AI OCH JÄMSTÄLLDHET I STATLIG FÖRVALTNING

JÄMSTÄLLDHETS MYNDIGHETEN

VINNOVA
Sveriges innovationsmyndighet

anch.AI

INNEHÅLL

Foto: Fauxels/Pexels

INLEDNING

Jämställdhetsmyndigheten och företaget anch.AI har genomfört pilotprojektet ”AI i statlig förvaltning”. Projektet har finansierats av myndigheten Vinnova och Skatteverket och Försäkringskassan har varit testbäddar och miljöer där metod och lösningar testats för vidareutveckling.

Syftet har varit att utforska hur införandet av artificiell intelligens (AI) i statlig förvaltning kan bidra till att de jämställdhetspolitiska målen uppfylls samt bidra till ett lärande för att skapa bättre förutsättningar för statliga myndigheter att främja jämställdhet genom användning av AI.

För att generera insikter har anch.AI:s ramverk för hållbar AI som metod använts. Ramverket syftar till att identifiera, mäta och hantera potentiella risker med användning av AI.

I den här skriften sammanfattas de kunskaper och insikter som projektet bidragit till. Skriften innehåller även praktiska metoder för hur risker med ojämställdhet i AI kan identifieras och hur AI med tillämpning av ett jämställdhetsperspektiv kan utvecklas. Projektet har genomförts utifrån ett antagande om att AI både kan riskera att skapa och reproducera ojämställdhet. Användning av AI kan därmed också riskera att leda till diskriminering på grund av kön.

Innehållet vänder sig till funktioner som bidrar till att utveckla AI med ett jämställdhetsperspektiv inom statlig förvaltning, exempelvis verksamhetsutvecklare, IT-utvecklare och jämställdhetssamordnare.

ARTIFICIELL INTELLIGENS OCH JÄMSTÄLLDHET

AI kan förenklat beskrivas som robotar och datorprogram som lär sig av den data de får in. Många myndigheter använder AI för att till exempel hjälpa till att sammanställa beslutsunderlag för handläggare. Två exempel är Försäkringskassan som utvecklar en digital plattform som används för att träna kognitiva modeller för olika verksamhetsbehov samt Skatteverket som använder AI för textanalys av upplysningar i deklarationen.

AI programmeras och tränas för att utföra sin uppgift och lär sig vanligtvis av historiska data. Om denna data bygger på ojämställdhet kan AI riskera att förstärka samma strukturer. Det finns alltså en risk att AI kan återskapa algoritmer som speglar oavsiktliga och stereotypa föreställningar eller väljer ut data och variabler som motverkar jämställdhet. Ett tredje sätt är att ojämställdheten i AI reproduceras genom den mänskliga faktorn, genom de individer som designar och tränar AI- lösningar. AI kan också bidra till ökad jämställdhet. Då gäller det att få AI att upptäcka preferenser eller gamla fördomar och stereotyper i data. AI blir då i stället ett verktyg som skapar jämställdhet.

En nyckel för att identifiera AI:s påverkan är att analysera momenten i införandeprocessen ur ett jämställdhetsperspektiv.

