockas upp? (Kvalitetsförbättring eller tids/kostnadsbesparing)
    - Tror du att PoCen kommer att leda till fler liknande projekt?

Appendix C

Intervjumall för semistrukturerade intervjuer tillsammans med handläggare på Sida.

- Beskriv dina arbetsuppgifter
- Hur ser en ansökningsprocess ut?
- Vilka verktyg använder du?
- Vilka aspekter tar du med i en ansökan?
- Finns det information och erfarenheter som inte finns i skrift som ändå tas med i ansökan?
- Vilka beslut kräver subjektiv bedömning?
- Vilka beslut innefattar behandling av kvantitativa data?
- Hur säkerställer ni att informationen ni får är korrekt?
- Hur konsekventa skulle du säga att ni är i era beslut?
- Hur digitaliserade skulle du säga att Sida är?
- Hur digitiserade skulle du säga att Sida är?
- Hur tänkte du när du besvarade frågorna angående digitisering och digitalisering?
- Vad tror du skulle kunna automatiseras?
- Vilka möjligheter finns med automatisering?
- Vilka förutsättningar finns för automatisering?
- Vilka risker ser du med en eventuell automatisering av dina arbetsuppgifter?
- Tror du att det skulle förekomma mer eller mindre snedvridning i en automatisering av beslut?

Appendix D

Intervjumall för semistrukturerade intervjuer tillsammans med chefen för IT-enheten på Sida.

Generellt

- Hur arbetar ni på Sida?
- Hur är organisationen uppdelad?
- Hur ser ansvarsfördelningen ut?

Digitalisering

- Vad har Sida för digitaliseringsmål?
- Sveriges mest modernaste myndighet?
- Vart tror du att kan komma att digitaliseras i framtiden?
- Hur ligger Sida till i jämförelse med andra myndigheter?
- Vilka är nyckelpersoner på Sida som kommer att ta beslut angående digitalisering?
- Vilka brister finns det med det nuvarande systemet?
- Både det specifika exemplet och IT-säkerhet
- Blir ni reviderade av IT-revisorer?
- Hur går det till?
- På vilka områden?

AI

- Hur skulle ni kunna tänkas implementera AI?
- Hur går ett beslut till i ett AI-projekt?
- Finns det några pågående AI-projekt andra än PoCen?
- Tror du att man kan använda sig av helt automatiserade processer inom offentlig sektor eller är det för riskabelt?
- Är det viktigaste att målgruppen täcker alla i Sveriges befolkning eller att den är exakt?
    - Skulle Sida kunna ta sådana risker?
- Vilka skulle du intervjua om du skulle göra ett ramverk för AI-processer?

Appendix E

Frågeformulär till maskininlärningsexperter på Cybercom.

How often are already trained open source models used in ML development at Cybercom according to you?

- ○ 0 - 20 %
- ○ 21 - 40%
- ○ 41 - 60%
- ○ 61 - 80%
- ○ 81 - 100%

How often do you use data augmentation according to you?

- ○ 0 - 20%
- ○ 21 - 40%
- ○ 41 - 60%
- ○ 61 - 80%
- ○ 81 - 100%

What fraction of gender representation do you usually see in Swedish development teams in general?

	0-20%	21-40%	41-60%	61-80%	81-100%
Women	○	○	○	○	○
Men	○	○	○	○	○
Other	○	○	○	○	○

What fraction of age representation do you usually see in Swedish development teams in general?

	0-20%	21-40%	41-60%	61-80%	81-100%
20 - 30	○	○	○	○	○
31 - 40	○	○	○	○	○
41 - 50	○	○	○	○	○
51 and older	○	○	○	○	○

What fraction of nationality representation do you usually see in Swedish development teams in general?

	0-20%	21-40%	41-60%	61-80%	81-100%
Swedish	○	○	○	○	○
Other Europpean	○	○	○	○	○
Non Europpean	○	○	○	○	○

How often do you take statistical bias into consideration in ML projects?

- ○ Never
- ○ Occasionally
- ○ Sometimes
- ○ Often
- ○ Always

How often do you take ethical bias into consideration in ML projects?

- ○ Never
- ○ Occasionally
- ○ Sometimes
- ○ Often
- ○ Always

Appendix F

Deltagare i intervjustudierna.

Intervjuperson	Företag/Organisation	Befattning	Refereras i texten som
Jon Söråker	Cybercom Group Stockholm	Competence team lead	Maskininlärningsexpert
Ahmed Assal	Cybercom Group Stockholm	AR, VR, MR, ML expert	Maskininlärningsexpert
Lef Filippakis	Cybercom Group Göteborg	Full-Stack Data Scientist	Maskininlärningsexpert
Francisco Rodes	Cybercom Group Stockholm	Data Scientist	Maskininlärningsexpert
Hakan Karaoguz	Cybercom Group Stockholm	Senior Data Scientist	Maskininlärningsexpert
Björn Annergren	Cybercom Group Stockholm	Data Scientist	Maskininlärningsexpert
Antti Hahto	Cybercom Group Helsingfors	Head of Data Science	Maskininlärningsexpert
Jenni Wigholm	Sida	Chef IT-enheten Sida	Chef IT-enheten Sida
Melinda Cuzner	Sida	Systemförvaltare	Systemförvaltare
Elies Bazine	Sida	Statistiker	Statistiker
Marcus Qvarnström	Sida	Statistiker	Statistiker
Sofie Berghald	Sida	Handläggare	Handläggare
Stephen Mwakifwamba	Sveriges ambassad Dar es Salaam Tanzania	Handläggare	Handläggare

Appendix G

Frågeformulär till handläggare på Sida.