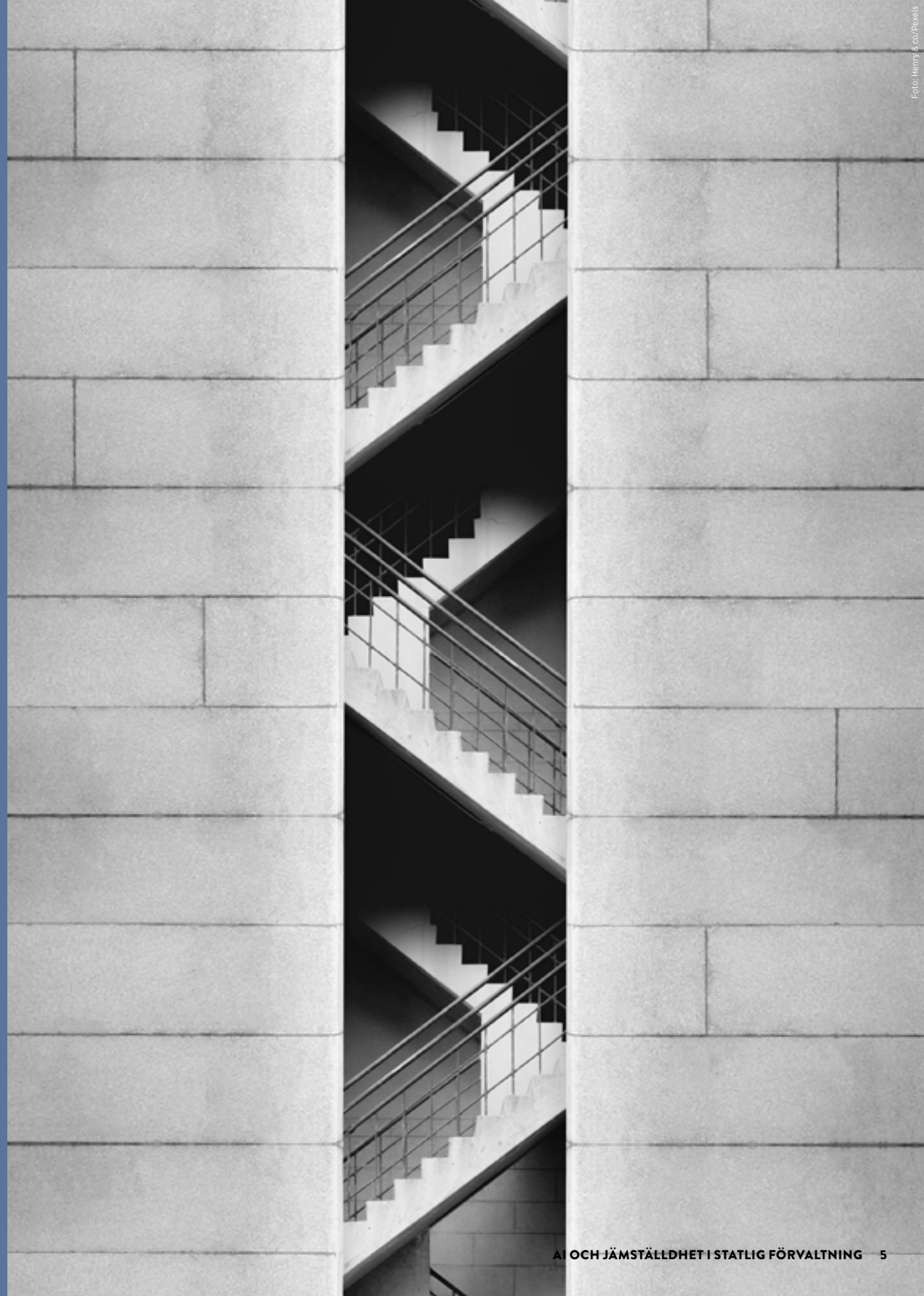


Foto: Henry & co/Pexels

Foto: Shutterstock

RISKER – TRE JÄMSTÄLLDHETS-SPECIFIKA RISKER

Tre riskområden för användning av AI:

- Att användning av AI skapar ojämställdhet.
- Att användningen av AI reproducerar ojämställdhet.
- Användning av AI leder till diskriminering på grund av kön.

Pilotprojektet har haft ett riskbaserat angreppssätt. anch. AI:s metod Ramverk för hållbar AI har tillämpats i syfte att skapa insikter om hur användningen av AI kan innebära risk för ojämställdhet och hur en myndighet kan hantera dessa. Ramverket har anpassats till tre övergripande jämställdhetsspecifika risker:

- Risk 1. Att användning av AI skapar ojämställdhet:

Nya former av strukturella skillnader mellan könen kan uppstå genom att använda AI. Strukturell ojämställdhet kan bland annat riskera att uppstå när användningen av AI och automatisering leder till förändringar på arbetsmarknaden. Ett exempel gäller om särskilda yrkesgrupper påverkas i större utsträckning än andra och om de grupperna är mer kvinno- eller mansdominerade.

Användningen av AI kan även medföra att existerande ojämställdhet tar sig nya uttryck eller får effekt på andra samhällsområden. Ett vanligt tillämpningsområde är att analysera och klassificera text. Det används till exempel för att sortera inkommande mejlkorrespondens, för att definiera och upptäcka hatisk retorik på sociala plattformar, eller för att sortera stora mängder informationsunderlag, som exempelvis domar eller läkarutlåtanden.

Språkbruk kan vara färgat av vilket kön en individ tillhör. Det kan leda till att AI-lösningar systematiskt gynnar eller missgynnar en viss grupp vilket kan få konsekvenser på andra samhällsområden. I de fall där individers könskodade språkbruk ligger till grund för diagnos i mötet med exempelvis en läkare finns risk för att texten blir färgad av föreställningar och stereotyper om kön.

Foto: Scandinav

Underlaget som AI ska tolka kan därmed vara färgat av bias, det vill säga att informationen inte avspeglar faktiska förhållanden, i flera steg eller på olika nivåer. Ett exempel är när ett AI-baserat verktyg skulle upptäcka och nå ut till potentiella jobbkandidater genom att läsa av tillgänglig information på internet. AI-lösningen visade sig systematiskt gynna kandidater som beskrev sig själva med verb som oftare fanns på manliga ingenjörers CV[1].

- **Risk 2. Att användningen av AI reproducerar ojämställdhet**

Om AI tränas på data som av olika anledningar är skev, det vill säga innehåller bias, riskerar det att reproduceras i lösningens utfall. I samhället finns olika former av strukturell ojämställdhet, vilket innebär att det finns en risk att informationen om samhället speglar dessa skillnader mellan kvinnor och män. Vanligtvis tränas AI-lösningar på historiska data, vilket innebär att det finns en risk för att historisk bias reproduceras i utfallen. Hur data väljs ut och analyseras är även källa till att bias introduceras i träningsdata.

Användningen av AI kan medföra att ojämställdhet reproduceras, genom att mänsklig bias reproduceras i enskilda lösningar. Bias kan introduceras av de individer som designar och tränar AI-lösningar. Algoritmer skapas därmed som återspeglar oavsiktliga, stereotypa, föreställningar eller introducerar bias genom hur data väljs ut och analyseras.

- **Risk 3: Användning av AI leder till diskriminering på grund av kön**

Det har under de senaste åren på flera håll uppmärksammats att användningen av AI kan leda till diskriminering. Diskrimineringsombudsmannen har konstaterat att diskriminering som sker genom AI-system omfattas av diskrimineringslagen (2008:567) vars förbud är teknikneutrala[2].

Likaså har andra europeiska likabehandlingsmyndigheter, Europeiska Kommissionen, Ministerrådet och Europeiska unionens byrå för grundläggande rättigheter uppmärksammat att AI kan medföra risk för diskriminering[3].

Användning av AI kan leda till direkt diskriminering, genom att diskriminerande beslut fattas på grund av variabeln kön. Diskriminering kan även ske indirekt, genom att andra variabler fungerar som en Proxy för variabeln kön, det vill säga att de speglar information i en annan variabel. Längd korrelerar exempelvis starkt med kön och kan därför fungera som en Proxy.

Det har också uppmärksammats att användningen av AI har resulterat i diskriminerande utfall inom rekrytering. Ett exempel visar på att fler män än kvinnor rekryterades historiskt till en viss typ av tjänst och när en AI-lösning tränades på information om historiska rekryteringar blev resultatet att manliga kandidater förutspåddes vara mer produktiva i rollen än kvinnor[4].

Diskriminering på grund av kön ingår som en riskkategori i det här projekt på grund av den existerande koppling som finns mellan diskriminering och ojämställdhet. Orsaken till könsdiskriminerande eller ojämställda utfall av en AI-lösning bero på överlappande faktorer som bias i data eller att lösningen inte är tillräckligt tränad alternativt används på fel sätt.

1 Amazon scraps secret AI recruiting toll that showed bias against women, 2018. Jeffry Dastin, Reuters.
2 DO yttrande över Europeiska kommissionens förslag till förordning om harmoniserade regler för artificiell intelligens. Proposal for a Regulation laying down harmonised rules on artificial intelligence (artificial intelligence act) and mending certain union legislative acts (COM (2021)206). Diskrimineringsombudsmannen; Dnr LED 2021/290
3 Se exempelvis: Når algoritmer sagsbehandler: risiko för diskrimination i det offentliges brug av profileringsmodeller. 2021. Institut for menneske rettigheter. Discrimination, artificial intelligence and algorithmic decision making 2018. Prof Frederik Zuiderveen Borgesius, Council of Europe, FRA. #Big data:
4 Amazon scraps secret AI recruiting tool that showed bias against women. 2018. Jeffrey Dastin, Reuters.

Diskriminering på grund av kön ingår som en riskkategori i det här projektet på grund av den existerande koppling som finns mellan diskriminering och ojämställdhet. Orsaken till könsdiskriminerande eller ojämställda utfall av en AI-lösning bero på överlappande faktorer som bias i data eller att lösningen inte är tillräckligt tränad alternativt används på fel sätt.

Sammanfattning risker:

- Användningen av AI resulterar i att strukturella skillnader mellan kvinnor och män skapas.
- Existerande strukturell ojämställdhet sprider sig till andra samhällsområden genom användningen.
- Strukturell ojämställdhet skapas även om ojämställdhet inte reflekteras i informationen som AI-lösningen är tränad på.
- Användningen av AI reproducerar eller upprätthåller existerande strukturella skillnader mellan kvinnor och män då myndigheter är en del av och agerar i ett samhälle där ojämställdhet finns.
- Användningen av AI orsakar diskriminering på individnivå på grund av kön i enlighet med Diskrimineringslagen (2008:567).

Foto: Shutterstock

METOD FÖR ATT INFÖRA AI MED ETT JÄMSTÄLLDHETSPERSPEKTIV

I projektet har anch.AI:s metod Ramverk för hållbar AI[5] använts för att bidra till insikter om hur införandet av AI inte riskerar att medföra negativ inverkan på utvecklingen av jämställdhet och uppfyllelse av de jämställdhetspolitiska målen.

Ramverket appliceras på en AI-lösning för att identifiera, mäta och hantera risker med denna. Det görs utifrån en hållbar så kallad AI-lins som innefattar ett legalt, tekniskt och samhälleligt perspektiv. I projektet har ett ytterligare perspektiv lagts till: jämställdhet. Metoden är utvecklad tillsammans med representanter från akademisk forskning och experter från olika discipliner. I det här avsnittet beskriver vi hur metoden Ramverk för hållbar AI fungerar.

Ramverk för hållbar AI bygger på att aktivera tvärfunktionellt samarbete i samband med införandet av AI. Därför engageras flera funktioner inom organisationen i att applicera ramverket. Representanter med kompetens inom juridik, teknik, verksamhet och styrning, etik och värdegrund, samt jämställdhet har involverats i arbetet.

Nedan beskrivs respektive ramverksfas med exempel från projektet.

5 Fellander, A, Rebane, J. Larsson, S. Wiggberg, M & Heintz, F. (2021) Achieving a Data-driven Risk Assessment Methodology for Ethical AI

Steg 1: definiera användarfall

Ramverket inleds med en fas där ett användarfall identifieras och beskrivs. Ett användarfall är en specifik användning eller applicering av AI.

När användarfallet definieras, beskrivs följande:

Syftet med användarfallet

- Vilka organisatoriska mål syftar användarfallet till att uppnå?
- Vilka problem löser användarfallet?

Användare och intressenter:

- Vilka är de avsedda användarna av AI-lösningen?
- Finns det några andra intressenter som påverkas av AI-lösningen (både positivt och negativt)?

Teknisk specifikation

- Vilken data används? Hur genereras denna data?
- Vilka modeller och metoder används vid behandlingen av data?
- Resultat: Vad är resultatet av AI-lösningen? (till exempel beslut eller rekommendation)

Att tydligt definiera och avgränsa användarfallet som ramverket ska appliceras på gör att analyser av risker och hur de hanteras blir så precisa som möjligt. Genom att göra en genomlysning av en AI-lösning som redan finns, eller som ska införas, blir diskussionerna konkreta vilket i sin tur avgör hur väl det går att gradera nivån av risk i kommande ramverksfaser. Det var också något som tydligt framkom i projektet.

Steg 2: riskskanning

I ramverkets andra fas identifieras potentiell riskexponering genom en riskskanning.

Projektet har använt verktyg som består av självskattning med hjälp av enkäter. Enkäterna riktar sig till olika funktioner inom organisationen i syfte att skapa tvärfunktionellt samarbete och förstå de olika funktionernas perspektiv på användarfallet.

Resultatet av att applicera verktyget är en riskprofil i form av en rapport som visualiserar potentiell riskexponering. Riskprofilen ger även en bild av hur väl etiska principer är implementerade såsom styrning, ansvar, transparens och förklarbarhet. Sammantaget ger riskprofilen en nulägesbedömning av användarfallet och organisationens mogenhet att hantera etiska och jämställdhetsspecifika risker med AI. Riskskanningsverktyget har även en underliggande insiktsmotor, som med hjälp av historiska aggregerade data kan förutspå och rekommendera åtgärder. Det här steget kan genomföras oberoende av de resterande faserna i ramverket.

Riskskanningsverktyget kan exempelvis användas för att följa upp arbetet med att reducera risker och för att kontrollera att en AI-applikation i utvecklingsfas är mogen nog att lanseras. Vidare kan verktyget användas för att se till att övergripande etiska principer och riktlinjer är implementerade i praktiken.

I projektet inleddes riskskanningsfasen med att tre jämställdhetsspecifika risker kategoriserades och utifrån det skapades nya enkätfrågor. En insikt är att frågeställningarna som inkluderas i riskskanningsverktyget behöver tas fram tidigt i utvecklingsprocessen eftersom de utgör en grund i arbetet med att analysera AI-lösningen ur ett jämställdhetsperspektiv.

Här beskrivs ett urval av riskskanningsfrågorna som utvecklades i projektet.

Utöver dessa användes även frågor ur anch.AI:s verktyg.

RISK	PRINCIP	FUNKTION/PERSPEKTIV FRÅGAN BERÖR	RISKSKANNINGSFRÅGA
Diskriminering på individnivå	Organisation & styrning (AI governance)	Tech	Har ni en process/rutin för att bedöma om AI-lösningen kan vara diskriminerande mot individer baserat på någon av följande grunder: Kön | Könsöverskridande identitet eller uttryck | Etnisk tillhörighet | Funktionsnedsättning | Sexuell läggning | Ålder | Religion eller annan trosuppfattning
	Organisation & styrning (AI governance)	Tech	Har ni en process/rutin för att förebygga att AI-lösningen inte resulterar i könsdiskriminering? (till exempel genom att använda olika fairness-tekniker)
	Organisation & styrning (AI governance)	Verksamhet/ Hållbarhet	Har ni fastställt en åtgärdplan om AI-lösningen resulterar i könsdiskriminerande utfall?

RISK	PRINCIP	FUNKTION/PERSPEKTIV FRÅGAN BERÖR	RISKSKANNINGSFRÅGA
Användning av AI skapar ojämställdhet	Ansvar	HR	Har den funktion som styr och leder utvecklingen av AI-lösningen kunskap för att kontrollera att ojämställdhet inte reproduceras till följd av att AI används?
	Ansvar	HR	Har den funktion som styr och leder utvecklingen av AI-lösningen mandat för att kontrollera att ojämställdhet inte skapas till följd av att AI används?

RISK	PRINCIP	FUNKTION/PERSPEKTIV FRÅGAN BERÖR	RISKSKANNINGSFRÅGA
Användning av AI reproducerar ojämställdhet	Organisation & styrning (AI governance)	Tech	Är data som AI-lösningen har tränats på representativ för hela populationen?
	Organisation & styrning (AI governance)	Tech	Inkluderar ni personer med olika kompetenser och roller i arbetet med att utveckla AI-lösningen?
	Transparens	Verksamhet/ Hållbarhet	Har ni involverat de som direkt påverkas av AI-lösningen (till exempel kunder, medborgare) för att höra hur de påverkas, i syfte att granska / utvärdera jämställdhetseffekterna av AI-lösningen?
	Organisation & styrning (AI governance)	Verksamhet/ Hållbarhet	Finns det ett jämställdhetsperspektiv i ordinarie uppföljningar av myndighetens AI-lösningar?

Steg 3. Riskbedömning

Resultaten i riskprofilen är grunden för att ta fram olika riskscenarier och används som utgångspunkt i en workshop bestående av tvärfunktionella team från organisationen. Riskscenarier utgör konkreta händelser som inträffar om riskerna realiseras. I workshopen beskrivs varje identifierat riskscenario mer ingående.

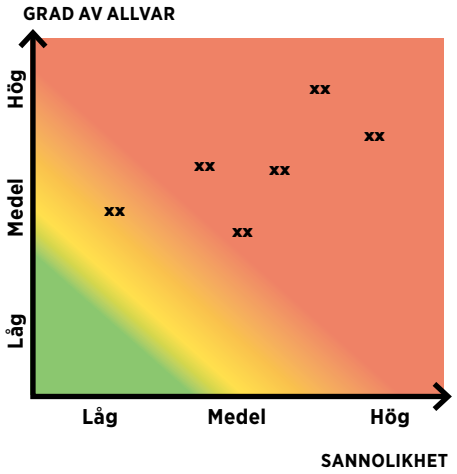
Orsaken till att scenariot inträffar definieras, samt effekterna på jämställdhet och hur det påverkar de jämställdhetspolitiska målen. När effekterna ska beskrivas, betraktas både hur och vem som riskerar att påverkas:

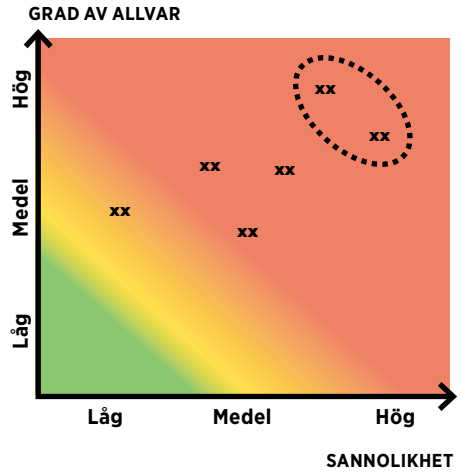
- Vem riskerar att påverkas av riskscenariot? Beskriv primära och sekundära intressenter.
- Vilka grupper av intressenter är mest sårbara om de påverkas?
- Vad blir effekten?
- Kan riskscenariot påverka jämställdhet mellan könen när det kommer till:
  - Fördelning av makt och inflytande
  - Ekonomi
  - Utbildning
  - Obetalt hem- och omsorgsarbete
  - Hälsa
  - Mäns våld mot kvinnor

Resultatet av att identifiera och beskriva respektive riskscenario dokumenteras och presenteras på ett överskådligt vis.

Riskscenario	Möjlig orsak till att scenariot uppstår	Effekter: Vem riskerar att påverkas och hur? (direkt och indirekt)
A. [Scenario]	[Rotorsak]	• [Vem / Vad] – [Hur] • [Effekt på jämställdhet]
B. [Scenario]	[Rotorsak]	• [Vem / Vad] – [Hur] • [Effekt på jämställdhet]

Vidare utvärderas riskscenarierna baserat på sannolikheten att de inträffar, och graden av allvar om de skulle inträffa. Slutligen prioriteras de mest kritiska riskscenarierna till riskmitigeringsfasen utifrån graderingen.





Steg 4: riskreducerande åtgärder

Riskreducering syftar till att minska sannolikheten av att ett riskscenario uppstår genom motverkan av själva rotorsaken, eller att effekterna dämpas om det skulle inträffa

I riskreduceringsfasen av ramverket mitigeras riskscenarierna som har prioriterats från föregående fas. Resultatet är en plan med konkreta åtgärder och utpekade funktioner som är ansvariga för den fortsatta implementeringen av AI-lösningen.

Åtgärderna för riskreducering kan vara långsiktiga initiativ som implementeras i organisationens befintliga processer och strukturer. Syftet är då att systematisera arbetet med att minska jämställdhetsspecifika risker med AI.

Åtgärder kan vara:

- Att inkludera krav på att det finns insyn i en AI-lösnings beslutsfattande (modelltransparens) när myndigheten upphandlar AI.
- Utbildningsinsatser för de som ska delta i att utveckla/upphandla AI.
- Att implementera tekniskt stöd för att kunna mäta hur fördelar och risker med AI-lösningars utfall påverkar olika grupper.

I en workshop med samma tvärfunktionella deltagargrupp som medverkat tidigare identifiera riskreducerande åtgärder och riskägare. Efter workshopen följer varje riskägare upp sina tilldelade åtgärder och ser till att de implementeras i organisationens befintliga processer och strukturer.

Steg 5: slutrapportering

I den sista fasen summeras genomförda aktiviteter och vilka insikter som har genererats i en slutrapport. Resultatet från att applicera ramverket och kunskapen bör spridas till andra delar av organisationen i syfte att skapa medvetenhet och att systematisera lärdomar.

Det kan även finnas anledning att konsultera berörda intressenter, både interna och externa. Involvering av externa intressenter och slutanvändare är viktigt för att samla in direkt återkoppling på hur de påverkas av och interagerar med AI-lösningen. Vidare är det ett sätt att stärka analysen av potentiella risker och hur de bäst hanteras.

Avslutningsvis bör en bedömning göras av hur slutsatserna ska kommuniceras, internt och externt. Det finns anledning att vara transparent och kommunicera kända risker och hur de har hanterats. Det stärker förtroendet för att myndigheten arbetar strukturerat med att implementera jämställda AI-lösningar.

FALLSTUDIER

Foto: Zac Lucero/Unsplash

Metoden som har använts i projektet är Ramverk för hållbar AI. Metoden har applicerats på två testbäddar: AI-lösningar på Skatteverket och hos Försäkringskassan.

Resultaten av fallstudierna ligger till grund för insikterna i den här publikationen. I projektet identifierades endast potentiella risker som teoretiskt skulle kunna uppstå. Appliceringen av ramverket syftade primärt till ett lärande kring AI och jämställdhet.

Skatteverket: textklassificering av övriga upplysningar i inkomstdeklarationer

Skatteverket använder AI för att klassificera text som privatpersoner och företag skrivit i fältet övriga upplysningar i inkomstdeklarationen. Klassificeringen av textinnehållet i fältet sker till någon av de cirka 50 ämnesområden (klasser) som modellen innehåller. Deklarationen fördelas sedan ut till rätt handläggare med hjälp av en fördelningsnyckel som hanterar upplysningen. Vissa övriga upplysningar skulle potentiellt kunna fördelas automatiskt till klassen ingen åtgärd i framtiden.

Tre nyttiggöranden identifierades med användarfallet i början av projektet:

- Ökad kunskap om innehållet i övriga upplysningar.
- Enhetligare, snabbare och mer effektiv hantering av övriga upplysningar. Något som framför allt handlar om att fördela ärenden snabbt till rätt kompetens.
- Potentiellt automatisera hantering av vissa typer av upplysningar

I efterhand har det framkommit att den största nyttan är enhetlig och snabbare hantering.

Genom att applicera ramverket på det aktuella användarfallet identifierades potentiella risker och möjliga åtgärder för hur de kan hanteras. Flera riskscenarier kretsade kring att det potentiellt kan finnas könsbias i data som lösningen har tränats på i form av att kvinnor och män uttrycker sig olika i skrift. Det kan medföra att lösningen fungerar bättre eller sämre på att klassificera upplysningar från kvinnor eller män.

Vidare identifierades även risker kopplade till valet av klasser som ingår i modellen. Eventuellt är kvinnors eller mäns upplysningar mer kategoriövergripande, och således svårare att klassificera till de utvalda klasserna som ingår i modellen. Det kan i sin tur bero på mänsklig bias eller brist på analys när ämneskategorierna (klasserna) definierades.

Avslutningsvis identifierades mer organisatoriska risker, exempelvis att jämställdhetsperspektivet inte tillvaratas tillräckligt i existerande rutiner, processer och styrning kopplat till utveckling av AI-lösningar.

I fasen för att minska risker i ramverket identifierades åtgärder för att förbättra processer kopplade till urval- och analys av träningsdata, samt modellering. Likaså identifierades åtgärder för att integrera ett jämställdhetsperspektiv i utvärdering av AI-lösningars utfall, samt myndighetens uppföljning och kvalitetssäkring av AI-lösningar. Andra åtgärder för att minska risker fokuserade på utbildning och kunskapsupphöjande aktiviteter, samt etablerandet av en permanent intern funktion med uppdrag att stötta med tvärfunktionell expertis.

Foto: Shutterstock

Försäkringskassans Automl-Plattform

Verktyget som Försäkringskassan använder, AutoML-plattform, används till exempel för att identifiera funktionsnedsättningar och aktivitetsbegränsningar. AutoML-plattformen är utvecklad för att stödja "end to end" machine learning pipelines[6].

Plattformen används av flera olika verksamheter för att realisera olika verksamhetsbehov kopplat till AI. Det innebär bland annat integration till olika datakällor, berikning av data, livscykelhantering av datamodeller, träning av modeller, utvärdering, publicering, analys samt efterlevnad.

Genom att applicera ramverket på Försäkringskassans plattform identifierades potentiella risker och åtgärder för hur de bör hanteras. Flera risker handlade om att jämställdhet inte ingår som ett specifikt perspektiv i analyser och bedömningar kopplade till data och modell, vilket kan leda till att analyser och bedömningar endast omfattar det strikt rättsliga, som till exempel diskrimineringslagen eller dataskydd enligt GDPR. En möjlig orsak är att det saknas insikt om att jämställdhet är centralt i den statliga förvaltningen, och att det således kräver en bredare ansats än det strikt rättsliga. När det saknas ett uttalat jämställdhetsperspektiv i riktlinjer och rutiner kopplade till införande av AI finns risken att personal med särskild kompetens inom jämställdhet inte involveras i tillräcklig utsträckning. Konsekvensen kan bli att det förs in fel data när träningsdata väljs ut och features och variabler skapas i modelleringen.

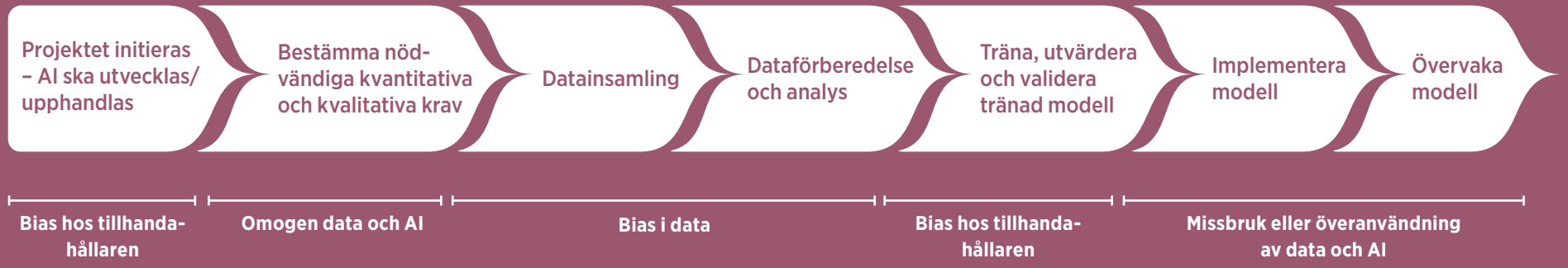
I riskreduceringsfasen av ramverket identifierades ett behov av jämställdhetsperspektiv i styrdokument och processer kopplade till införande av AI.

Följande bör ingå:

- Instruktioner kring hur träningsdata ska tas fram för att minska könsbias.
- Information om hur data ska analyseras/testas för könsbias.
- Tvärfunktionell kompetens, inklusive jämställdhet, ska involveras i de olika faserna att införa AI.
- Information om hur kvalitet ska mätas.
- Information om hur AI-lösningar ska utvärderas, kontrolleras och eventuell modifieras efter implementering.

6 End to end machine learning innebär att träna en modell och applicera den i en produkt som skapar värde för organisationens verksamhet

För att upptäcka och undvika fallgroparna behöver vi förstå när i utvecklingsprocessen de bör kontrolleras för. I modellen visualiseras en konceptuell utvecklingsprocess som visar hur fallgroparna kan uppstå i de olika faserna i utvecklingen.



Fallgropar för jämställdhet i AI

Ramverket som har använts i projektet bygger på att synliggöra grundorsaker till att risker med AI uppstår. I ramverksmetoden kontrolleras AI-lösningar utifrån fyra fallgropar

1. Missbruk eller överanvändning av data eller AI: AI-lösningen eller data överanvänds, eller används för oavsiktliga ändamål.

2. Omogen data och AI: AI-lösningen tränas på data som inte är validerat för ändamålet, eller så är AI-lösningen inte tillräckligt tränad.

3. Bias hos tillhandahållaren: Oönskade värderingar och partiskhet byggs avsiktligt eller oavsiktligt in i AI-lösningen på grund av partiskhet hos organisationen i stort, eller dem som utvecklar lösningen. Brist på tillräcklig kunskap i hur etiska principer och värdegrunder ska kodas.

4. Bias i data: AI-lösningen tränas på data som inte återspeglar verkligheten eller den föredragna verkligheten på ett korrekt sätt.

Principer för att undvika fallgropar

Ramverksmetoden utgår från fyra övergripande etiska principer för att undvika fallgroparna. Principerna återfinns i de etiska riktlinjerna för AI som Europeiska kommissionens expertgrupp tagit fram.

Organisation & styrning: Etablerandet av policyer, principer och/eller protokoll samt kontinuerlig övervakning att de implementeras korrekt. Det innebär att skapa skalbara kontrollsystem.

Ansvar: Vikten av att stå ansvarig för beslut och rekommendationer från AI-lösningar till partners, användare och andra intressenter som systemet interagerar med eller påverkar. Ansvarstagande för införande av AI, insamling och behandling av data, utveckling och design av algoritmer, policyer, riktlinjer och rutiner (liksom administration, styrning och implementering).

Transparens: Det ska vara möjligt att kunna identifiera, spåra och upptäcka hur och varför ett system tog ett visst beslut eller agerade på ett visst sätt och, om systemet gör skada, kunna upptäcka grunden till problemet. Det innebär också att organisationen är öppen gentemot intressenterna i att beskriva AI-lösningens roll och funktion.

Förklarbarhet: Att algoritmiska beslut, liksom den data som driver besluten, kan förklaras för slutanvändare och andra intressenter i icke-tekniska termer. Att skapa den rätta nivån på förklarbarhet till den specifika intressenten i fråga.

PROJEKT INITIERAS – AI SKA INFÖRAS

Det här avsnittet beskriver hur statlig förvaltning kan beakta jämställdhet vid införandet av en AI-lösning.
Resultatet omfattar situationer både när AI utvecklas internt och när det upphandlas.

Projekt initieras – AI ska införas

När en verksamhet beslutar att AI ska införas behöver verksamheten där AI ska användas analyseras ur ett jämställdhetsperspektiv. Den mottagande verksamheten redogör för vilka jämställdhetsrisker som finns inom det specifika verksamhetsområdet.

Analysen behöver involvera tvärfunktionell kompetens där jämställdhets-kompetens ingår för att identifiera jämställdhetsspecifika risker. En vanlig risk är att endast rättslig kompetens involveras då jämställdhet antas ingå i rättslig bedömning.

Slutligen behöver det bli klart vem som är ansvarig för att jämställdhet ingår i analyser och bedömningar längs hela införandeprocessen.

Bestämma nödvändiga kvantitativa och kvalitativa krav

Tidigt i processen bör det specificeras hur AI-lösningen ska utvärderas för att den ska anses vara värd investeringen. AI-lösningar utvärderas vanligtvis utifrån ett kombinerat verksamhets-, tekniskt- och finansiellt perspektiv. För att lösningen inte ska få oönskade konsekvenser på jämställdhet bör traditionella utvärderingskriterier kompletteras med analyser och utvärderingar specifika för jämställdhet.

Exempelvis bör organisationer:

- Definiera sitt förhållningssätt till hur likvärdiga/rättvisa utfall av AI-lösningar ska mätas på gruppnivå (på engelska kallat AI Fairness). Det finns idag många definitioner av vad fairness är. Generellt syftar AI fairness på att fördelar och risker med AI-lösningar fördelas jämnt mellan olika användargrupper. Det vill säga att undvika scenarion där någon grupp systematiskt gynnas eller missgynnas av AI-lösningens utfall.

- Bestämma vilka mått som AI-lösningens kvalitet ska utvärderas på och vilket tröskelvärde som ska passeras för att modellen ska ses som godkänd ur ett jämställdhetsperspektiv.

- Bestämma vad den ”acceptabla” nivån av eventuella skillnader i AI lösningens utfall är mellan grupper om det skulle visa sig att lösningen fungerar bättre/sämre på gruppen kvinnor/män och hur denna skillnad rättfärdigas till olika intressenter.

- Om AI upphandlas ska kravställningen till leverantörer kunna redogöra för hur lösningen uppfyller motsvarande kriterier som hade gällt om AI utvecklats internt på myndigheten.

Foto: Magda Ehlers/Pexels

Datainsamling och analys

När data samlas in, förbereds och analyseras ska eventuell oönskad könsbias i data identifieras. Bias i data kan uppstå på grund av flera orsaker.

Bias i urvalet: Datamängden som AI-lösningen tränas på är antingen inte tillräckligt stor eller representativ för populationen.

Historisk bias: Myndigheter är en del av, och agerar i, ett samhälle där ojämställdhet finns. Därför kan data återspegla befintliga fördomar och stereotyper om kön, vilket riskerar att reproduceras i AI-lösningens utfall. Det kan hända att myndighetsutövningen historiskt varit biased, vilket riskerar att reproduceras om det inte upptäcks och hanteras.

Bias i mätningen: Den data som använts mäter inte det avsedda syftet på att objektivt sätt, vilket riskerar att leda till att datapunkterna inte är sanna. Detta kan exempelvis ske om data är självrapporterat.

Exkludering av data: Bias kan även uppstå om det saknas datapunkter för särskilda variabler, eller om vissa datapunkter inte ingår i modellen. Ett vanligt sätt att hantera könsbias är att exkludera kön från modellen.

Här behöver man uppmärksamma att det kan finnas andra variabler som fungerar som Proxy variabler för kön. Därför behöver en tränad modells utfall mätas och analyseras utifrån ett jämställdhetsperspektiv. Det beskrivs i nästa avsnitt.

Checklista för tidig validering av data (innan modellutveckling):

- Har ni upptäckt och potentiellt tagit bort variabler som starkt korrelerar med kön, och som introducerar bias?
- Är könsfördelningen i urvalet av träningsdata representativt för populationen?
- Är könsfördelningen i urvalet av testdata representativt för populationen?

Om data inte är representativt för populationen bör myndigheten utreda om det finns möjlighet att samla in mer data för att få bättre representation, alternativt att se över möjligheten att använda urvalstekniker eller syntetiska data. I projektet har potentiella risker identifierats för ojämställdhet om man tränar AI på data från verkligheten som innehåller strukturell ojämställdhet och inte från den önskvärda verkligheten. Om risken är identifierad finns en möjlighet vid införandet av AI att på ett systematiskt sätt optimera för den ”önskvärda verkligheten”. Det är ett exempel på hur AI är ett medel för att öka jämställdhet.

Modellering och analys av utfall

Från tillgängliga rådata skapas variabler (features) som ska ingå i modellen. Här kan mänsklig bias introduceras, därför behöver det tvärfunktionella teamet delta och att särskild kompetens inom jämställdhet finnas representerad.

I modelleringsfasen behöver modellens utfall kunna förstås och förklaras av olika intressenter, inte enbart av personer med god teknisk förståelse. Det kräver transparens i modellen, det vill säga möjlighet att förstå hur stor inverkan olika variabler har på utfallet av AI-lösningen. Det varierar hur bra olika AI-tekniker kan förklaras. Deep learning är vad som i dagsläget är svårast att förklara på grund av komplexiteten i beräkningarna, medan linjära regressioner och beslutsträd är mer förklarbara. Det finns även metoder för att i efterhand förklara komplexa modeller, exempelvis LIME eller SHAP. Förklarbarhet säkerställer också att modellen inte diskriminerar. Genom att kontrollera förklaringarna över vilka variabler som driver ett utfall kan diskriminerande utfall upptäckas.


Foto: Shutterstock

Implementering och kompetens

När en tränad modell anses vara kvalitativ nog att implementeras i en lösning ska det finnas transparens. Berörda intressenter bör få information om hur AI används särskilt om det används i myndighetsutövning, men även om det används i servicesyften. Dokumentation över hur AI-lösningen har tagits fram bör finnas. Om det finns några kända brister eller svagheter bör de kommuniceras till relevanta intressenter.

När AI används i ett beslutsstöd för exempelvis handläggare på en myndighet behöver handläggarna få en gedigen träning i att använda och förstå AI-lösningen. Det gör att personal kan kommunicera information, till slutanvändarna, om AI-lösningens funktion i beslutsfattandet. De som interagerar med AI-lösningen ska kunna granska, och upphäva, utfallen vid behov.

Här etablerar projektet en åtgärdsplan för hur myndigheten ska hantera en situation om det upptäcks att AI-lösningen diskriminerar, skapar eller reproducerar ojämställdhet efter implementering.

Uppföljning och modifiering

Det bör finnas tydliga instruktioner hur en AI-lösning ska följas upp efter implementering. Här ska själva modellen fortsättningsvis hålla hög kvalitet, det vill säga att de mätvärden som den utvärderats utifrån inte förändras över tid. Den andra delen är att kontinuerligt bevaka området jämställdhet i syfte att förstå om ny forskning, trender och nya upptäckter innebär att AI-lösningar bör modifieras. Det bör ingå som ett perspektiv i ordinarie uppföljningar av implementerade AI-lösningar

SAMMANFATTANDE INSIKTER

Foto: Cedric Dhaenens/Unsplash

I pilotprojektet har vi undersökt potentiella risker för ojämställdhet när AI ska införas på en myndighet. Inom projektet har vi kommit fram till följande insikter.

Analysera verksamheten ur ett jämställdhetsperspektiv

Innan en myndighet utvecklar en verksamhet med AI är det viktigt att tidigt analysera den verksamhet som berörs av projektet ur ett jämställdhetsperspektiv. Att ha med ett jämställdhetsperspektiv innebär att man genomför en målgruppsanalys och en kartläggning över hur myndighetens verksamhet påverkar kvinnor och män i syfte att upphäva strukturella orättvisor.

Då kan myndigheten lättare ta hänsyn till eventuella risker när AI-lösningen designas och när data samlas in och valideras.

Om verksamheten deltar tidigt och redogör för befintliga risker för ojämställdhet ökar möjligheten att bygga bort ojämställdhet i AI-lösningen. En sådan analys kan även vara hjälpsam i att förstå om användning av AI kan bidra till att identifiera risker för ojämställdhet i delar av verksamheten som reproducerar eller skapar ojämställdhet.

Tvärfunktionella grupper skapar förutsättningarna för införandet av jämställd AI

Jämställdhetsintegrering innebär att ett jämställdhetsperspektiv införlivas i alla steg av AI-processen av de aktörer som deltar i den. Det innebär att olika funktioner och perspektiv aktiveras vid införandet av AI i syfte att förbättra verksamheten. Testbäddarna i detta projekt vittnar om att tvärfunktionella team skapar intressanta och lärande diskussioner och ger möjligheter att upptäcka olika risker som påverkar utfallet för jämställdhet i införandet av AI. När testbäddarna genomförde riskanalyser med ett team som bestod av kärnverksamhet, stödfunktioner samt ledningss- och verksamhetsutvecklingsfunktioner ökade deras förmåga att identifiera risker.

Jämställdhetsintegrering ger således en kvalitativ skillnad jämfört med en enklare jämställdhetsbedömning i slutet av en införandeprocess av AI.

Arbeta systematiskt

För ett systematiskt angreppssätt är det grundläggande att etablera en metod som identifierar och hanterar potentiella jämställdhetsrisker med enskilda AI-lösningar. Varje steg i införandeprocessen av AI och dess livscykels olika faser bör inkludera ett jämställdhetsperspektiv. En del av ett systematiskt angreppssätt är att identifiera och hantera potentiella jämställdhetsrisker med enskilda AI-lösningar, synliggöra hur kvinnor och män påverkas på olika sätt av användningen av en AI-lösning samt göra ändringar vid behov.

Jämställd AI utgår från god förvaltningskultur och statlig värdegrund

Jämställdhet skapas där resurser fördelas, beslut fattas och normer skapas och bör därför vara en tydlig del i verksamhetsplaneringen. En del i myndighetens verksamhetsplanering är att bestämma var ansvaret för uppföljning av mål för verksamheten ligger, samt se till att rutiner och processer efterlevs. I Skatteverkets och Försäkringskassans arbete med jämställdhetsintegrering framgår det till exempel att AI-utveckling är en del i främjandet av mer jämställt utfall i verksamheten.

AI som statliga myndigheter använder i sin verksamhet ska vila på och utgå från den statliga värdegrunden. Den statliga värdegrunden består av ett antal principer som syftar till att säkerställa en god förvaltningskultur. Principerna, som utgår från grundlagen och andra lagar och föreskrifter, beskriver hur de anställda i staten ska agera för att säkerställa en effektiv, rättssäker och fungerande förvaltning. Mot bakgrund av den statliga värdegrunden kan det uppstå målkonflikter mellan till exempel effektivitet och att uppnå jämställdhet i verksamheten. Men behöver dock beakta att kvalitetsutveckling ofta tar tid och resurser i anspråk.

Foto: Matteo Vistocco/Unsplash

BILAGA – BEGREPPSLISTA

AI
Artificiell intelligens avser system som uppvisar intelligent beteende genom att analysera sin miljö och vidta åtgärder, med viss grad av självständighet, för att uppnå särskilda mål.

AI-baserade system kan vara helt programvarubaserade och fungera i den virtuella världen (till exempel röstassistenter, bildanalysprogram, sökmotorer, tal- och ansiktsigenkänningssystem), eller inbäddas i hårdvaruenheter (till exempel avancerade robotar, självkörande bilar, drönare eller applikationer för sakernas internet).

AI Fairness
Ett tillstånd där fördelar och risker med en AI-lösning är jämnt fördelade mellan olika användargrupper. Syftar även på metoder för att utvärdera och maximera fairness i AI-lösningar7.

Användarfall
En specifik tillämpning av AI som planeras att utvecklas, håller på att utvecklas eller är implementerad i verksamheten.

Automatiserat beslutsfattande
Automatiskt beslutsfattande omfattar i vår definition algoritmiska och automatiserade beslutsprocesser, med eller utan AI. I begreppet inkluderar vi både helt automatiserat beslutsfattande och automatiserade processer som används som beslutstöd för handläggare.

Bias
Bias i data: Informationen återspeglar inte verkligheten, eller den föredragna verkligheten. Vissa delar i informationen är systematiskt mer viktade och/eller bättre representerade.
Mänsklig bias: Partiskhet, subjektivitet eller fördomar för eller emot en person eller grupp, i regel på ett sätt som anses vara orättvist.

Könsbias
Antaganden om kvinnor och män som bygger på stereotypa föreställningar, finns representerat i data, hos människor, i samhällsstrukturer etcetera.

Proxyvariabel
En variabel som speglar information i en annan variabel. En persons längd kan till exempel vara en proxy för personens kön.

Testbädd
I detta projekt: deltagande myndighet som använder AI i sin verksamhet, i detta fall avgränsat till en enskild AI-lösning.

7 End to end machine learning innebär att träna en modell och applicera den i en produkt som skapar värde för organisationens verksamhet

JÄMSTÄLLDHETS MYNDIGHETEN

www.jamstalldhetsmyndigheten.se

VINNOVA
Sveriges innovationsmyndighet

www.vinnova.se

www.anch.ai